Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
Premier ministre
SECRÉTARIAT GÉNÉRAL POUR LA MODERNISATION DE L'ACTION PUBLIQUE

Administrateur général des données

Rapport au Premier ministre sur la gouvernance de la donnée 2015

Les données au service de la transformation de l'action publique

Décembre 2015

Le décret du 16 septembre 2014 institue, auprès du Premier ministre, un Administrateur général des données (AGD), rattaché au secrétaire général pour la modernisation de l'action publique.

L'Administrateur général des données coordonne l'action des administrations en matière d'inventaire, de gouvernance, de production, de circulation et d'exploitation des données.

Il organise, dans le respect de la protection des données personnelles et des secrets protégés par la loi, la meilleure exploitation de ces données et leur plus large circulation, 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.

Dans la poursuite de ces objectifs, l'Administrateur général des données propose au Premier ministre toutes mesures, y compris, le cas échéant, des évolutions législatives ou réglementaires.

Aux termes de ce décret, l'Administrateur général des données est donc chargé de stimuler le meilleur usage des données par l'administration - notamment en soutenant la diffusion des méthodes dites « de datasciences » - et d'encourager la meilleure circulation des données. L'Administrateur général des données peut être saisi par toute personne de toute question portant sur la circulation des données. Les collectivités territoriales, les personnes morales de droit public et les personnes morales de droit privé chargées d'une mission de service public peuvent le saisir, pour avis, de toute question liée à l'utilisation par leurs services des données des administrations.

La site de l'Administrateur général des données permet de réaliser ces saisines et de découvrir différents premiers résultats issus des datasciences : http://agd.data.gouv.fr

Enfin, le décret précise que l'Administrateur général des données remet chaque année au Premier ministre un rapport public sur l'inventaire, la gouvernance, la production, la circulation, l'exploitation des données par les administrations.

Ce rapport fait notamment état des données existantes, de leur qualité ainsi que des exploitations innovantes que ces données autorisent. Il présente les évolutions récentes de l'économie de la donnée. Il contient des propositions visant à améliorer l'exploitation et la circulation des données entre les administrations.

Le présent rapport est le premier de ces rapports annuels.

TABLE DES MATIÈRES

INTRODUCTION

Prédire et empêcher les vols de voitures ; optimiser les temps d'attente aux urgences ; mieux cibler les contrôles douaniers ; détecter les immeubles passoires énergétiques ; repérer les entreprises qui vont prochainement recruter et les signaler aux demandeurs d'emploi concernés ; affecter les remplaçants aux académies qui vont manquer d'enseignants ; optimiser les feux de circulation pour désengorger et dépolluer les centres-villes ; réviser la formule de calcul des prix des médicaments pour les optimiser ; négocier les achats d'électricité en anticipant en en contrôlant les pics de consommation ; mieux négocier les achats publics ; prédire les effets microéconomiques d'une réforme fiscale ; anticiper les besoins d'investissement médical grâce à l'analyse de la littérature scientifique… tous ces **usages de l'analyse prédictive sont à portée de main de la puissance publique.** Ils recèlent un immense potentiel d'efficacité, de maîtrise des dépenses et de justice de l'action publique.

L'analyse prédictive n'est que l'une des modalités d'un ensemble de nouvelles pratiques, les « stratégies fondées sur la donnée », qui permettent par exemple :

- **de réguler un secteur industriel** par la mise en circulation judicieuse des données pertinentes – comme l'expérimente le gouvernement en utilisant les données de géolocalisation des taxis pour leur permettre de bénéficier des apports de clientèle issus de nouveaux services numériques ;
- d'organiser l'information pour que les **agents puissent prendre individuellement de meilleures décisions ;**
- **d'améliorer l'action quotidienne des agents de guichet** en leur donnant plus d'informations temps réel ;
- **d'augmenter l'autonomie et la liberté de choix des usagers du service public** – en prédisant par exemple l'espérance de succès d'une démarche en justice.

Cette promesse des datasciences, qui est aujourd'hui au cœur de la transformation numérique de grandes entreprises et de grandes villes dans le monde entier, est l'un des leviers de la modernisation de l'action publique. En créant la fonction d'Administrateur général des données, le 16 septembre 2014, le Premier ministre a souhaité intégrer ces approches dans la palette des outils de la réforme de l'État, quelques mois avant que les États-Unis, puis la Grande-Bretagne, ne fassent de même.

Cette ambition suppose l'**intégration dans l'État de nouvelles compétences :** les datascientists, ces statisticiens au profil innovant, férus d'informatique et de nouvelles méthodes de traitement de la donnée, et attentifs à la traduction en actes de leurs résultats mathématiques. Elle suppose des **données de qualité** – que la France, précisément, produit et manipule depuis longtemps grâce à sa grande tradition de statistique publique et à son attachement à la qualité du service public. Elle suppose enfin une culture accrue des **« stratégies fondées sur la donnée »**, une volonté de conduire ce type de changement et la patience de tester et de vérifier sans relâche si de petites améliorations permettent de produire de grands résultats.

La mise en œuvre correcte de ces méthodes suppose cependant au préalable une véritable **gouvernance de la donnée,** c'est-à-dire une organisation globale des données produites ou détenues par l'État permettant d'en assurer la qualité, la fraîcheur, l'interopérabilité, la disponibilité dans des formats techniques en facilitant l'utilisation rapide et la meilleure circulation possible afin que chaque agent public – de l'État comme des collectivités locales - bénéficie des informations nécessaires à l'exercice de ses missions, et ce, dans le respect des secrets légaux qui protègent d'importantes libertés fondamentales et les intérêts fondamentaux de la nation. Une organisation qui assure à l'État la maîtrise et la souveraineté sur ses données, ses processus et ses systèmes, et au citoyen les points de transparence qu'il est fondé à revendiquer.

Lutter contre le vol de voiture à l'aide des datasciences

L'Administrateur général des données et le Service des technologies et des systèmes d'information de la sécurité intérieure (ST(SI)²) collaborent à l'élaboration d'un modèle prédictif des vols de véhicules. L'objectif est d'aboutir à une allocation optimale des patrouilles de police et de gendarmes dans le département de l'Oise. A l'aide de plus de 600 variables géographiques et socio-économiques ainsi que d'autres indicateurs tels que la météo et l'occurrence de vols les jours précédents et dans les quartiers voisins, l'équipe de l'AGD*[1] a mis au point un modèle dont les résultats sont prometteurs. Allouer quotidiennement des patrouilles dans les 10 % de quartiers les plus risqués (probabilités estimées par le modèle) permettrait de couvrir plus de la moitié des vols des 5 derniers mois de 2014. Une collaboration directe avec les services de la sécurité intérieure présents sur le terrain est maintenant envisagée afin de construire un outil sur-mesure et adapté aux problématiques métiers.

[1] *Tous les termes suivis d'un * sont définis en annexe de ce document.*

Pour être mises au service d'une ambition de réforme de l'État, ces méthodes doivent être intégrées dans une « **politique de la donnée** ». La place croissante que prennent les données et les algorithmes dans la société peut parfois inquiéter. Mais elle représente essentiellement une capacité accrue pour l'action publique, qu'il convient de savoir mobiliser et de savoir encadrer. Elle nécessite des points de transparence et des contre-pouvoirs indépendants. Elle nécessite la garantie d'un usage démocratique et contrôlé de ces données de l'État, la transparence sur les objectifs et les résultats, la coopération avec la société civile. En 1978, le législateur avait organisé les principales sécurités permettant de protéger la vie privée, créant un cadre d'analyse et de régulation qui est toujours d'actualité quarante ans plus tard. En 2016, d'autres sécurités, portant sur l'éthique des algorithmes, sur l'ouverture de la décision et de l'action publique, seront sans doute nécessaires[2]. Cette démarche nécessitera un important effort pédagogique et générera d'importants débats, notamment sur le poids de l'intérêt individuel face à l'intérêt général dans les algorithmes d'optimisation.

C'est pourquoi, en instaurant la fonction d'Administrateur général des données, le Premier ministre lui a confié le soin de préparer chaque année un **rapport annuel sur la gouvernance de la donnée,** permettant de mesurer les progrès réalisés dans la qualité, la circulation et l'utilisation des données de l'État, mais aussi de mesurer l'appropriation administrative, politique et sociale de ces méthodes, et la construction démocratique des stratégies fondées sur ces méthodes.

Ce premier rapport, fondé sur une année d'enquêtes, d'échanges et d'expérimentations avec de nombreux agents publics et de nombreuses administrations a vocation à poser le cadre d'analyse, à cerner les promesses et les illusions des sciences de la donnée, à présenter de premiers résultats, à signaler les premières difficultés rencontrées et à suggérer de premières orientations.

Il vise aussi à inscrire la révolution en cours dans une histoire : celle de l'administration et de l'État français. Sans reprendre la longue genèse, déjà bien documentée, de la statistique publique[3], il importe, en effet, de garder à l'esprit que ces nouvelles méthodes n'apparaissent pas, comme le prétendent certains prophètes du numérique avec le *big data**. Elles prennent place dans une longue histoire des relations entre le pouvoir et la statistique, marquée par la création d'opérateurs importants, par l'instauration d'une gouvernance et d'une économie spécifique, sécurisée par l'instauration de processus, de règlements et de législations[4], profondément travaillée, également, par une organisation globale de l'informatique de l'État. Que cette histoire soit aujourd'hui bouleversée par l'apparition de grandes quantités de nouvelles données numériques, de nouveaux outils et de nouveaux acteurs, mais aussi de nouvelles philosophies de l'action elle-même, n'empêche pas qu'il faille la connaître, la comprendre et l'estimer à sa juste valeur.

C'est pourquoi, **ce premier rapport annuel** de l'Administrateur général des données s'efforcera, en gardant son ambition pratique et opérationnelle, de remettre l'actualité de l'usage des données dans sa perspective historique.

La première partie permettra d'analyser le rôle des données dans l'action publique, et notamment :

- le développement progressif de données de plus en plus variées et précises, avec le tournant numérique de la prolifération des données issues de l'informatisation des systèmes de gestion ;
- les utilisations de plus en plus nombreuses de ces données, et donc les ambitions légitimes qui peuvent être fondées sur ces nouvelles pratiques.

La deuxième partie permettra d'identifier les principaux freins constatés, après un an d'expérimentation, au bon usage de ces données et d'interroger ainsi les principes actuels – même s'ils sont tacites – qui sous-tendent la gouvernance par défaut qui existe aujourd'hui. En effet, en l'absence de gouvernance explicite, les règles qui président aux relations entre administrations, ou les choix d'architecture qui organisent l'informatique de l'État, dessinent, de facto, une gouvernance par défaut.

Enfin, la troisième partie de ce rapport s'efforcera de présenter les principales pistes d'organisation d'une nouvelle gouvernance de la donnée, orientée vers l'efficacité globale du service public, et de présenter de premières dispositions pouvant être mises en œuvre, parfois à très court terme.

Les datasciences au service de l'emploi

Pour aider les demandeurs d'emploi à retrouver un emploi, on peut essayer de trouver l'offre qui leur convient le mieux. On peut aussi repérer les entreprises susceptibles de les embaucher et les encourager à candidater dans ces entreprises. C'est le projet de La Bonne Boîte, une startup d'État développée par Pôle emploi et le SGMAP. En utilisant des données économiques décrivant l'entreprise et notamment l'historique des embauches, le modèle permet de prédire dans chaque secteur et dans chaque département la probabilité qu'une entreprise embauche dans les six prochains mois. Le demandeur d'emploi peut alors cibler sa recherche sur les entreprises ayant la plus grande probabilité d'embauche dans son secteur.

[2] A cet égard, la France peut se féliciter d'avoir introduit, suite à la concertation publique sur le projet de loi pour une « République numérique » un principe général de transparence sur les algorithmes ayant des conséquences dans la vie des citoyens, qui sera prochainement débattu au Parlement.

[3] Desrosières A. (2000) : La Politique des grands nombres : histoire de la raison statistique, Editions La Découverte (2nde édition)

[4] On pourrait souligner par exemple l'importance accordée par le Conseil National de la Résistance à l'existence d'une statistique publique indépendante et de qualité, qui a directement donné naissance à la loi de 1951 sur la statistique publique, et donc à la fois à l'INSEE que nous connaissons aujourd'hui et au secret statistique qui structure aujourd'hui les échanges de données statistiques entre l'État et les entreprises.

Rédigé sous la responsabilité de l'Administrateur général des données, il a bénéficié des contributions significatives des équipes du SGMAP, et notamment de l'équipe d'Etalab au sein de la DINSIC, ainsi que des échanges avec l'ANSSI, l'APIE, le CGEIET, la CNIL et l'INSEE et les DSI ministérielles, ainsi que du concours essentiel de Simon Chignard, au sein de la mission Etalab.

L'Administrateur général des données, un an d'action

Au cours de sa première année d'exercice, l'Administrateur général des données a travaillé pour atteindre trois objectifs : mettre les datasciences au service des politiques publiques, faciliter la circulation des données au service des politiques publiques et faire levier sur les écosystèmes, tant au sein qu'à l'extérieur des administrations publiques.

1/ Mettre les datasciences au service des politiques publiques

En 2015, l'Administrateur général constitué une équipe de quatre « datascientists », proposant ses services aux administrations, a réussi, en moins d'un an, à produire plusieurs résultats encourageants avec des ministères volontaires. On notera en particulier les résultats suivants :

- une mission avec le Service des achats de l'État, permettant d'analyser en détail la consommation d'électricité de l'État et de nourrir ainsi une stratégie d'achat optimisée[5] ;
- une mission avec le Service des technologies et des systèmes d'information de la sécurité intérieure ayant permis de développer un modèle de prédiction des vols de voiture à l'échelle d'un département ;
- une mission menée avec les équipes de Pôle emploi permettant de prédire avec une probabilité de 80 % une entreprise qui recrutera un profil donné dans le trimestre à venir, et qui a permis à Pôle emploi appuyé par le SGMAP de développer le service « La bonne boîte[6] ».

2/ Faciliter la circulation des données au sein des administrations

L'Administrateur général des données a mis en place une procédure de saisine qui permet à quiconque de faire connaître des difficultés liées à une mauvaise circulation des données.

Au cours de l'année 2015, une douzaine de saisines, en provenance d'agents des services de l'État ou des collectivités, de chercheurs, de journalistes, d'entreprises ou d'individus portant des projets d'utilisation de données, ont été traitées. Par ailleurs, le Ministère de la Santé et des Affaires sociales a sollicité un avis formel de l'AGD dans le cadre de la préparation du projet de loi Santé[7].

Un soutien a été apporté à la mission en charge d'évaluer les ventes de données entre administrations pilotée par Monsieur Antoine Fouilleron de la Cour des Comptes.

3/ Faire levier sur les écosystèmes

Au cours de l'année 2015, l'Administrateur général des données a préparé un marché cadre d'appui en datasciences. Ce marché entre dans la stratégie générale du SGMAP (Secrétariat général pour la modernisation de l'action publique) d'appui aux administrations. Ces dernières auront donc la possibilité, en 2016 de bénéficier de ressources supplémentaires pour mener des projets de datasciences.

Henri Verdier
Administrateur général des données
Décembre 2015

[5] *Ce travail est documenté par le SAE et l'AGD sur le site de l'AGD : https://agd.data.gouv.fr/2015/05/17/analyser-les-consommations-energetiques-des-batiments-publics/*
[6] *http://labonneboite.pole-emploi.fr/*
[7] *https://agd.data.gouv.fr/2015/04/02/avis-portant-sur-la-publication-la-rectification-et-la-reutilisation-des-informations-portant-sur-les-professionnels-de-sante/*

1

Les données au cœur de l'action publique

1. L'ÉTAT, PRODUCTEUR DE DONNÉES

L'État est depuis longtemps producteur de données pour ses besoins propres et ceux de la société. L'excellence de ses grands opérateurs producteurs de données est reconnue. L'importance économique et sociale des données géographiques, cadastrales ou météorologiques, l'importance scientifique et démocratique d'une statistique publique fiable et indépendante, tout comme celle des données légales qui incarnent le principe constitutionnel de publicité de la justice, sont aujourd'hui difficiles à quantifier tant leurs usages sont omniprésents.

Avec la baisse des coûts de production de ces données, l'apparition de nouvelles stratégies de production – incluant éventuellement les citoyens eux-mêmes – et la démultiplication de données de gestion – à visée initiale non scientifique – la nature et la portée de ces données produites par l'État changent d'échelle et peut-être de nature. Elles appellent alors de nouvelles organisations.

Une très brève histoire des données de l'État

La construction progressive de l'État moderne s'accompagne de celle d'un ensemble de données de références nécessaires à son organisation et à son fonctionnement, ainsi que de données indispensables au bon fonctionnement de l'économie et de la société. Dès le XVIIe Siècle, le pouvoir royal a cherché à codifier et standardiser la tenue des registres paroissiaux qui constituaient à l'époque la meilleure source de **connaissance** sur la population. Parallèlement à la mise en place de ces registres, le pouvoir a développé une connaissance de plus en plus fine du **territoire,** à des fins de défense ou de calcul de l'impôt. Ainsi l'entreprise cartographique lancée sous Louis XIV aboutira à la première carte complète du royaume. Ainsi également le cadastre napoléonien permet de répartir l'impôt entre les citoyens, tandis que le premier bureau de la statistique organise le recensement général de la population en 1801. Dès l'Ancien Régime, les cartes et relevés effectués par le service hydrographique officiel sont mises à disposition des « vaisseaux tant de guerre que de commerce[8] ».

La plupart de ces données ont été construites pour répondre à des **besoins** de la puissance publique et pour accompagner son développement. Leur finalité initiale détermine bien souvent, aujourd'hui encore, leur ministère de tutelle : le cadastre est aujourd'hui géré par la direction générale des finances publiques, et le service hydrographique et océanographique (SHOM) dépend du ministère de la Défense.

A l'issue de la seconde guerre mondiale, statistique publique et planification vont de pair : à l'une revient le soin de décrire la population et l'économie du pays, à l'autre celui de guider et d'orienter son développement par une série de grands projets.

L'informatisation de l'administration, réalisée à grande échelle depuis les années 1970, donne un nouvel essor aux registres. En 1974, l'administration ne dispose que de **200 ordinateurs[10].** Les modèles sont très couteux, centralisés, et leur usage est réservé aux fonctions les plus complexes : sécurité sociale, gestion du système de santé. Les registres sont alors devenus des **« systèmes d'information »,** permettant des capacités de traitement, de calcul et de recoupement jusqu'alors inédits. **L'État est le premier,** avec les grands acteurs bancaires, à saisir l'opportunité du développement de l'informatique.

Les données produites par l'État, et plus globalement par l'ensemble des services publics sont de plusieurs natures et sont utilisées pour un grand nombre de finalités. Cette variété de nature et d'usage explique en partie la complexité de la gestion et de l'utilisation de ces données

Qu'est-ce qu'une donnée numérique ?

Une donnée numérique est la description élémentaire de nature numérique, représentée sous forme codée, d'une réalité (chose, évènement, mesure, transaction, etc.) en vue d'être :

- **collecté,** enregistrée,
- **traitée,** manipulée, transformée
- **conservée,** archivée
- **échangée,** diffusée, communiquée.

Selon leur destination, les données peuvent être fermées (réservées à quelques personnes ou à des organisations), partagées (sous réserve de respecter des contraintes contractuelles – licences spécifiques – ou des conditions générales d'utilisation) ou ouvertes (ouvertes à tous utilisateurs et tous usages légaux)[9].

Une information est un ensemble de données agrégées en vue d'une utilisation par l'homme. Pour être utiles, les données doivent le plus souvent être accompagnées de **métadonnées** (littéralement des données sur les données) qui permettent de les décrire le plus finement possible (origine, mode de production, destinations, règles juridiques d'utilisation, etc.).

[8] *Arrêté du Conseil du Roi, 1773*
[9] *Rapport Nora Minc sur l'informatisation de la société, 1978*
[10] *Source : Data Spectrum, Open Data Institute (https://theodi.org/data-spectrum)*

et la difficulté de définir une gouvernance simple s'appliquant à des données et des informations aussi variées (information statistique, météorologique, géographique, budgétaire, administrative, informations issues des grands systèmes de gestion, informations coproduites avec des entreprises, des chercheurs voire des citoyens, etc.). L'objectif n'est pas ici de classifier parfaitement toutes les données produites par l'État, mais uniquement de sensibiliser les différents décideurs à l'ampleur et à la complexité du sujet.

Il est courant d'organiser les données selon trois grandes catégories, à savoir les données permettant:

- **d'identifier ou nommer des personnes ou des choses :** des individus dans le système d'information des ressources humaines (SIRH), des équipements (identification des véhicules), des infrastructures (identification des routes), des personnes (l'état civil des personnes, le répertoire national d'identification des personnes physiques), des organisations (le répertoire des unités légales), des règles (la loi est structurée et organisée pour pouvoir identifier telle ou telle partie), des évènements, des motivations ou encore des budgets (la loi organique loi de finances - LOLF). Cette identification joue un rôle fondamental dans toute la société. Et l'État y joue un rôle central, a minima dans la définition des règles ou des standards applicables, ou pour lui-même identifier les choses ou les personnes (ex. l'état civil),
- **de décrire, de caractériser des choses pour pouvoir les utiliser, interagir avec elles, les étudier, etc.** il sera question ici, par exemple, de données pour décrire une entreprise : son activité, ses implantations géographiques, sa taille, son chiffre d'affaire, ses relations avec l'administration, ses obligations légales, etc.
- **ou encore de décider, de conduire, de piloter…** de faire des choix[11].

Les grands producteurs de données publiques

Pour faire face à ses missions, l'État a mis en place des opérateurs dédiés à la production de données. L'INSEE, l'IGN, Météo France ou encore l'INSERM, l'INED, l'ONEMA sont réputés pour leur savoir-faire et la qualité de leur production.

L'institut national de la statistique et des études économiques (INSEE) produit, analyse et diffuse des informations statistiques sur l'économie, la société et les territoires français. Ces informations relèvent des domaines macroéconomique, sectoriel, démographique et social. L'INSEE a notamment en charge la tenue des registres d'état civil, de répertoires des entreprises (Sirene, Système informatisé du répertoire des entreprises et des établissements, utilisé pour l'identification des entreprises) et le recensement annuel de la population.

L'Institut géographique national (IGN) a pour mission, depuis l'immédiat après-guerre, de cartographier la France et ses territoires. Il n'aura de cesse d'adapter sa production au défi de la numérisation, notamment par la production du référentiel à grande échelle ou la mise à disposition du GéoPortail. L'**Institut national de la santé et de la recherche médicale** (INSERM) est le seul organisme public de recherche français entièrement dédié à la santé humaine et ses chercheurs produisent de très grands volumes de données, notamment en matière d'épidémiologie.

La réputation de ces grands producteurs est établie au niveau international et ils participent aux travaux de normalisation et standardisation européens et mondiaux. L'INSEE représente ainsi la France auprès du Système Statistique Européen et le directeur de l'IGN fait de même au sein du comité d'experts des Nations Unies sur la gestion de l'information géographique (UN-GGIM).

De nouvelles sources et de nouveaux modes de production de données à mobiliser

Comme les entreprises et les citoyens, l'État constate aujourd'hui en son sein et dans son environnement l'irruption de quantités considérables de données d'un nouveau genre.

On ne retient trop souvent de la révolution en cours que l'**explosion des volumes** des données produites[12]. Pourtant, la diversité des sources et des données produites est probablement un phénomène encore plus marquant.

[11] On peut aussi classer les usages des données selon trois axes. Le premier correspond à l'identification et la description des usagers. Le deuxième regroupe l'ensemble des données nécessaires à l'exécution des services et missions de l'État (défense, justice, éducation, santé, travail, etc.). Le dernier axe a trait aux ressources que les services publics utilisent pour leurs missions, que ce soient des actifs matériels (mobilier, équipement, locaux) et immatériels, et des personnes (agents, partenaires, prestataires). Enfin, il faut distinguer les données de flux et les données permanentes.

[12] 2,5 trillions d'octets sont ainsi produits quotidiennement dans le monde et 90 % du stock de données existants a été produit au cours des deux dernières années selon IBM.

« *Google en sait plus que l'INSEE sur la France* » affirmaient en 2013 deux chercheurs français en informatique. La formule, largement reprise, fait mouche. Elle sera même citée en préambule de la session du **Conseil national de l'Information statistique** consacré à l'usage du big data* pour les statistiques. La formule masque pourtant la complémentarité entre les sources. L'INSEE produit les statistiques par des méthodes scientifiques robustes, de manière objective et transparente. Les statistiques publiques présentent une cohérence interne et dans le temps. Elles permettent les comparaisons entre régions et le plus souvent aussi entre pays.

Les grands acteurs du numérique, pour leur part, collectent des données de **manière non scientifiquement contrôlée,** grâce à des capteurs, au recueil de traces d'utilisation, ou encore à la contribution des internautes. Elles n'ont ni la robustesse, ni la complétude des données scientifiques. Elles créent cependant progressivement, de par leur seul volume, une forme d'**empreinte du réel** qui peut à son tour être interprétée et être utilisée produire un savoir activable.

Pour utiliser ces données, le statisticien doit apprendre à faire face à une problématique nouvelle. Il doit renverser la logique habituelle qui consiste à adapter la collecte des données au niveau de précision souhaité, pour, désormais, déterminer le degré de confiance que l'on peut accorder à des données dont le but premier n'était pas l'information statistique. La société Waze, qui propose un navigateur GPS repérant les embouteillages, a créé les cartes de nombreuses villes de pays émergents en analysant simplement les déplacements des téléphones portables de ses utilisateurs. Apple recueille de nombreuses données biométriques, comme le font de nombreuses entreprises offrant des appareils d'aide aux sportifs. Les opérateurs téléphoniques, les entreprises de e-commerce et de nombreux acteurs du numérique commencent à construire une connaissance fine de l'économie et de la société.

Les données produites par les administrations et les données produites en externe ne sont pas antagonistes, elles peuvent même se compléter. L'INSEE développe ainsi un projet de suivi de l'indice des prix à la consommation basé sur des tickets de caisse anonymes. De nombreux chercheurs ont montré que, soigneusement analysées, des données issues du web social pouvaient apporter des éclairages essentiels à des questions extrêmement difficiles – ou coûteuses à analyser – sur le seul fondement de données d'enquête.

Dernière dimension - et non la moindre - des changements en cours : produire une donnée essentielle n'est plus aujourd'hui la seule prérogative de l'État. Les contributeurs bénévoles de l'association OpenStreetMap cartographient le pays à grande vitesse. La **co-production des données essentielles** avec la multitude n'est pas un scénario futuriste : le projet de **Base adresse nationale**[13] est le fruit d'une collaboration entre l'Institut national géographique, La Poste, l'Administrateur général des données et l'association OpenStreetMap France.

Nouvelles sources donc, mais aussi nouvelles manières de produire des données. **L'État n'est pas absent de cette révolution :** les systèmes de gestion qu'il utilise produisent aussi de la donnée sous forme de traces qui peuvent être maintenant mobilisées, ses agents peuvent embarquer sur le territoire des terminaux légers permettant de recueillir de multiples informations, et il démontre progressivement sa capacité à entrer dans l'univers de création des biens communs.

Quatre exemples d'utilisations de nouvelles sources de données pour les politiques publiques

Pour détecter les effets secondaires inconnus des médicaments

Aux États-Unis, la Food and Drug Administration (FDA) vient de passer un accord avec Google pour repérer les effets secondaires inconnus des médicaments. Les requêtes anonymisées des utilisateurs du moteur de recherche permettent notamment de repérer des effets secondaires qui apparaissent tardivement après le début de traitement et qui sont aujourd'hui parfois sous-estimés par les dispositifs actuels de pharmacovigilance.

Pour surveiller la propagation des épidémies (malaria, dengue, virus ebola)

Les données issues de Twitter et de Google sont utilisées dans de nombreux pays (dont le Brésil et Singapour) pour surveiller la propagation des maladies transmissibles comme la dengue. Le service HealthMap, qui analyse en continu des milliers de sources de données a pu repérer le déclenchement de l'épidémie d'Ebola près d'une semaine avant que l'alerte ne soit officiellement déclenchée par les pays concernés.

Pour améliorer l'offre de transports dans une ville

Dans le cadre du programme Data4Development, l'opérateur Orange a mis à disposition de la communauté scientifique des données rendues anonymes et en particulier la localisation des utilisateurs de téléphone mobile en Côte d'Ivoire et au Sénégal, utilisées notamment pour en déduire les flux origine-destination au sein de la ville. Elles sont alors converties en parcours au niveau du réseau de transport existant. Une équipe d'IBM a ainsi permis d'améliorer le réseau de transport de la ville d'Abidjan, de manière à augmenter le nombre de lignes et la satisfaction des usagers, à la fois en termes de parcours et de temps d'attente.

Pour repérer les crises économiques et alimentaires

Le programme Global Pulse du secrétariat général des Nations-Unies analyse les données de Twitter pour repérer l'évolution des opinions dans chaque pays. En procédant ainsi, ils ont été en mesure de détecter en temps réel des crises alimentaires liées à l'explosion du prix de certaines matières premières. L'analyse ne remplace pas la mesure officielle de l'inflation, mais elle la complète en offrant une vue en temps réel.

[13] *Voir les données sur le site : http://adresse.data.gouv.fr/*

2. L'ÉTAT, UTILISATEUR DE DONNÉES

Si l'État est producteur de données, c'est d'abord parce qu'il les utilise lui-même. Là encore, une longue histoire de l'action publique fondée sur l'utilisation des données est aujourd'hui bousculée par de nouvelles logiques d'action issues de la révolution numérique. Les « stratégies fondées sur les données » représentent de nouvelles opportunités pour qui sait **transformer les données en actions** (« data to action »), et les utiliser comme **outil de régulation.** La production ou l'utilisation de données ouvertes, support de nouvelles logiques d'action, élargissent la palette d'options stratégiques disponibles pour l'État.

Les données au service de la politique du logement

42 milliards d'euros d'argent public sont consacrés annuellement au logement. La somme consacrée à produire des données, notamment statistiques, permettant de piloter ces politiques publiques est d'environ 30 millions d'euros[15]. Cette proportion traduit bien la difficulté à disposer de données dans un champ où, de plus, les collectivités locales jouent un rôle de plus en plus important.

Le débat thématique sur l'ouverture des données du logement, mené conjointement par le Conseil national de l'Habitat et la mission Etalab a permis de constater que :

- les données utiles n'existent pas toujours, faute notamment d'une coordination entre de multiples acteurs intervenants et compte tenu de la complexité des sujets traités,
- les données existantes ne sont pas toujours disponibles au niveau le plus fin, ce qui ne permet pas de prendre en compte la diversité des réalités locales,

Il apparait ainsi urgent de mieux organiser la collecte, la remontée et le partage des données du logement, et d'en garantir la qualité et la facilité de réutilisation[16].

Organiser l'État, organiser la société, des usages traditionnels en pleine évolution

Les données sont nécessaires au fonctionnement quotidien de l'État et des services publics

Pour **remplir ses missions de manière efficiente,** l'État doit mobiliser un ensemble de données de qualité et à jour. Elles interviennent à chaque stade de l'action publique : le diagnostic, la programmation, la mise en œuvre et l'évaluation. Ces usages sont innombrables, allant de la gestion des effectifs en milieu scolaire à la préparation d'une réforme fiscale en passant par la planification de travaux ou de nombreuses décisions d'investissement.

L'État, comme la plupart des grandes organisations, a massivement automatisé et optimisé un grand nombre de processus grâce à l'informatique et manipule donc, de ce fait, un nombre de données en croissance accélérée. Aucune administration n'y a échappé. Cette informatique gère au quotidien des données nécessaires au fonctionnement des administrations, comme par exemple les données sur l'identité des personnes et sur les véhicules qui sont utilisées au quotidien par les forces de l'ordre.

La réutilisation en masse de fichiers administratifs est pratiquée de longue date par les administrations dont la finalité première est de mettre à disposition du public des informations. Ainsi les statisticiens publics font appel de manière croissante aux fichiers de données détenus par les administrations sociales ou fiscales pour produire des statistiques relatives à l'emploi, l'activité des entreprises, les revenus. Les fichiers administratifs évitent, en effet, d'avoir recours à des enquêtes, coûteuses pour les répondants comme pour les services enquêteurs. Ils permettent aussi de répondre à la demande croissante de données à des niveaux géographiques ou de nomenclature détaillés, notamment en appui aux politiques publiques.

Ainsi, les informations finement localisées dont dispose l'INSEE dans de nombreux domaines, en particulier en matière d'équipements ou de revenus (ceux-ci sont diffusés depuis plusieurs années au « carreau » de 200m de côté), lui permettent d'apporter un éclairage quantitatif très précis en appui aux politiques et aux administrations concernées. A titre d'exemple, on peut citer l'apport essentiel des données de revenus fiscaux localisés à la conception de la réforme de la géographie prioritaire de la politique de la ville, puis au suivi dans la durée de cette politique. La mobilisation par l'INSEE d'une méthodologie innovante fondée sur des données carroyées a, en effet, permis l'identification des nouveaux quartiers prioritaires sur la base exclusive du critère de revenu des habitants. Cette approche a été traduite dans la loi, et ses deux décrets d'application[14]. Ces quartiers ont vocation à être suivi dans le temps, au moyen d'un ensemble d'indicateurs statistiques.

Les nouveaux défis auxquels est confrontée la puissance publique - par exemple la sécurité, l'aménagement urbain ou encore la transition énergétique - sont en effet de plus en plus complexes et font intervenir un nombre important d'acteurs différents : services de l'État, collectivités, entreprises et acteurs associatifs. Le partage des données doit suivre ce mouvement de travail inter-administration.

[14] Voir notamment la loi de programmation pour la ville du 21 février 2014 et le Rapport d'activité 2013 de l'INSEE, pages 20 à 22
[15] Inspection générale de l'INSEE N°1.7.25 - Conseil général de l'Environnement et du Développement Durable N°009075-02 « Rapport sur l'organisation du service statistique dans le domaine du logement » - http://www.ladocumentationfrancaise.fr/var/storage/rapports-publics/144000532/0000.pdf
[16] Conseil national de l'habitat, Etalab (2015) : ouverture des données publiques dans le champ du logement, synthèse des débats

Les données de référence sont essentielles pour le fonctionnement de la société

De même, l'État produit de longue date des **données de référence** (des données utilisées par un grand nombre d'acteurs, qui y recourent fréquemment, comme par exemple le code officiel géographique, le répertoire SIRENE, le cadastre, etc.). De nombreuses activités économiques et sociales reposent sur la qualité et la disponibilité de ces données, qu'il s'agisse de nomenclatures, de référentiels ou de données essentielles.

Avec la révolution numérique, et l'intensification du recours aux données par les nouveaux services, de nouvelles données de référence se feront jour. La géolocalisation précise des bâtiments est par exemple devenue essentielle pour de nouveaux services. Parallèlement, avec la baisse du coût de production de ces données, la capacité de l'État à **définir un référentiel** est en partie remise en cause. Ce qui fait référence c'est ce qui est **reconnu comme tel** par les utilisateurs, et non ce qui est défini comme tel de manière unilatérale.

Dans ce monde de **standards de fait** (et non plus uniquement de normes), préserver la capacité d'agir de la France c'est, par exemple, faire en sorte que l'identifiant des entreprises reste le numéro SIREN, fourni par l'INSEE, et non un identifiant attribué par un acteur tiers, par exemple une société d'informations financières. Fournir la donnée « officielle » ne suffit plus : il faut qu'elle soit de qualité, complète, à jour et mise à disposition 24/7 via des API* avec un haut niveau de qualité de service. De très nombreux référentiels clés sont, en quelques années, devenus obsolètes. Il y a vingt ans encore, le savoir de l'humanité était classé selon les typologies définies par la Library of Congress (Dewey) et celle définie par la BNF (Rameau). Ces deux typologies se sont affrontées brièvement lorsque Yahoo ! a tenté d'indexer le web selon une logique similaire. L'approche de Google, utilisant des algorithmes fondés sur les liens hypertexte définis par les utilisateurs du web a modifié ce raisonnement : les référentiels ne sont pas devenus obsolètes comme moyens de classer et de hiérarchiser, mais seulement comme moyen de retrouver un document.

La donnée est, en effet, un **actif stratégique** dont la valeur est liée à la réutilisation plus encore qu'à l'utilisation première. L'exemple du GPS est à cet égard éclairant. Ce système de localisation par satellite développé par les États-Unis dès la fin des années 1970 est devenu opérationnel en 1995. Son usage était initialement réservé à l'armée américaine, puis fut étendu progressivement aux usages civils, notamment sur décision du président Clinton. Aujourd'hui, le GPS est devenu la plateforme essentielle au fonctionnement de nombre d'industries, de l'aviation à l'agriculture en passant par les transports. L'Europe, la Chine et la Russie s'efforcent de déployer leurs propres réseaux de satellites pour que leurs économies ne dépendent pas uniquement de cette infrastructure entièrement contrôlée par une seule grande puissance.

Les nouvelles stratégies d'action

« Le logiciel dévore le monde », comme aiment à le rappeler les acteurs de la révolution numérique citant la célèbre tribune de Marc Andreesen[18]. Les entreprises, les personnalités, les stratégies et les outils qui ont fait la révolution numérique bouleversent en effet de nombreux champs de l'action humaine. La puissance publique, pour remplir ses missions et maîtriser ses coûts, doit s'approprier à son tour ces outils, ces méthodes et ces stratégies.

L'exemple du référentiel cadastral

Une abondante littérature internationale a souligné l'impact économique d'un référentiel clé très familier en France : le cadastre.

Sans cadastre, la puissance publique se trouve très démunie pour prélever l'impôt et sécuriser les droits de propriété, comme le montre l'incapacité de la Grèce à prélever l'impôt foncier, difficultés parfois invoquées comme déterminant de la crise de la dette publique grecque.

Sans cadastre, par ailleurs, il est quasiment impossible aux propriétaires fonciers de mobiliser leur propriété pour emprunter et donc pour investir. De nombreuses analyses montrent que l'absence de système cadastral est un frein à l'émergence d'une classe moyenne dans les pays en voie de développement[17].

Les usages ont aussi fortement évolué : autrefois exclusivement destiné à la levée de l'impôt par l'État, le cadastre moderne sert maintenant aux collectivités locales (aménagement et urbanisme, gestion des infrastructures) mais aussi aux entreprises (sécurisation des transactions immobilières, support du crédit hypothécaire, etc.).

Le cadastre n'est donc plus seulement un outil pour un usage précis (lever l'impôt foncier) il est devenu un élément stratégique, un standard sur lequel de nombreux acteurs se synchronisent afin de pouvoir échanger. En cela, le cadastre est une donnée de référence qui rend possible la coordination entre acteurs, qui leur permet de conjuguer leurs forces et d'atteindre le meilleur équilibre possible.

La donnée de référence possède des similarités avec la monnaie, qui est produite et garantie par l'État, pour permettre des échanges entre acteurs et le bon fonctionnement de l'économie.

[17] *De Soto H. (2005) : Le mystère du capital : « Le mystère du capital : pourquoi le capitalisme triomphe en Occident et échoue partout ailleurs ?, trad. française Flammarion*
[18] *Andreessen M. (2011) : Why software is eating the world, The Wall Street Journal*

Les datasciences : de nouveaux champs d'action

Qu'est-ce que les datasciences ?

Le terme de datascientist a été forgé par Jeff Hammerbacher (Facebook) et DJ Patil (LinkedIn et désormais Chief data scientist à la Maison Blanche) en 2008. Il désigne des personnes qui analysent les données, non pas pour produire des rapports ou des statistiques, mais pour améliorer le produit ou le service de l'organisation pour laquelle ils travaillent. Les datascientists ont donc à la fois la capacité d'analyser les données, la capacité de développer du code informatique et la capacité d'imaginer de nouveaux usages.

Par exemple, en 2006, Jonathan Goldman, fraîchement embauché chez LinkedIn, remarque qu'il est capable de prédire le réseau d'un utilisateur. Il imagine alors le module « People you may know » et le teste dans l'interface de LinkedIn. Ce modèle a rencontré un grand succès et a joué un rôle important dans le développement du réseau social. De même, le Français Paul Duan, alors chez Eventbrite, a développé des approches extrêmement originales et efficaces de détection de risque de fraude, fondées sur des modèles algorithmiques.

Ainsi, les grandes entreprises du numérique ont développé de nouveaux usages des méthodes statistiques et des algorithmes d'apprentissage statistique. Facebook utilise les données des utilisateurs pour prédire les amis de l'utilisateur, LinkedIn pour prédire les contacts professionnels, Netflix pour prédire les films que l'utilisateur aime et Amazon pour prédire les produits qu'il est susceptible d'acheter.

Les datasciences utilisent donc l'ensemble des méthodes des statistiques et du « machine learning* », la régression linéaire, la régression logistique, les arbres de décisions, les forêts aléatoires, les algorithmes de segmentation et l'ensemble des méthodes de visualisation de données[19] pour concevoir de nouvelles applications[20].

Les datasciences permettent donc notamment d'assister les administrations dans leur **prise de décision.** Pour cela, il faut disposer des données pertinentes, mais aussi que les solutions apportées soient **activables** et **mesurables :**

- **activables,** en travaillant sur des questions ayant une **application opérationnelle concrète :** il ne s'agit pas d'illustrer, d'observer ou même de comprendre en tant que tel mais bien d'utiliser les données pour aider une prise de décision. Dois-je acheter ce produit et à quel prix ? Comment planifier les ressources pour faire face aux demandes ? Par où commencer ?
- **mesurable** en s'organisant pour travailler sur des décisions dont l'impact est quantifiable et pour faire en sorte que cette mesure nourrisse à son tour (« éduque ») l'algorithme utilisé.

Avec les datasciences, les données ne servent donc pas seulement à **décrire le réel,** ni même à nourrir une décision : elles entrent de plain-pied dans les processus d'action.

Les évolutions qui se passent à cette frontière de la donnée et de l'action sont sans **doute la dimension essentielle de la révolution des datasciences.** Pour en saisir la portée, il faut réaliser à quel point ces nouvelles capacités élargissent la palette des actions qui s'offrent au décideur.

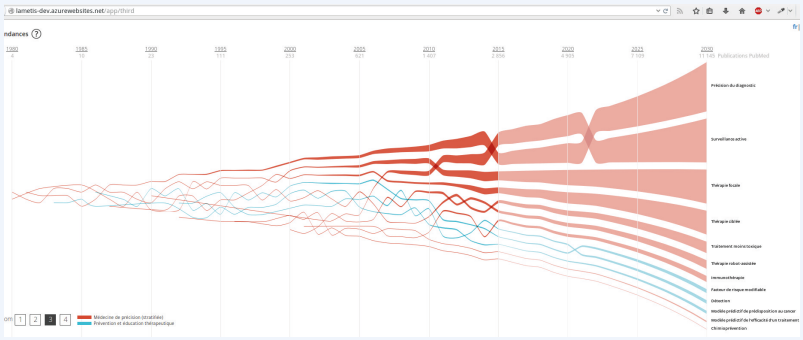
Un simple exemple. Dans un régime d'action classique, le statisticien essaye d'identifier des causalités vérifiées pour intervenir dans un processus linéaire. Si, par exemple, il peut prouver que la vitesse est réellement cause d'accidents de la route, limiter la vitesse produira automatiquement une baisse des accidents de la route. Dans ce régime d'action, il est essentiel de bien comprendre que « corrélation ne signifie pas causalité », car, dans le cas contraire, on agit sur des facteurs qui ne sont pas de véritables déterminants. Avec la capacité de traiter des masses de données en temps réel, et donc de mesurer quotidiennement les effets d'une décision, il devient moins important de séparer finement les corrélations des causalités. Il est possible en effet de tester au quotidien l'efficacité d'une action et de la modifier dès qu'elle ne semble plus produire les effets désirés. Cette approche n'est pas sans poser d'importantes questions épistémologiques et parfois éthiques, qui ont commencés à être discutés dans un célèbre et provocateur article[21] de Wired en 2008. Toujours est-il que, dans l'ordre de l'action, elle *fonctionne*.

Cette focalisation sur le « ***data-to-action*** » explique un certain nombre de stratégies observées dans les villes nord-américaines. La Ville de New-York a ainsi identifié les immeubles à risque d'incendie pour aider les pompiers dans leur action de prévention (passant ainsi en quelques semaines, de 10 % de contrôles positifs à 78 % de contrôles positifs) ou encore aidé à la prise en charge des conséquences de l'ouragan Sandy avec des méthodes inspirées de cette approche probabiliste[22].

[19] Press G. (2013) : A very short history of data sciences, Forbes.com
[20] Davenport T., Patil DJ (2012) : Data Scientist, the sexiest job of the 21st century, Harvard Business Review
[21] Anderson C. (2008) : The end of theory, the data deluge makes the scientific method obsolete, Wired Magazine
[22] Flowers M. (2013) : NYC by the numbers, annual report to the mayor of New York

Un exemple de Big Data Santé qui n'utilise pas de données personnelles

Le développement des datasciences c'est aussi le développement de l'analyse de nouveaux types de corpus, comme des images ou des textes. Par exemple, pour prédire les technologies dans la prise en charge du cancer de la prostate et imaginer l'hôpital de demain, le cabinet La Métis, appuyé par l'Administrateur général des données, a analysé l'ensemble des publications de la bibliothèque américaine de médecine (Medpub) librement disponibles et réutilisables sur le sujet depuis 1980. Cette analyse permet d'identifier des courbes de diffusion des innovations et de prédire les volumes de publications sur chacune des innovations identifiées pour les 15 prochaines années, ce dont il est possible d'inférer des tendances en matière de pratiques, et donc des recommandations en matières d'investissements hospitaliers.



Lecture : L'analyse montre que la précision du diagnostic et la surveillance active sont les deux grandes tendances dans la littérature scientifique sur la prise en charge du cancer de la prostate.

La visualisation[23] de données développée par La Métis avec l'Administrateur général des données permet aux décideurs publics d'avoir une meilleure idée des solutions pour détecter, diagnostiquer et de traiter le cancer de la prostate en 2030 et aux praticiens et spécialistes d'avoir une vision globale de la littérature scientifique sur le sujet.

En juillet 2015, le Conseil général de l'Économie, de l'Industrie, de l'Énergie et des Technologies (CGEIET) a remis au Ministre de l'Économie, de l'Industrie et du Numérique, à la Secrétaire d'État chargée de la réforme de l'État et de la simplification et à la Secrétaire d'État chargée du Numérique un rapport sur **« Les meilleures pratiques du big data et de l'analytique dans l'administration ».** Sous l'entrée **« big data* »,** ce rapport développe largement les perspectives importantes en matière de nouveaux usages de la donnée.

Il documente en particulier l'émergence de bonnes pratiques, par exemple en matière de lutte contre la fraude, de l'archivage électronique, de recueil de données non structurées au service de la statistique publique. Il montre ainsi qu'il existe dans l'administration centrale, déconcentrée et territoriale un ensemble potentiel d'utilisations des données et des pionniers qui pourraient servir de ferment à cette révolution des politiques publiques s'ils étaient soutenus (avec parfois de réels besoins en financements et en effectifs), accompagnés, et mis en réseau. Il propose donc d'organiser une expérimentation approfondie et ciblée sur quelques politiques publiques (identifiées par exemple grâce à la LOLF). La mission propose en outre, une méthode d'évaluation de l'expérimentation après un temps limité de l'ordre de deux années. Le rapport propose aussi l'animation d'une communauté du big data ou du traitement de la donnée ainsi que l'élaboration d'un vademecum juridique à disposition des ministères.

[23] *https://agd.data.gouv.fr/2015/11/25/cancer-de-la-prostate-a-quoi-ressemblera-le-parcours-de-soin-en-2030/*

Rapport du CGEIET « Meilleures pratiques pour le big data et l'analytique dans l'administration : une nouvelle étape »

Dans un contexte fortement évolutif et innovant, la mission a retenu une définition large du « big data » reposant sur une nouvelle exigence de valorisation des données, qu'elles soient internes ou externes. La mission a procédé à un état des lieux auprès des responsables de programme LOLF et bénéficié de la coopération d'un groupe de responsables de projets venus de différents ministères.

Les nombreuses réponses aux questionnaires complétées par des entretiens montrent un intérêt et une appétence pour des usages innovants des données. Par ailleurs, des réalisations pionnières permettent d'identifier des bonnes pratiques au sein de l'État. Les réalisations existantes sont dans le cœur de métier avec des ressources (modérées) et dédiées.

Les échanges de données entre administrations sont peu développés et il n'y a pas de cadre de référence pour le faire. Il existe au moins une expérience d'utilisation de données externes (privées) en substitut de données internes qui pose des questions spécifiques. D'autres aspects des « big data » devraient être explorés plus précisément, notamment une meilleure représentation visuelle plus compréhensible des résultats pour permettre des décisions ou le partage de la connaissance. Un champ très important (criticité, volume, …) est en train de s'ouvrir avec les objets connectés.

Il existe un gisement de problématiques important au sein des services déconcentrés de l'État qui sont un ferment fort pour des opportunités de mise en œuvre de « big data ». Cependant, la révélation du potentiel du « big data » nécessite de réunir des ressources et un contexte fonctionnel adapté, dans les domaines juridique, managérial, technologique et culturel.

Le retour en termes de valeur semble assuré dans le domaine de la lutte contre la fraude. Dans d'autres domaines, le retour en termes de valeur est d'une autre nature : par exemple des externalités positives, un meilleur service au citoyen ou une meilleure efficacité des politiques publiques. L'utilisation des données tierces dans les domaines fiscaux ou grâce à « Dites-le nous une fois » a généré ou a vocation à générer des économies de fonctionnement, tant pour l'administration que pour les administrés. Les processus de l'administration qui reposent sur un ciblage ou un criblage sont également des candidats potentiels pour démontrer l'utilité économique du « big data » tant pour la société que pour l'État.

L'État a donc intérêt à s'approprier la culture « big data » dans des situations variées (production de connaissance, optimisation de processus, services rendus aux usagers, …) en s'appuyant tant sur ses données par un décloisonnement maîtrisé qu'en ayant recours, dans des cadres sécurisés, à des données de tiers.

L'apprentissage à partir de cette diversité nécessite une gouvernance et de la souplesse.

Enfin, il faut souligner que le terme « données » ou « data », conformément à l'usage, désigne tant des données structurées que des informations non structurées.

En conclusion, la mission recommande de s'intéresser en priorité aux modalités d'échanges de données entre ministères par l'établissement d'un référentiel adapté ; elle recommande également de mener une opération de consolidation de l'usage des « big data » sur deux ans, puis d'en faire une évaluation avant de l'étendre. Cette étape de deux ans permettrait d'identifier une communauté ouverte de décideurs et professionnels de l'administration du « big data ».

L'ouverture et le partage de données pour créer de la valeur économique et sociale

L'ouverture des données publiques a longtemps été analysée en France comme une conséquence du droit à l'information des citoyens, ce qu'elle est en partie, indéniablement. Mais il est important de souligner qu'elle peut également représenter une forme extrêmement efficace d'action publique. En partageant des données sur www.data.gouv.fr, l'État et les collectivités en encouragent la **réutilisation** par l'ensemble de la société et peuvent ainsi stimuler l'innovation, faire entrer l'État dans une démarche d'innovation ouverte, voire corriger des imperfections de marché ou même améliorer les données publiques.

La mise à disposition de données libres et ouvertes, et leur réutilisation, génèrent de la valeur économique et sociale, par le biais de cinq mécanismes générateurs de valeur[24] : la réduction des coûts de transaction, l'innovation, la réduction des asymétries d'information, la collaboration et les boucles de rétroaction. Ces mécanismes, explicités ci-après, ne sont pas exclusifs, ils peuvent se combiner pour une même donnée.

L'efficience : réduire les coûts de transaction

L'ouverture des données permet une **meilleure utilisation** des ressources disponibles par les acteurs publics et privés. La théorie des coûts de transaction postule que toute transaction économique engendre des coûts (coût de recherche d'information notamment). En mettant à disposition librement et gratuitement les données publiques, on réduit ces **coûts de transaction**, tant dans leur phase amont que dans la transaction elle-même. La mise à disposition des données est source d'efficience et d'efficacité, tant pour les administrations que pour les acteurs privés. Plusieurs expériences, en France et à l'étranger confirment ce mécanisme de création de valeur. En **Australie,** les coûts de transaction induits par la vente et la distribution des données géographiques australiennes ont été évalués, avant leur mise à disposition libre et gratuite en 2002, entre 17 % et 33 % des revenus. Le gain annuel de cette ouverture a été évalué à 1,7 million de dollars par an pour la seule réduction des coûts de transaction[25]. Au Danemark, le gouvernement a lancé un programme nommé **« Basic Data ».** Il s'agit de mettre en place une infrastructure informationnelle libre et gratuite autour de trois bases de données de référence dont le registre des entreprises et les données géographiques essentielles. Les gains de ce projet sont estimés à **35 millions d'euros** annuels pour le secteur public (meilleure efficience) et 70 millions d'euros pour le secteur privé (production de nouveaux services)[26].

L'innovation et la transformation

Le second mécanisme de création de valeur est lié à l'utilisation, par les secteurs public et privé, des données ouvertes pour créer de nouveaux produits et services. La capacité globale d'**innovation** joue un rôle économique déterminant. Elle permet non seulement la croissance mais suscite aussi d'importantes modifications structurelles.

Les données ouvertes sont un facteur d'innovation pour ceux qui les réutilisent. **Aux Pays-Bas,** l'ouverture des données météorologiques a permis la création d'un **écosystème de ré-utilisateurs professionnels** très dynamique : le revenu des acteurs privés a augmenté de 400 %, le nombre d'utilisations de ces données de 300 %. Ces activités ont généré un retour de 35 millions d'euros pour les finances publiques néerlandaises, sous la forme d'impôts et de taxes additionnels.

Plusieurs études européennes montrent que la baisse d'une redevance ou sa suppression entraînent mécaniquement une augmentation de la réutilisation des données concernées[27]. Par exemple, le passage à la gratuité du référentiel à grande échelle de l'IGN pour les organismes chargés d'une mission de service public administrative, a entraîné une multiplication par 20 des volumes de données téléchargées, soit un **bénéfice social estimé à 114 M€/an,** pour un manque à gagner de 6 M€/an de redevances environ[28].

En France, **le portail data.gouv.fr** accueille et anime une communauté de plus de 600 organisations, dont la moitié de services publics et plus de 10 000 utilisateurs, qui ont publié un total de 90 000 ressources représentant 21 000 jeux de données, et plus de 1 300 réutilisations. De plus, le concours Dataconnexions, organisé par Etalab depuis 2012, a permis d'identifier 200 startups à fort potentiel. Ce dispositif permet de donner une première impulsion à ces projets en contribuant à leur croissance et consolidation. C'est le cas, par exemple de Snips, projet lauréat de la 3ème édition du concours, qui, en quelques mois, est devenu une entreprise de 35 salariés.

[24] Jetzek T.,Avital, M. (2013) : The Generative Mechanisms Of Open Government Data, ECIS 2013 Proceedings.
[25] De Vries M. (2012) : Re-use of public sector information, report for Danish Ministry for Housing, Urban and Rural Affairs.
[26] Ministère des finances du Danemark (2012) : Good Basic Data for Everyone – A Driver for Growth and Efficiency.
[27] Vickery G. (2010) : Review of Recent Studies on PSI Re-Use and Related Market Developments.
[28] Trojette A. (2013) : Ouverture des données publiques : les exceptions au principe de gratuité sont-elles toutes légitimes ?, rapport au Premier ministre.

La réduction d'asymétrie de l'information

Le troisième mécanisme générateur de valeur est lié à la réduction de l'asymétrie d'information par la transparence. On parle d'asymétrie d'information quand un acteur possède une information plus complète, ou de meilleure qualité, que les autres acteurs participant à une transaction ou une communication. L'**asymétrie d'information** aboutit à des situations non optimales. Les données ouvertes permettent de réduire ces asymétries à plusieurs niveaux. Au niveau macroéconomique, la transparence est un outil de **lutte contre la corruption** reconnu notamment par la Banque mondiale. Au niveau microéconomique, la mise en ligne de données détaillées sur les **marchés publics** permet à tous les acteurs de disposer du même niveau d'information. Les répondants peuvent connaître le dernier attributaire d'un marché public et les conditions du marché, leur permettant ainsi de mieux dimensionner leur réponse. Le nombre et la qualité des réponses s'améliorent, ce qui est aussi une condition d'**efficacité de l'achat public.** En France, le service des achats de l'État et l'Administrateur général des données ont ainsi mené une analyse de la consommation énergétique des bâtiments qui a permis d'identifier des profils de consommation homogènes. Cette analyse, ainsi que les données sous-jacentes, ont été mise à disposition des fournisseurs potentiels d'énergie[29].

Les boucles de rétroaction pour agir sur les comportements

Le partage d'une donnée encourage enfin le mécanisme de **boucle de rétroaction.** Partager une **information en temps réel** sur l'état d'un système permet à ses acteurs de modifier leurs comportements, de constater l'effet de cette modification de comportement et de l'ajuster dynamiquement en la renforçant ou en l'atténuant. Les panneaux indicateurs de vitesse, que l'on installe à l'entrée des villes et sur certains tronçons routiers fonctionnent selon ce principe. Leur présence permet de réduire d'environ 10 % la vitesse moyenne, de manière durable. De la même manière, indiquer à des automobilistes une information prédictive à une heure sur les embouteillages attendus sur un axe routier permet de les contourner et, *in fine* d'en diminuer l'intensité. Les boucles de rétroaction ont de multiples applications pour l'action publique. Le **département du travail** américain publie ainsi chaque trimestre depuis 2010 la liste des 500 entreprises les plus récalcitrantes à appliquer la réglementation en matière de **sécurité au travail**[30]. Ne pas figurer sur cette liste représente un enjeu majeur pour les employeurs et constitue une incitation forte à mieux protéger leurs salariés. En France, le Ministère de l'Economie, de l'Industrie et du Numérique a fait de même, en publiant en novembre 2015 la liste des 5 grandes entreprises qui se sont vu infliger les plus importantes amendes pour leur politique de paiements tardifs répétés[31].

La collaboration pour produire, enrichir et améliorer des données

La production de données **en mode collaboratif** n'est pas à proprement parler une nouveauté : les **sciences participatives** bénéficient d'une longue tradition dans le domaine de la botanique, de l'observation de la biodiversité[32] ou même de l'astronomie. Le numérique donne un nouvel essor à ces pratiques et élargit leur champ d'action.

La mise à disposition de données en open data crée les conditions d'une collaboration entre de multiples acteurs, tant publics que privés. Cette collaboration autour des données constitue une nouvelle stratégie d'action.

En effet, la collaboration génère des **économies d'échelle.** Ainsi la plateforme data.gouv.fr permet à chacun d'enrichir, d'améliorer et de repartager un jeu de données. Depuis fin 2013, de nombreux exemples **d'enrichissement** ont été documentés. Le fichier des accidents corporels de la circulation a fait l'objet de multiples améliorations par les utilisateurs du fichier : nettoyage, correction des doublons, ajout des codes géographiques (INSEE et codes postaux). De même, les utilisateurs du site ont pu signaler les erreurs aux producteurs et proposer des corrections (signalement d'erreurs de géocodage, d'adresses absentes ou incomplètes, de données manquantes), enclenchant ainsi une **dynamique d'amélioration continue** de la qualité des données.

La collaboration permet d'améliorer la qualité de données existantes. Le mode collaboratif peut être aussi le levier de la production de données. Le projet BAN (Base Adresse Nationale) est ainsi la mise en commun des données de l'IGN, de la Poste et des données produites par les contributeurs d'OpenStreetMap[33].

[29] *Administrateur général des données (2015) : Analyser les consommations énergétiques des bâtiments publics, disponible sur agd.data.gouv.fr*
[30] *« Severe Violator Enforcement Program », US Department of Labor : https://www.osha.gov/dep/*
[31] *http://www.economie.gouv.fr/dgccrf/sanctions-delais-paiement*
[32] *Voir notamment le programme Vigie Nature piloté par le Muséum national d'histoire naturelle (http://vigienature.mnhn.fr)*
[33] *http://adresse.data.gouv.fr*

Cette dynamique positive s'organise : l'utilisateur doit pouvoir accéder à une documentation, il peut même participer à sa rédaction. L'organisation collaborative est aussi vertueuse pour tous les éléments de nettoyage des données. Il est en effet peu efficace que chaque utilisateur s'affaire à redresser des valeurs aberrantes isolément. On aurait tout intérêt, pour chaque jeu de données partagées, à créer un espace de dialogue et d'échange de codes pour que ceux qui veulent jouer le jeu du travail collaboratif puissent le faire.

La régulation par la donnée, une nouvelle forme d'action publique

Les pouvoirs publics sont confrontés à de nouveaux défis : des services déployés en ligne ou depuis un mobile ont maintenant un **impact direct et immédiat** sur les territoires et des filières économiques bien établies, qu'il s'agisse de **mobilité** (Uber, Blablacar) ou d'**hébergement** de courte durée (Airbnb, Bedicasa).

La plupart de ces entreprises n'existerait pas sans les données, tant la place qu'elles occupent est centrale dans leurs modèles d'affaires. Elles fournissent d'abord l'ingrédient indispensable aux échanges : la confiance, par l'analyse des transactions et la notation réciproque des intervenants. La confiance que le client potentiel doit avoir avant d'acheter le service offert par un tiers ne passe pas des certifications, des labels ou des diplômes, mais par l'analyse des transactions et la notation réciproque des intervenants - des données traduites sous forme de notes.

Les données sont également utilisées pour améliorer en permanence le service : Uber est capable de prédire les zones où la demande sera la plus forte à un instant T, et donc d'encourager les chauffeurs à s'y rendre ou de modifier les prix en fonction de l'offre et de la demande. Airbnb analyse en permanence les recherches et l'historique d'un client, et sait donc quel bien il faut lui présenter en premier lieu pour répondre à ses goûts. Il peut aussi aider les hôtes à fixer le meilleur tarif de location (tout au moins celui qui maximise le revenu de la plateforme).

Ces nouvelles activités appellent une **nouvelle forme de régulation.** Les régulateurs, tant nationaux que locaux, prennent progressivement conscience de l'importance des données dans ces modèles d'affaires. De nouvelles formes de régulation apparaissent, que l'on peut qualifier de « régulation par la donné e » ou « régulation 2.0 »[34].

La première forme consiste à échanger des données contre l'**autorisation d'exercer** sur un territoire. La ville de New-York a ainsi assoupli les conditions d'exercice d'Uber en échange des données sur les trajets, les chauffeurs et la demande de mobilité en chaque point de la ville et à chaque instant. Ainsi armée de ces données, la ville peut passer d'un système de contrôle a priori (autorisation ex-ante d'exercer par la délivrance de licences) à une modération a posteriori (maintien de cette autorisation par l'analyse des données).

La mairie de San Francisco cherche à lutter contre la « gentrification » de certains quartiers imputée à Airbnb. Elle a récemment mis en place un bureau dédié à la location de très courte durée. Son objet : encourager les propriétaires à se conformer au droit local qui prévoit que les hôtes ne puissent pas louer leur logement plus de 90 jours par an sans y être présents sur place. Mais, pour contrôler cela, il faudrait pouvoir accéder aux données de l'entreprise, ce qu'Airbnb refuse jusqu'à présent. Ici, les données sont un **instrument de négociation** dans le rapport de force qui s'établit entre ces plateformes et les territoires sur lesquels elles opèrent.

La seconde forme de régulation par la donnée consiste, pour le régulateur, à jouer un rôle actif dans l'émergence des plateformes numériques. En France, l'article 1er de la **loi n° 2014-1104 du 1er octobre 2014** relative aux taxis et aux voitures de transport avec chauffeur prévoit la mise en place d'un **registre géolocalisé des taxis.** L'idée est de permettre aux taxis d'avoir accès à la maraude électronique. Grace à ce registre - et aux applications qui l'utiliseront, il va devenir possible de réserver son taxi immédiatement sur son smartphone, indépendamment de son appartenance à telle ou telle centrale de réservation (« tous les clients peuvent voir tous les taxis »[35]). La plateforme Le.Taxi, développée par le Secrétariat général à la modernisation de l'action publique et le Ministère de l'Intérieur est en cours de déploiement. Elle marque une étape importante en matière de régulation par la donnée. En organisant la circulation d'une donnée (la géolocalisation des véhicules) et en posant des règles fortes (neutralité et gratuité de la plateforme), l'**État adapte son rôle à la révolution numérique.**

[34] Grossman N. (2015) : White Paper : Regulation, the Internet Way. A Data-First Model for Establishing Trust, Safety, and Security | Regulatory Reform for the 21st Century, Mimeo.
[35] Cf. http://le.taxi/

2

Le manque de gouvernance des données comme frein au potentiel des données

Si l'État a, de longue date, organisé son action autour de données scientifiques et administratives, il faut reconnaître que les choix d'organisation, les stratégies technologiques et les règles de gouvernance de ces données qui prévalent aujourd'hui encore résultent de choix organisationnels, de technologies et de cadres juridiques antérieurs à la révolution en cours. Pour des raisons historiques bien compréhensibles, l'État, comme toutes les grandes organisations, s'est plus intéressé à la construction de savoirs incontestables qu'à la diffusion de données exploitables par le plus grand nombre. Focalisé sur la fiabilité, la sécurité et la maîtrise des coûts, il a négligé l'interopérabilité, l'accessibilité et la capacité d'usage, et a donc toléré une culture de silos, des divergences de formats avec des qualités excessives ou au contraire dégradées, une sous-traitance excessive et une perte globale de souveraineté et d'autonomie sur ses propres données.

1. LA MÉCONNAISSANCE DES DONNÉES DISPONIBLES

Nul n'est aujourd'hui en mesure de connaître avec précision l'étendue des données que l'administration publique possède. Cette **méconnaissance** constitue le premier frein à une pleine et entière exploitation du potentiel des données par la puissance publique : ce qui ne se connait pas ne se maîtrise pas.

Cet état de fait peut conduire à une perte d'opportunité car l'on se prive d'informations intéressantes et *de facto* disponibles. Cela est d'autant plus prégnant à l'heure où il est de plus en plus facile de croiser des données et de multiplier le potentiel d'intérêt. Nul n'est en mesure aujourd'hui, de recenser l'ensemble des informations détaillées au niveau des communes et cela restreint les capacités d'analyse des territoires et la possibilité d'observer des corrélations croisées originales.

De plus, la méconnaissance des données possédées par l'administration - ou la difficulté à y accéder- conduit parfois à la duplication du travail de production. C'est ainsi qu'une base de géolocalisation des adresses s'est construite parallèlement à l'INSEE, à la DGFiP et à La Poste.

Ce constat ne doit pas masquer les efforts entrepris, depuis plusieurs années, pour tenter de dresser un tel inventaire.

Dès 1978, l'article 17 de la loi CADA introduit l'obligation de tenir un **répertoire des informations publiques.** Il s'agit de lister en un point unique l'ensemble des documents produits par une administration. Les répertoires aujourd'hui publiés le sont majoritairement dans une optique documentaire : très pertinents pour celui qui recherche une étude ou une statistique, ils le sont beaucoup moins pour identifier les différentes bases de données existantes au sein d'une administration. L'approche d'urbanisation centrée sur les systèmes d'information (« Plan d'Occupation des Sols ») recense les principales applications et les bases correspondantes par périmètre fonctionnel. D'un côté, une approche par le document, de l'autre, une approche par le SI, elles sont toutes les deux nécessaires, mais aucune n'est suffisante.

Pourquoi est-il difficile d'obtenir une vision globale des données que maîtrise l'État ? Une première année d'échanges et de coopérations avec les administrations a permis d'identifier les principales difficultés qui freinent un tel recensement exhaustif :

- **une prise de conscience** variable selon les acteurs. Les grands producteurs ont pour mission de produire des données et se sont organisés en conséquence. Mais nombre d'administrations produisent aujourd'hui, parfois de manière incidente, des données sans les considérer comme telles et sans s'interroger sur leur utilité potentielle pour des tiers ;
- une **difficulté à distinguer la donnée du système qui la produit :** la donnée est parfois si intimement intégrée au système d'**information** qu'il devient très difficile, voire impossible de l'en extraire. Cette problématique, liée à la conception des systèmes d'information sera développée au chapitre suivant. Il est courant, au sein des ministères, de désigner les bases de données par le nom de l'applicatif qui les produit (SIV, PATRIM) ;
- une **très grande diversité** dans la manière dont les données sont stockées et présentées, d'un fichier tabulaire à des bases de données associées à des progiciels métiers. Cette diversité rend aussi parfois plus difficile l'identification des données pertinentes par les producteurs eux-mêmes ;
- une **même source** peut alimenter plusieurs bases de données différentes, gérés par des acteurs différents. Il n'est pas toujours aisé d'établir cette traçabilité de l'origine des données ;
- de **multiples bases** sur le même sujet : il existe autant de bases que d'angles d'approche du sujet et de finalités possibles. La cartographie des données de santé établie par Etalab recense ainsi pas moins d'une demi-douzaine de bases sur le thème du cancer, sans pour autant qu'on puisse évoquer une redondance. Certaines ont trait au suivi des patients à long terme (cohortes), d'autres à la gestion des soins en milieu hospitalier. Chacune a son utilité propre.

De l'importance des données de gestion

Les politiques publiques sont souvent établies à partir des « données d'autorité », ces données employées par les décideurs qui en connaissent l'existence, l'usage, la portée, et éventuellement aussi les risques de mésusage.

Or, du fait de la révolution numérique, la plupart des données existantes sont aujourd'hui produites dans de grands systèmes de gestion informatisés, et ne sont pas connues ni repérées comme telles.

Une histoire connue dans les communautés open data concerne cette grande municipalité qui souhaitait ouvrir son portail d'open data et recherchait dans ce but des données concernant les pratiques culturelles. Il lui fallut près d'un an pour réaliser que l'application de gestion des bibliothèques municipales recelait un trésor : la liste des ouvrages empruntés quotidiennement dessinait une sociologie des pratiques culturelles, permettait de comprendre la saisonnalité des pratiques, d'identifier des corrélations inédites entre types d'ouvrages, de recommander des livres à emprunter, etc.

L'ouverture de ces données soulevait à son tour de nouvelles questions. Ces données en effet, même anonymes, pouvaient éventuellement révéler des phénomènes de communautarisme, dévoiler par accident des informations à caractère personnel. Il s'agissait de données brutes, non utilisées dans les travaux scientifiques ni les communications politiques. Il était nécessaire de se les approprier et d'en analyser la portée exacte avant de pouvoir les partager.

De telles données, issues des grands systèmes de gestion, représentent aujourd'hui un sujet central de la gouvernance de la donnée. Ce sont celles qui posent aujourd'hui le plus de questions ouvertes. Il est essentiel d'apprendre à les « domestiquer » et à les intégrer dans les processus usuels de l'administration.

2. LE SI DE L'ÉTAT N'EST PAS AU SERVICE DE L'USAGE DES DONNÉES

Les choix d'architecture sont antérieurs à la révolution de la donnée

Au cours de l'année écoulée, l'Administrateur général des données a coopéré avec de nombreuses administrations, et a engagé des projets concrets avec plusieurs d'entre elles. Cette approche concrète a révélé combien l'accès aux données pertinentes était difficile pour les administrations elles-mêmes.

Dans la plupart des cas, les difficultés proviennent de choix de conception du système d'information opérés en fonction des contraintes et priorités antérieures au développement des nouveaux usages des données. En particulier, les applications informatiques et leurs bases de données sont souvent conçues et optimisées pour remplir un ou plusieurs objectifs métiers dans des délais contraints, au détriment d'autres objectifs liés à la réutilisation des données. Ces difficultés sont notamment :

- **Une lourde dette technologique,** les données étant produites, ou traitées, dans des systèmes souvent très anciens pour lesquels les compétences de développement se font rares, et donc chères, et dans lesquels la réversibilité des données n'a pas été pensée dès la conception ;
- **Une structuration n'anticipant pas les besoins de partage,** mêlant par exemple les données susceptibles d'être partagées et réutilisées et des données couvertes par des secrets (par exemple quand des données à caractère personnel sont mêlées avec l'ensemble du système), ou omettant des métadonnées essentielles (comme les droits associés aux données), ce qui empêche la fouille et l'extraction de données sur les questions, les secteurs ou les données ne posant pas de problèmes ;
- **Une structuration** organisée autour des besoins des applications de gestion, et omettant la nécessité de prévoir une extraction des données non structurées pour les usages imprévus à venir ;
- **Le manque de référentiels accessibles 24 heures sur 24.** Même lorsque les administrations acceptent de transmettre leurs données, celles-ci sont bien souvent perçues comme des informations (voire des fichiers) à transmettre de façon périodique, solution lourde et ne garantissant pas une fraîcheur optimale des données. Rares sont les administrations qui ont pris le virage des stratégies de plateforme et d'API*, et qui s'organisent pour mettre à disposition leurs données de référence à d'autres applications (à des fins de consultation et synchronisation) en temps réel, 24h/24.

Cette situation n'est ni une spécificité publique ni une spécificité française. Elle tient également à l'histoire de l'informatique, qui vit, elle aussi, une révolution numérique constante.

Il n'y a pas si longtemps encore, l'informatique était conçue sur papier, orientée pour minimiser les investissements d'innovation, et utilisée comme une ressource statique au service de l'organisation. En particulier, sa valeur était rarement estimée (et donc pilotée) en fonction du potentiel de transformation de l'organisation ou de la chaîne de valeur. Cette situation originelle a suscité un portefeuille d'applications fragmentées, avec une dépendance excessive envers les prestataires de services, une opacité croissante des dépenses, et, de ce fait, l'échec retentissant de certains « grands projets informatiques ».

Cet état de fait n'est pas seulement nuisible au bon usage des données dans la conduite des politiques publiques. Il induit une perte globale de capacité d'action de l'État. Il suscite une complexité qui diffuse largement au-delà des enjeux informatiques. Comme l'a souligné Michel Volle[36], les systèmes d'information reflètent tout autant qu'ils modèlent les processus des organisations. Aujourd'hui, la nécessité de pouvoir échanger des informations entre agents publics, de coopérer à la production ou à l'amélioration de ces données, est profondément étrangère aux principes de design des systèmes d'information tout comme aux relations usuelles entre administrations.

La création de la DISIC, en 2011, a engagé le processus de réponse à cette situation. La création de la fonction d'Administrateur général des données en 2014, puis l'intégration de la DISIC dans une Direction interministérielle du numérique et du système d'information et de communication de l'État (DINSIC), qui inclut aussi l'Administrateur général des données, en septembre 2015, va permettre d'allier dans une même stratégie le « back-office » et le « front-office », l'infrastructure et l'expérience utilisateur, ouvrant ainsi de nouvelles perspectives. Elle permettra en particulier de faire émerger, par une action interministérielle, une informatique d'État conforme aux meilleures pratiques actuelles : extrême disponibilité des applications critiques, capacité d'innovation agile, ouverture des systèmes, flexibilité des architectures, amélioration continue, management des dépenses organisé autour des services…

La stratégie d'« État plateforme[37] », élaborée par la DISIC et les DSI ministérielles en 2014 et 2015, pose les fondamentaux de cette révolution numérique de l'informatique publique, en privilégiant l'interopérabilité des systèmes fondée sur une logique de plateforme et d'API. Il reste cependant un long chemin avant qu'elle ne devienne un principe de base de l'informatique d'État.

L'État ne conserve pas suffisamment la maîtrise de son Système d'information

Diverses contraintes auxquels sont soumis les acteurs du Système d'Information de l'État (délais fixés pour la réalisation, difficultés liées aux ressources humaines, etc.) ont parfois conduit à une perte d'indépendance de l'État au profit d'acteurs tiers dans la maîtrise de son système d'information :

- Le recours à la sous-traitance se fait parfois sans **créer ou maintenir une équipe interne** disposant de connaissances sur les produits créés, capables de piloter le prestataire et de mettre en question ses propositions, de mettre en œuvre un changement de sous-traitant sans risque sur la continuité de service, et de choisir de ré-internaliser certaines tâches si nécessaire (par exemple réaliser des extractions et manipulations de ces données, sans qu'elles soient facturées à l'État à la tâche avec des tarifs et délais prohibitifs)[38].
- La propriété de l'État sur ses données fait parfois l'objet d'un **renoncement implicite ou explicite :** quand, techniquement, il devient impossible de récupérer des données stockées dans un système d'information, quand la localisation des données de référence n'est plus connue ou accessible, quand les outils permettant de les traiter n'existent plus, quand leur migration dans des formats ouverts et actuels n'est plus possible, voire quand l'État accepte d'en perdre une partie des droits de propriété intellectuelle (abandonnée aux **fournisseurs, qui parfois revendiquent même une « propriété intellectuelle » de certaines dimensions essentielles à l'usage des données :** métadonnées, données d'usage, structuration des données, et l'enfermement dans des logiques propriétaires qui finissent par geler la capacité de l'État à utiliser lui-même ses propres données).

Cet État de fait a également des conséquences en matière de sécurité du système d'information de l'État.

Ne pas pouvoir récupérer aisément ses données est une perte globale d'efficacité et de réactivité regrettable et parfois dangereuse. Ainsi, dans le traitement d'une attaque informatique ayant touché un ministère régalien, les données de sécurité (« logs ») n'étaient disponibles que chez le sous-traitant, à titre onéreux et moyennant un délai important, ce qui a retardé d'autant leur analyse. La sécurité des systèmes d'information peut appeler une capacité de traitement massif de données[39].

L'externalisation de l'hébergement ou du traitement des données, peut également receler des failles de sécurité qu'il faut maîtriser. Autant il est essentiel de soutenir la plus large ouverture possible pour les nombreuses données qui doivent être partagées (« open data »), autant il est important de s'interroger en permanence sur les conditions d'hébergement et de traitement des données qu'il a été décidé de ne pas ouvrir, au nom de l'un des secrets légaux ou de l'une des sécurités prévues par la loi CADA. Cette attention au transfert de données, au nom de choix d'hébergement ou de traitement, est essentielle. Elle deviendra de plus en plus importante avec le développement des usages du cloud, mais aussi avec le développement imminent de services de machine learning « as a service », quand il sera possible, d'un simple clic, et pour quelques milliers d'euros, de faire traiter des masses de données massives à distance[40].

La question de l'autonomie de l'État et de la capacité d'action doit redevenir la question centrale du design des systèmes d'information, et, notamment, de la construction du cadre juridique et contractuel.

Autonomie ne signifie pas autarcie. L'autonomie c'est la capacité à définir soi-même ses objectifs et ses moyens pour atteindre ces objectifs. Chaque décision d'achat public, chaque délégation de compétence, chaque décision de sous-traitance devrait être pensée à l'aune de cette question centrale : est-ce que la puissance publique conserve toute capacité d'agir ? Récupérer les données produites par l'activité déléguée, pouvoir accéder au code source du logiciel, pouvoir construire des passerelles d'interopérabilité entre des systèmes, conserver la capacité humaine indispensable à pouvoir piloter le devenir d'un projet… toutes ces ambitions doivent être au cœur des décisions d'achat et des spécifications des projets informatiques. La capacité à pouvoir mettre les données au service de l'action publique en dépend. Et de nombreuses autres capacités en dépendent tout autant…

[36] Volle M. (2006) : De l'Informatique, savoir vivre avec les automates, Economica
[37] http://references.modernisation.gouv.fr/strategie-du-si-de-letat
[38] Des actions correctives ont été engagées, il s'agit notamment d'un des critères d'évaluation des projets sur la base desquels le DINSIC rend son avis à l'occasion des audits de projets qu'il conduit.
[39] Source : entretien avec l'ANSSI
[40] Morin-Desailly C. (2013) : L'Union européenne, colonie du monde numérique ?, rapport fait au nom de la commission des affaires européennes du Sénat

3. LA CULTURE ADMINISTRATIVE N'ENCOURAGE PAS LE PARTAGE NI LA COOPÉRATION ENTRE LES ADMINISTRATIONS

Le difficile partage des données au sein des administrations

Personne ne sera surpris de découvrir dans ce rapport combien nous sommes loin d'une culture administrative **fondée sur la logique de coopération et de partage de données.**

Plusieurs raisons à cela : la crainte de violer certains secrets légaux, le sentiment général que le « croisement de fichiers[41]» serait proscrit, la crainte de réactions hiérarchiques, la crainte d'être débordé par une demande alors que les moyens nécessaires aux missions courantes tendent à diminuer… Toutes ces raisons sont légitimes et méritent des réponses sérieuses.

Mais il ne faudrait pas qu'elles fassent perdre de vue l'objectif central. **La culture de coopération interministérielle, de coopération avec les services déconcentrés, et avec les collectivités locales, est encore insuffisante.** Les systèmes d'information ont été conçus dans une logique de silos ministériels, financés dans des budgets distincts et cette situation finit par pénaliser réellement le fonctionnement de l'État.

Car **c'est le propre des données :** même si elles sont produites par une administration pour ses besoins propres, elles peuvent servir à de nombreuses autres fins. Les succès de l'open data – cette forme de partage radical - en témoignent : la base des accidents de la route horodatés géolocalisés n'intéresse pas seulement les forces de l'ordre, mais aussi les urbanistes ; la pollution des nappes phréatiques gagne à être croisée avec la pollution des rivières ; les prix de l'immobilier intéressent le ministère du logement tout autant que l'administration fiscale. On pourrait démultiplier les exemples.

L'enjeu principal n'est pas tant le comportement individuel des agents - qui sont pour grand nombre d'entre eux **convaincus des vertus de la collaboration** - mais bien dans la **culture administrative** elle-même, cet ensemble d'habitudes, de normes et de pratiques dans lequel les agents évoluent tout au long de leur carrière.

Les administrations sont souvent réticentes à partager des données car elles considèrent, souvent à tort, que la production de données est **si étroitement liée à leur mission** et à leur activité qu'elles n'ont guère d'utilité pour des tiers. Il n'est pas rare d'entendre un producteur revendiquer **un droit moral sur « ses » données,** à défaut d'un droit de propriété stricto sensu. La crainte du mésusage ou d'une mauvaise compréhension des données - ou la peur de mettre en lumière leur qualité insuffisante - sont souvent invoquées.

Ce manque de partage des données a **des conséquences très importantes pour l'État,** tant en perte d'efficience que d'efficacité. Des données indispensables au pilotage et à l'évaluation des politiques publiques ne sont ainsi pas mises à disposition de l'ensemble des acteurs concernés. Des administrations doublonnent parfois des bases qui existent déjà mais dont elles n'ont pas la connaissance. Plus généralement, l'**État se prive lui-même,** et dans le cadre de son fonctionnement, des **bénéfices attendus d'une meilleure circulation** et utilisation des données.

En situation d'urgence, comme nous en avons connu récemment, cette capacité d'échanger et d'interconnecter rapidement des données peut devenir extrêmement critique, notamment dans le domaine de la sécurité civile, de l'information du public ou du recueil d'informations en temps réel. **Il semble indispensable d'organiser rapidement un travail interministériel sur la manière de préparer une capacité d'échange rapide de données en contexte de crise.**

[41] S'il est encadré par la loi « Informatique et libertés », pour des raisons bien compréhensibles de protection de la vie privée, il n'est pas interdit. De plus, cet encadrement ne porte que sur les fichiers comportant des informations à caractère personnel. En soi, le principe de partage voire de croisement d'informations non personnelles ne pose aucun problème. Quant aux croisements des données personnelles, il n'est pas impossible mais soumis à autorisation préalable de la CNIL.

Les saisines de l'Administrateur général des données (AGD)

Le décret n°2014-1050 du 16 septembre 2014 précise que l'Administrateur général des données peut être saisi par toute personne de toute question portant sur la circulation des données. Les collectivités territoriales, les personnes morales de droit public et les personnes morales de droit privé chargées d'une mission de service public peuvent le saisir pour avis de toute question liée à l'utilisation par leurs services des données des administrations.

Une douzaine de saisines ont été reçues au cours de cette première année d'exercice :

- la moitié d'entre elles proviennent d'individus ou d'entreprises développant des projets utilisant des données publiques,
- quatre saisines sont issues des administrations publiques (principalement des collectivités),
- le solde des demandes provenant d'une association, d'un journaliste et de la Cour des comptes.

Ces saisines illustrent des cas de mauvaise circulation des données ayant eu pour conséquence une sous-efficience des administrations :

- des collectivités se voient refuser l'accès à des données sur les parcelles agricoles qui leur permettraient de mieux coordonner la lutte contre la pollution environnementale ;
- des acteurs en charge de la prévention de l'habitat indigne ne disposent pas de données sur les incendies d'immeubles qui leur seraient utiles pour mieux évaluer le risque associé à chaque adresse ,
- l'établissement public en charge de la gestion des avoirs saisis et confisqués (AGRASC) n'a pas accès aux données judiciaires, administratives et financières indispensables à l'efficacité de sa mission de service public. Cela a pour effet de ralentir le traitement des dossiers et de réduire le potentiel de l'établissement à abonder le budget général de l'État[42].

Par ailleurs, l'Administrateur général des données a remis un avis sur la publication, la rectification et la réutilisation des informations portant sur les professionnels de santé. Cet avis fait suite à une saisine du Ministère de la Santé et des Affaires sociales dans le cadre de la préparation du projet de loi Santé[43].

[42] Extrait du rapport annuel 2014 de l'Agence pour la gestion et le recouvrement des avoirs saisis et confisqué

[43] Le processus de saisine ainsi que les avis rendus sont publiés sur le site de l'Administrateur général des données : agd.data.gouv.fr . En 2015, une dizaine de saisines ont été reçues et traitées.

4. LES LOGIQUES DE GESTION BUDGÉTAIRE FREINENT LE PARTAGE ET LA COOPÉRATION ENTRE LES ADMINISTRATIONS

Le **bénéfice** d'un partage de données est **souvent collectif** (le collectif incluant d'autres administrations, voire d'autres acteurs de l'économie), alors que les efforts nécessaires au partage des données sont **souvent portés par leur producteur** (ou diffuseur), qui le plus souvent ne verra pas clairement l'intérêt qu'il pourra tirer à court et moyen terme de ce partage.

Comment concilier partage et gestion budgétaire ?

Mettre à disposition des données à des tiers représente un coût, généralement contrôlé, et parfois un manque à gagner, pour des entités administratives. Or ces entités agissent souvent dans un **cadre très contraint et très vertical.** C'est notamment l'**esprit de la LOLF** qui cadre les programmes sur un périmètre d'actions, des objectifs, et des moyens associés. Or ces objectifs n'incluent quasiment jamais la mise à disposition de données à des tiers. Pour le responsable du budget de l'entité administrative, le partage de données à titre gracieux ou à faible coût signifie une augmentation de ses coûts, sans quasiment aucun revenu complémentaire, et **sans impact sur l'atteinte des objectifs** dans le cadre des missions qui lui sont confiées.

En outre, si les coûts liés à l'ouverture des données peuvent être anticipés dans le cas de l'ouverture en open data des données (l'export de fichier ne générant pas de surcoût si les données sont consultées voire utilisées par un grand nombre d'acteurs), il est plus difficile d'anticiper le succès possible d'une ouverture contrôlée et nombre d'administrations craignent d'être victimes de ce succès. Une réponse possible à cette tension consiste à **inclure la mission de diffusion des données dans les missions de l'administration productrice,** afin de responsabiliser les décideurs en charge de la gestion budgétaire. Cette approche peut également nécessiter de **compléter la dotation budgétaire** des entités administratives concernées afin de leur donner les moyens d'assurer cette mise à disposition.

L'administration se vend des données à elle-même

La vente de données par des administrations de l'État ou ses opérateurs à d'autres autorités administratives représente un **manque à gagner important,** tant en terme d'**économie,** d'**efficacité que d'efficience.**

Cette situation est tout d'abord improductive d'un strict point de vue comptable. En effet, **il ne s'agit pas d'un jeu à somme nulle** au niveau de l'État, mais bien d'une perte nette liée aux coûts engendrés par ces transactions. On constate, au-delà de la vente de quelques grands référentiels à quelques grands clients internes, une **multitude de transactions de faible montant** (moins de 500 euros) qui sont réalisées chaque année entre administrations, opérateurs et collectivités. Le traitement comptable et administratif (facturation, règlement, suivi) de telles transactions engendrent des coûts sans commune mesure avec les montants engagés.

L'État lui-même n'est pas étranger à cette situation, où les grands producteurs sont contraints de trouver par eux-mêmes des sources de financement pour compléter les financements publics qui leur sont accordés. C'est le modèle économique de ces mêmes producteurs qu'il faut revoir pour faire en sorte que ce soit bien le bénéfice commun qui soit maximisé. La décision de rendre le référentiel à grande échelle (RGE) gratuit pour les missions de service public a constitué un premier pas en ce sens.

La vente des données à soi-même représente aussi une **perte d'opportunité** pour l'État. Certaines administrations renoncent à acquérir des données qui seraient pourtant utiles à l'exécution de leurs missions, d'autres **construisent leurs propres bases** pour ne plus dépendre du modèle économique de tiers, au risque de créer des doublons et de gâcher des ressources limitées.

Enfin, cette organisation entrave l'**ambition de doter l'État d'un système d'information unique,** cohérent et de qualité. Elle génère en particulier une quantité de doublons, de copies de jeux de données, et, de ce fait, de risques d'erreurs.

Les conclusions de la mission Fouilleron

À la demande de la directrice de cabinet du Premier Ministre, M. Antoine Fouilleron, auditeur à la Cour des comptes, a mené une étude approfondie concernant ces ventes de données entre les administrations[44], remise au gouvernement à la fin du mois de novembre 2015, qui aboutit aux recommandations suivantes :

Proposition 1 : Fixer le principe de gratuité des échanges de données entre les administrations au titre de leur mission de service public dans la loi, et ne prévoir une possibilité de maintien de tarification de ces échanges que pour les données à façon issues d'un traitement complexe et le cofinancement d'enquêtes.

Proposition 2 : Réaffirmer, au sein d'une circulaire du Premier ministre, le principe de gratuité totale des échanges de données entre les services de l'État, y compris pour les données à façon complexe, et étendre ce principe aux relations réciproques entre l'État et ses opérateurs pour l'exercice des missions de service public.

Proposition 3 : Assurer la neutralisation des flux budgétaires constatés au titre des échanges de données à titre onéreux par des transferts en base dans le projet de loi de finances pour 2017.

Proposition 4 : Accompagner la mise en œuvre du principe de gratuité des échanges de données entre administrations par le déploiement d'infrastructures et de services propres à favoriser la standardisation et la normalisation de ces échanges. Rédiger une licence-type pour les échanges de données entre administrations.

Proposition 5 : Approfondir l'analyse sur les freins non budgétaires à la bonne circulation des données entre les administrations, réaliser le répertoire des bases de données des administrations et objectiver les contraintes juridiques pouvant restreindre la diffusion aux administrations des données couvertes par un secret protégé par la loi.

[44] Fouilleron A. (2015) : Les échanges de données réalisés à titre onéreux entre les administrations, rapport au Premier ministre

5. DES FREINS ISSUS DES MODALITÉS D'APPLICATION DES « SECRETS LÉGAUX »

« La CNIL ne voudra jamais. » « Il faut respecter le secret statistique. » « Ici on approche du secret fiscal. » « On risque d'enfreindre le secret défense. » « La transparence s'oppose au secret médical. » « J'ai besoin d'une instruction hiérarchique. » « Les données ne sont pas assez bonnes et je vais être responsable de vos erreurs. » « Je pense être autorisé à vous transmettre des données mais je n'y suis pas obligé et je ne sais pas ce qu'en pense ma hiérarchie. » « Cette demande est instruite par le tribunal administratif. »… Quiconque a tenté, un tant soit peu, de publier ou simplement d'utiliser des données publiques s'est heurté, souvent, à ces réponses et à beaucoup d'autres.

Il est regrettable et préjudiciable que la CNIL soit ainsi instrumentalisée avant toute saisine par les administrations concernées, et ce, alors même qu'elle autorise la quasi-totalité des traitements, pourvus que leurs conditions de mise en œuvre, y compris en termes de sécurité informatique, soient satisfaisantes.

Les **« secrets légaux »** qui s'imposent, sont nombreux. Chacun d'entre eux a son histoire propre, son cadre de référence, sa portée et ses limites. **Protecteurs de libertés fondamentales, des intérêts fondamentaux de la nation ou nécessaires à l'exercice de certaines fonctions régaliennes, ils sont légitimes.** Certains d'entre eux méritent même probablement d'**être renforcés** à l'heure où la puissance de calcul s'est disséminée dans la société du fait de la diffusion de l'informatique individuelle, et où les méthodes du big data*, conjuguées à l'apparition de données massives d'un nouveau genre permettent de déduire de nouvelles informations de données que l'on croyait anodines. La plupart des craintes exprimées sont donc légitimes. Il n'en demeure pas moins que leurs **fondements juridiques sont souvent discutables,** qu'elles tendent à mêler des questions différentes et des problématiques hétérogènes, et que **ce climat d'inquiétude et de sécurité approximatives** devient un frein à l'établissement d'une bonne gouvernance de la donnée au service de l'efficacité de l'action publique.

Plusieurs rapports et études ont pointé d'éventuelles lacunes, voire quelques contradictions, du cadre législatif et règlementaire. Ce constat est peut-être fondé, mais après quelques années d'expériences, il semble secondaire face au constat de l'application approximative des secrets légaux.

Les flottements dans l'application des secrets légaux

Le secret, ce n'est pas la destruction de l'information.

Au contraire, un secret c'est une information qui est connue de certains, et qui ne doit pas être transmise à d'autres.

En démocratie, cette barrière est érigée pour protéger quelqu'un ou les intérêts fondamentaux de la Nation.

Le plus important, dans le cas d'un secret légal, est donc de **bien savoir à qui il s'oppose,** et dans quelles conditions. Ainsi, le secret médical ne s'oppose pas aux patients, mais à leurs proches : il ne concerne pas la relation médecin-malade. Le secret professionnel ne protège pas le secret du professionnel mais celui qui est contraint de se dévoiler devant un professionnel. L'habilitation au **secret de la défense** ne donne pas de droit d'accès à l'ensemble des documents classifiés – l'accès se fait sur la base du besoin d'en connaître, afin de limiter le risque s'agissant d'information dont la compromission sont susceptibles de porter atteinte aux intérêts de la Nation. Le **« secret statistique »** n'est pas une interdiction de produire des résultats statistiques portant sur l'individu. Il a été fondé sur une question spécifique : les obligations de transmission d'informations par les entreprises, édictées par la loi de 1951, en contrepartie desquelles l'État s'engageait d'une part à ne pas utiliser ces informations pour contrôler les entreprises et d'autres part à ne pas les dévoiler d'une manière qui mettrait en péril le secret des affaires desdites entreprises.

Or, on constate qu**'au fil du temps,** la protection de ces secrets légaux s'est installée **dans certaines habitudes,** qui ont petit à petit élargi le cercle strict des limites initiales du secret concerné, ou qui n'ont pas suivi les évolutions des données, des pratiques et des usages.

Au cours de l'année écoulée, l'Administrateur général des données s'est vu refuser de nombreuses données au motif que la CNIL n'accepterait pas cette transmission. Outre le fait qu'il lui appartenait d'obtenir l'autorisation de la CNIL pour les traitements qu'il entendait mettre en œuvre, la vérification effective a montré que, neuf fois sur dix, la CNIL n'avait pas été consultée et ne se serait pas opposée à l'accès à ses informations.

Et la liste de ces approximations pourrait être étendue sans difficulté au secret fiscal, au secret des affaires, à la responsabilité pénale d'un agent public en cas de divulgation d'un secret commercial, au secret statistique, etc.

Aujourd'hui, l'**ombre perçue des secrets légaux** a plus d'impact sur le fonctionnement de l'État que les secrets eux-mêmes. Il devient essentiel au bon fonctionnement de l'État de **remettre le droit, simplement le droit,** au cœur des pratiques d'échange de données et de réapprendre aux administrations que la **coopération devrait être la règle** et que les limites posées par les secrets légaux doivent être appliquées dans le strict respect de la volonté du législateur.

D'éventuels ajustements nécessaires

Après une année d'expérimentation, l'Administrateur général des données considère que la grande majorité des difficultés rencontrées en matière de partage de données proviennent surtout de précautions excessives qui **étendent indûment** la portée des secrets légaux. Il n'est cependant pas impossible que certaines difficultés proviennent de **frottements légaux** qui pourraient être corrigés.

Les avis divergent, par exemple, sur l'articulation entre la loi sur l'accès aux documents administratifs et la loi Informatique et Libertés.

Ces deux textes, en effet, **proviennent de raisonnements différents** et manipulent des concepts qui se recoupent mais ne se confondent pas.

La loi **Informatique et libertés,** adoptée le 6 janvier 1978, encadre le traitement automatique de données à caractère personnel. Elle répond à des **risques de dérive de l'informatique** alors naissante.

La **loi sur l'accès aux documents administratifs,** adoptée le 17 juillet 1978, organise le droit d'accès aux documents administratifs, droit qui s'enrichira progressivement d'un droit de réutilisation. Cette **transparence de l'action publique** s'arrête naturellement à la préservation d'autres intérêts garantis par la Loi, comme la propriété intellectuelle d'un tiers, les intérêts fondamentaux de la nation et la vie privée. Pour garantir cette protection de la vie privée, elle prévoit que ne sont communicables **qu'à l'intéressé** les informations dont la communication porterait atteinte à la protection de la vie privée, au secret médical et au secret en matière commerciale et industrielle. Au fil du temps, la jurisprudence de la CADA a précisé cette notion de vie privée et a notamment reconnu que certaines informations à caractère personnel concernant des personnalités publiques, par exemple, ne ressortaient pas de la vie privée.

La « loi CADA » place ensuite une deuxième sécurité portant sur la réutilisation en précisant que « les informations publiques comportant des données à caractère personnel peuvent faire l'objet d'une réutilisation soit lorsque la personne intéressée y a consenti, soit si l'autorité détentrice est en mesure de les rendre anonymes ou, à défaut d'anonymisation*, si une disposition législative ou réglementaire le permet. »

Les **autorités compétentes,** et notamment la CNIL et la CADA, ont développé un ensemble de règles leur permettant d'articuler ces deux textes dans **le respect de leurs objets respectifs.**

La question de leur application à l'open data a fait l'objet de nombreux travaux, notamment au sein du Conseil d'orientation de l'édition publique et de l'information administrative (COEPIA) qui a publié un « Mémento sur la protection des informations à caractère personnel dans le cadre de l'ouverture et du partage des données publique[45] », ou entre la CNIL, la CADA, la direction de l'information légale et administrative (DILA) et la mission Etalab pour garantir toutes les sécurités portant sur la réutilisation de grandes bases de jurisprudence[46].

Mais il n'en demeure pas moins que la **coexistence de ces deux logiques,** s'agissant de surcroît d'une tierce question, celle de la circulation de données au sein de l'administration, génère de nombreuses hésitations au sein de l'administration et mériterait une mise au point et une **véritable impulsion politique interministérielle.**

[45] Conseil d'orientation de l'édition publique et de l'information administrative (2013) : Mémento sur la protection des informations à caractère personnel dans le cadre de l'ouverture et du partage des données publiques
[46] voir par exemple : https://www.data.gouv.fr/fr/datasets/cass/

6. DIFFUSER LES PRATIQUES DES DATASCIENCES

La bonne gouvernance de la donnée est une condition nécessaire au développement des datasciences, mais, symétriquement, le développement des datasciences est sans doute indispensable aux efforts qu'appelle une bonne gouvernance de la donnée.

En effet, dans un contexte de fortes contraintes administratives et budgétaires, il est difficile d'engager les organisations, quelles qu'elles soient, dans un effort sans permettre à chacun de toucher du doigt, concrètement, la manière dont ces efforts peuvent rapidement permettre de simplifier ou d'optimiser les fonctionnements quotidiens et le succès des missions.

En d'autres termes, **la construction d'une meilleure gouvernance de la donnée sera tirée par les usages,** et le développement des usages fait partie intégrante d'une stratégie visant à construire une meilleure gouvernance de la donnée.

C'est pourquoi, l'Administrateur général des données a commencé sa mission en constituant une petite équipe de datascientists (composée de quatre personnes), qui a proposé ses services aux administrations et qui a réussi, en moins d'un an, à produire plusieurs résultats encourageants avec l'appui de ministères volontaires. On notera en particulier les résultats suivants :

- **un travail avec le Service des achats de l'État,** permettant d'analyser en détail la consommation d'électricité de l'État et de nourrir ainsi une stratégie d'achat optimisée[47] ;
- **un travail avec le Service des technologies et des systèmes d'information de la sécurité intérieure** ayant permis de développer un modèle de prédiction des vols de voiture à l'échelle d'un département ;
- **un travail mené avec les équipes de Pôle emploi** permettant de prédire avec une probabilité de 80 % une entreprise qui recrutera un profil donné dans le trimestre à venir, et qui a permis à Pôle emploi appuyé par le SGMAP de développer le service « La bonne boîte[48] ».

Ces premières expériences, qui seront poursuivies au cours de l'année 2016, ont permis de montrer l'importance du changement de paradigme que représente l'arrivée des datasciences et des « big data ». Ce changement est avant tout du côté des usages de la donnée, du « data to action ». Les mathématiques utilisées par la statistique traditionnelle et par les datasciences sont fondamentalement les mêmes, même si de nouveaux outils et de nouvelles méthodes se font jour régulièrement. Les écoles de l'Institut Mines Telecoms (IMT) ainsi que le Groupe des écoles nationales d'économie et de statistique (GENES) ont d'ailleurs commencé à enseigner ces approches « datasciences » et « big data*». C'est beaucoup plus dans la logique de l'action publique que se passe une transformation essentielle et parfois difficile à accepter.

Avec la révolution numérique, se fait jour une nouvelle logique de l'action fondée sur la donnée. Cette logique, orientée action, diffère parfois de la logique scientifique, orientée vers la production d'un savoir certain, vérifiable et reproductible. Ainsi, le monde des datasciences tolère plus que la science traditionnelle de travailler avec des données imparfaites, du moment que la logique de l'action intègre cette incertitude. Les pompiers de New-York, par exemple, patrouillent en fonction des recommandations d'un algorithme qui leur conseille certaines visites de prévention, à partir de corrélations qui ne sont peut-être pas des causalités. Mais ils vérifient chaque semaine si cet algorithme continue à susciter des visites efficaces. Ainsi également, les données sont de plus en plus souvent utilisées pour nourrir la décision de l'agent au guichet et non pas pour définir une stratégie imposée par la hiérarchie. Enfin, les données qui servaient à créer un savoir de décision sont de plus en plus souvent utilisées pour permettre une décision temps réel, ou pour nourrir les décisions des usagers du service public lui-même[49].

Ces attitudes et ces stratégies d'action sont encore rares dans les administrations (comme elles sont rares également dans les entreprises privées). Leur généralisation est à la fois l'objectif de l'AGD, et le plus puissant motif pour une administration de travailler à améliorer la gouvernance de la donnée.

[47] Ce travail est documenté par le SAE et l'AGD sur le site de l'AGD : https://agd.data.gouv.fr/2015/05/17/analyser-les-consommations-energetiques-des-batiments-publics/
[48] http://labonneBoîte.pole-emploi.fr/
[49] A titre d'exemple, le lauréat du Hackathon organisé par la CNAF et Etalab pour le 70e anniversaire de la Sécurité sociale est une équipe qui a proposé de travailler à partir des statistiques de visites aux guichets des Caisses d'allocations familiales pour proposer aux usagers de se regrouper en fonction de leurs attentes, afin que leurs visites soient à la fois l'occasion de développer une entraide entre ayant-droits et une occasion pour les CAF de mobiliser ce jour-là des agents particulièrement formés sur les problématiques en question.

3

Premières pistes pour une bonne gouvernance des données

La première partie de ce rapport a permis de souligner l'importance de la qualité et de l'accessibilité des données produites par l'État, et le potentiel que recèle l'application des méthodes des datasciences, non seulement pour évaluer ou piloter les politiques publiques, mais aussi pour les réformer en profondeur.

La deuxième partie a permis de désigner les principaux freins à cette ambition, et a montré que ces freins sont aussi bien techniques, organisationnels, culturels que juridiques.

Cette dernière partie montrera la nécessité d'un travail de long terme, interministériel et concerté, permettant d'aboutir à une transformation. De même que la DISIC a entrepris la construction, avec les ministères, d'une stratégie unifiée du SI de l'État, il sera nécessaire de construire un système des données de l'État.

Ce travail sera piloté par l'Administrateur général des données et rythmé par le rapport annuel sur la gouvernance de la donnée prévu par son décret fondateur.

Il est cependant d'ores et déjà possible d'ouvrir, dans les mois qui viennent, de premiers chantiers permettant de premières et substantielles améliorations de cette situation.

1. PARTIR DES DÉVELOPPEMENTS CONCRETS

En matière de numérique, l'action, concentrée sur la résolution de problèmes concrets et mesurables, est bien souvent, et de loin, la meilleure stratégie. Il y a de nombreuses raisons objectives à ce changement, dont certaines sont très bien présentées par Mike Bracken, la figure historique du Government Digital Service britannique, par exemple sur son blog[50]. Dans un célèbre article « *The strategy is delivery, again* », il explique que trop de projets numériques partent d'une vision politique abstraite, que l'on tente de traduire en processus, puis en systèmes d'informations, qui vont chercher à rencontrer des utilisateurs puis à s'installer dans la durée. À contrario, les processus agiles du numérique partent des besoins des utilisateurs, développent des services, construisent les systèmes à partir de ces services, travaillent ensuite ces services et ces systèmes en fonction des décisions politiques puis organisent le feedback permanent entre les ambitions politiques et les retours des utilisateurs.

C'est cette même conviction qui a guidé de nombreux projets portés au sein du SGMAP :

- la réalisation concrète et opérationnelle de **projets de datasciences** par l'équipe de l'Administrateur général des données ;
- le développement d'API et d'interfaces d'accès aux données géographiques (**api.carto**) ou d'entreprises (**api.entreprises**) désormais utilisés par plusieurs dizaines d'administrations et intégrés à de nombreux services ;
- le soutien au développement de modèles de simulation (**Open fisca**) ;
- des startup d'État comme **Marchés publics simplifiés, Mes Aides, La Bonne Boîte** ou **Le Taxi** qui, en essayant de régler de façon probante et mesurable un problème concret, s'efforcent également de créer une API ou une ressource ouverte utilisable par d'autres systèmes ;
- la création de **France Connect,** ressource ouverte à toutes les administrations, permettant concrètement aux citoyens de se connecter à tout service public, mais surtout, de déterminer lui-même des échanges de données entre différents systèmes pour simplifier ses relations avec l'administration ;
- le soutien à des projets collaboratifs (**Base adresse nationale**).

Ces développements, légers à l'échelle du système d'information de l'État, ancrent la stratégie de la donnée sur des pratiques réelles, et créent progressivement un réseau de pratiques et de possibilités d'interconnexions développant, au cœur même du système d'information, plus d'interopérabilité, plus d'agilité, plus de capacité d'action, et cela en lien permanent avec les usages.

La DINSIC poursuivra et amplifiera cette stratégie de développements concrets permettant de fluidifier le système d'information de l'État, notamment à travers la stratégie d'État plateforme, et amplifiera ses soutiens aux initiatives similaires émanant des différentes administrations avec lesquelles elle collabore.

[50] *Bracken M. (2013) : On Strategy : The strategy is delivery. Again., disponible sur mickebracken.com*

2. RÉVÉLER LA DONNÉE DISPONIBLE DANS L'ÉTAT

La cartographie complète des données disponibles dans les systèmes d'information de l'État est une **tâche ardue,** parfois jugée impossible.

La « loi CADA » avait prévu (article 17) la publication régulière par les administrations d'un **registre des « documents administratifs »** qu'elle détient. Mais bien que la jurisprudence de la CADA précise que les fichiers, bases de données, voire logiciels doivent être considérés comme des « documents » au sens de la loi, la plupart de ces registres portent sur des **documents publiés,** et omettent de traiter des données ou informations disponibles. Il faut dire que le contexte de 1978 était bien différent de celui de 2015. En quatre décennies, s'est opéré le passage d'une logique de documents, hébergés sur un nombre restreints d'ordinateurs, nettement validés par des agents publics en situation de responsabilité, à une logique de l'informatisation générale des systèmes, de distribution massive des capteurs et du calcul, ainsi que de l'explosion des données de gestion, des big data et des flux de données en temps réel.

La DISIC a engagé des travaux d'urbanisme du SI de l'État, et a pu dégager ainsi un **« plan d'occupation des sols »** qui désigne les grandes familles de données disponibles et fournit de premières orientations aux chercheurs de données.

Dans le cadre des travaux sur l'ouverture des données de santé, la mission Etalab a proposé un premier inventaire des principales bases de données disponibles dans le système de santé[51]. Cette approche a permis de mesurer la difficulté d'une **approche fondée sur l'enquête** auprès des administrations[52], ainsi que la difficulté de décrire dans un référentiel unique le contenu des bases de données, l'origine des données, leur précision, leur granularité, leur fréquence de mise à jour, les responsables et les différents secrets protégeant ces données.

Ces expériences ont montré la **difficulté de toute approche** qui se voudrait linéaire, **centralisée et exhaustive.**

Cette difficulté est renforcée par le fait, déjà souligné, que nombre des informations importantes proviennent d'une informatique de gestion, rarement perçue comme source de savoir par les autorités en charge.

C'est pourquoi l'Administrateur général des données lancera, dans l'année 2016, un projet de cartographie collaborative ouvert à toutes les administrations qui souhaiteront y participer et en bénéficier.

[51] Voir https://www.data.gouv.fr/fr/datasets/cartographie-des-bases-de-donnees-publiques-en-sante/
[52] Il a en effet fallu près de trois entretiens avec chaque administration concernée pour obtenir les résultats souhaités.

3. FAIRE ÉVOLUER LES SYSTÈMES D'INFORMATION DE L'ÉTAT

Ce rapport a souligné combien la capacité d'utiliser ses propres données dépend de la maîtrise de l'informatique d'État. Il a également souligné la naissance, dans toutes les grandes organisations, d'une informatique de nouvelle génération, qui n'est pas seulement pensée comme un outil statique au service de l'organisation antérieure, mais, au contraire, pour relever de nouveaux défis :

- capacité à délivrer rapidement et de manière itérative les nouvelles compétences numériques ;
- minimiser les coûts d'exploitation des anciennes applications ;
- répondre au niveau de service élevé exigé par l'ère numérique, par exemple en matière de simplicité, de disponibilité ou de sécurité.

La stratégie d'État plateforme, préparée par la DISIC et les DSI ministérielles, constituera le socle de cette évolution. Elle devra veiller à intégrer les exigences particulières liées à l'utilisation des données.

Elle devra, en particulier, **veiller à préparer « l'extractibilité » des données « by design ».** Les choix d'architecture et les métadonnées devront viser à conserver la capacité de séparer les données en fonction des limites à leurs usages[53]. Il lui faudra également veiller à développer des outils d'extraction ou d'interrogation efficients (sans être contraints de recourir à un marché de prestation ad-hoc forcément limitant) et modifier les processus qui président à la création et la tenue à jour des bases de données. L'ambition d'exportabilité des données doit devenir une préoccupation permanente du design des nouveaux systèmes.

Cette stratégie devra également privilégier les choix d'architecture et de gouvernance permettant d'avancer vers l'**utilisation en temps réel des données** (bien souvent, des données produites en temps réel ne sont « jugées » qu'une fois par an et ne sont donc pas utilisables au fil de l'eau).

De nouvelles règles d'audit des projets informatiques de l'État seront proposées aux DSI, puis systématiquement ajoutées aux processus de suivi des projets par la DINSIC, comme, par exemple, le contrôle de :

- la liberté juridique d'exploitation des données et des modèles associés : s'assurer que les intervenants sur le projet, et en particulier les prestataires et éditeurs logiciels ne revendiquent pas des droits de propriété intellectuelle susceptibles de limiter la capacité d'extraction, d'utilisation et/ou de réutilisation des données ;
- la capacité du SI à distinguer les données soumises aux secrets légaux et au respect de la vie privée et celles qui ne sont pas soumises à ces contraintes ;
- les conditions techniques de l'extractibilité : capacité du SI à permettre une extraction de données dans un format réutilisable, par le moyen d'un « dump » complet de la base et, le cas échéant, d'une interface ouverte de type API ou webservices non-propriétaires ;
- l'utilisation de référentiels, de nomenclature et d'ontologie existantes : pour favoriser la réutilisation des référentiels existants plutôt que la création de bases parallèles ;
- les conditions d'accès aux données pour des usages de type datasciences, mais aussi pour d'autres administrations (circulation de la donnée au sein de l'État) ;
- enfin, la capacité du SI à alimenter, si possible de manière directe et automatisée, des portails de données ouvertes de type www.data.gouv.fr

Enfin, la DINSIC devra également s'employer à **garantir l'accès de tous les ministères aux ressources leur permettant de tester concrètement le potentiel des datasciences** (par exemple, donner accès aux capacités de calcul : des ordinateurs/serveurs performants et la possibilité d'installer et de travailler avec des ressources libres comme le langage R ou le langage Python avec les librairies Pandas et Scikit-learn, langages massivement utilisés par les communautés de datasciences).

[53] *Par exemple, l'agrégation dans un même système de données couvertes par le secret défense et de données non couvertes a pu, récemment encore, empêcher des travaux de datamining qui auraient pu permettre de substantielles économies.*

4. DÉCLOISONNER LES ADMINISTRATIONS

Il est indispensable, pour réaliser la valeur des données, de décloisonner les administrations et d'encourager les **collaborations interministérielles.** Dans cette optique, l'Administrateur général des données s'appuiera au cours des prochains moins sur plusieurs chantiers en cours ou qui font l'actualité. Certains d'entre eux visent à **donner un cadre général** plus favorable à cette collaboration. Le premier d'entre eux concerne la vente de données entre administrations. La mission sur les ventes de données au sein de l'administration, confiée par le Premier ministre à Monsieur Antoine Fouilleron, a permis de quantifier ces ventes ainsi que les coûts de transaction et les pertes d'opportunités associés[54]. **L'Administrateur général des données, au sein du SGMAP, accompagnera les producteurs de données dans la transformation de leurs modèles économiques.**

Ce premier chantier donnera un cadre plus favorable à la circulation des données et à leur meilleure exploitation par la puissance publique. Cependant, le décloisonnement des administrations ne se décrète pas, il se pratique. Les financements liés au Programme d'Investissement d'Avenir (PIA) sont un outil privilégié pour encourager les collaborations interministérielles. 21 projets ont été sélectionnés dans le cadre de l'appel à projet sur l'industrialisation de la mise à disposition des données ouvertes[55]. Les ministères se sont fortement mobilisés sur ce type de dispositif, notamment à l'occasion d'un évènement « Project Camp » dédié.

L'Administrateur général des données anime un **réseau de correspondants et personnes ressources** dans les administrations. Ce réseau s'adresse à tous ceux qui pratiquent (ou souhaitent pratiquer) les datasciences au sein des ministères et, potentiellement, des collectivités territoriales. Il vise à encourager l'échange de bonnes pratiques, mais aussi à offrir à ces agents la bienveillance nécessaire à la réalisation de leurs missions. L'année 2016 permettra de le structurer et de le consolider. Diverses initiatives de création d'administrateurs des données dans les ministères ou chez les opérateurs sont en cours d'étude. L'Administrateur général des données y est favorable, surtout si elles respectent deux principes :

- penser cette fonction dans l'esprit des « chief data officers » de nombreuses entreprises privées et de nombreuses collectivités, à savoir une personnalité spécifiquement en charge de travailler à faire émerger des décisions fondées sur la donnée, ou des politiques publiques fondées sur la donnée. Il ne s'agit pas de créer une couche de contrôle supplémentaire, mais de pousser à une transformation de l'action, qui appelle de ce fait une qualité des données, une capacité à les extraire et à les manipuler, une sécurité juridique et technique, et une compétence en datasciences ;
- organiser ab initio la mise en réseau de ces initiatives autour de l'Administrateur général des données afin que ces différentes initiatives donnent naissance à une véritable capacité d'intelligence et d'action collective.

Quel financement pour les grands référentiels de données ?

On distingue trois modèles de financement des grands référentiels de données :

- le producteur fait payer celui qui doit s'enregistrer dans sa base : par exemple une entreprise ou une association à qui l'on facture des frais d'enregistrement et/ou de publication ;
- le producteur fait payer le réutilisateur, selon un modèle de redevances, la plupart du temps lié au niveau d'usage ;
- le producteur est financé directement par le service public, à l'instar de l'approche « Basic Data » retenue aux Pays-Bas et au Danemark.

Aucun pays européen n'applique strictement l'un ou l'autre de ces trois modèles, mais ils pratiquent plutôt des approches différenciées par type de données et/ou producteur.

L'étude POPSIS[56], commanditée par la Commission européenne s'est ainsi intéressée à calculer le ratio de couverture des coûts par les recettes liées à la vente de données d'un point de vue global (on considère l'ensemble du budget du producteur et non ceux directement lié à la production d'une base de données en particulier). Dans la moitié des cas étudiés par POPSIS, les revenus couvrent à peine 1 % des budgets totaux des producteurs, et la plupart du temps ce taux de couverture des coûts ne dépasse pas 5 à 15 %.

[54] Fouilleron A. (2015) : Les échanges de données réalisés à titre onéreux entre les administrations, rapport au Premier ministre.
[55] Dont la création de la base « J'accueille du public » qui recense les établissements recevant du public (ERP), l'industrialisation du processus d'anonymisation des données de santé, le projet Le.Taxi, la création d'une plateforme de l'information nautique géographique.
[56] http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=1158

5. UNE NOUVELLE DOCTRINE D'APPLICATION DES SECRETS LÉGAUX

Préciser la doctrine d'application des secrets légaux

« Il n'y a pas d'État sans secret ». La discrétion, la confidentialité, la protection des citoyens, des entreprises et de la sécurité nationale sont des devoirs fondamentaux de l'État. Mais ces principes étant rappelés, l'État a aussi le devoir d'appliquer les secrets légaux avec précision et discernement, dans le respect de l'esprit et de la lettre de la loi. Comme indiqué en deuxième partie de ce rapport, un secret n'est pas un mystère : il organise une partition entre ceux qui ont à connaître une information et ceux qui n'ont pas à en connaître. Cette partition doit être effectuée à la limite exacte souhaitée par le législateur ou l'autorité administrative qui a instauré un secret légal.

L'Administrateur général des données appelle donc les autorités administratives à prêter la plus grande attention à la doctrine de mise en œuvre des secrets légaux de leurs administrations respectives. Il peut accompagner autant que de besoin les autorités qui souhaiteraient préciser, consolider ou réexaminer cette doctrine.

L'Administrateur général des données propose que, pour chaque base de données créée et protégée par un secret, l'administration concernée documente, pour elle-même et pour des tiers autorisés, quelles sont les données précises qui sont protégées par le secret spécifique, et à quel titre.

Il pourrait par exemple être opportun, dans un cadre à définir, de demander au Conseil d'État de préciser cette doctrine et surtout les conditions de son application concrète par les administrations.

En revanche, dans le cadre précis pour lequel ont été définis les secrets légaux, l'Administrateur général des données considère que toutes les précautions doivent être prises et soutient en particulier l'idée selon laquelle les **données détenues par l'administration et relevant d'un secret entrant dans le champ du code pénal doivent être stockées et traitées sur le territoire national.**

Les « packs de conformité » de la CNIL

Parmi les différents secrets légaux, la protection de la vie privée tient probablement une place singulière. D'une part, il s'agit en effet d'une liberté fondamentale. D'autre part, il existe une autorité indépendante respectée, la CNIL, qui encadre le traitement automatique des données à caractère personnel et intervient de ce fait sur nombre de décisions ayant trait à la vie privée. Mais c'est également l'une des questions qui inquiètent le plus les citoyens, parfois à bon droit, parfois également parce que cette question est mélangée avec différentes questions différentes mais adjacentes, comme les écoutes illicites des affaires PRISM et autres, ou la prise de conscience croissante de la capacité de prédire certaines choses en croisant des données en apparence banales avec d'autres données. C'est donc probablement aussi le secret légal le moins bien connu, celui où l'on rencontre le plus d'assertions contradictoires et parfois infondées.

Consciente de cette difficulté, la CNIL a inauguré en 2014 un nouvel outil, les « packs de conformités ». *« Les packs de conformité constituent une réponse opérationnelle aux besoins des professionnels concernant l'application de la loi « informatique et libertés ». Il s'agit de travailler, dans le cadre d'une étroite concertation entre la CNIL et les acteurs d'un secteur, à la mise en place d'outils juridiques de simplification ou d'allégement des formalités (normes simplifiées, autorisations uniques, dispenses…) et de bonnes pratiques spécialement adaptées à un secteur professionnel. Ces packs permettent également d'anticiper sur les changements attendus avec le projet de règlement européen sur la protection des données. Il en résulte pour les responsables de traitement une simplification substantielle des formalités au profit d'une relation plus dynamique avec le régulateur. En ce sens, la capacité à rendre compte de la mise en conformité avec la loi devient un enjeu essentiel*[57]*. »*

[57] *http://www.cnil.fr/en/linstitution/actualite/article/article/les-packs-de-conformite-un-succes-grandissant/*

Cette démarche s'inscrit plus généralement dans la recherche de nouveaux outils de régulation, comme les labels qui permettent de valoriser la protection des données vis-à-vis de leurs clients, tout en étant juridiquement sécurisées vis-à-vis du régulateur. En trois ans, la CNIL a ainsi délivré 60 labels, et le label « gouvernance de la protection des données » est appelé, si l'on en croit le nombre de candidats (y compris, pour la première fois, des collectivités locales) à connaître un franc succès.

Le développement, avec des acteurs volontaires, d'un pack de conformité adapté à la puissance publique semble une direction extrêmement prometteuse. La CNIL et la DINSIC ont évoqué cette perspective, pour laquelle plusieurs DSI ministérielles ont manifesté leur intérêt. **La CNIL pourrait donc lancer cette démarche dès 2016, en concertation étroite avec l'AGD.**

Faciliter l'anonymisation

Un certain nombre de bases de données détenues par l'administration publique contiennent des informations à caractère personnel et ne peuvent donc être diffusée directement. En revanche, différentes méthodes existent pour faire disparaitre ce caractère personnel. On parle de « dépersonnalisation », d'anonymisation, ou, plus minutieusement, de retrait de la possibilité de ré-identification. En appliquant ces méthodes, on peut diffuser de l'information.

Aujourd'hui, la mise en place de ces techniques exige un investissement dans un domaine très spécifique. Ceci représente un coût non négligeable pour les administrations qui souhaitent mettre en place la stratégie de partage de la donnée, telle qu'elle a été décrite dans ce rapport. Elles doivent alors, seules, établir un processus d'anonymisation.

Pour une bonne gouvernance de la donnée, cette étape devrait être facilitée par la **mise en visibilité menée par l'AGD d'un pôle d'expertise sur ces questions.** Ce pôle devra inclure une meilleure maîtrise des questions liées à l'anonymisation des données, non pas uniquement sous l'angle juridique, mais également sous l'angle technique (capacité d'automatiser l'anonymisation à grande échelle). Il assurerait une veille technologique sur les questions d'anonymisation relatives aux différents types de données (tables données de réseau, données géolocalisées, etc.). Il devra, dans la mesure du possible, mettre à dispositions des outils permettant de mesurer le caractère personnel des données et aidant à l'éliminer. Il accompagnerait les administrations dans cette démarche et veillerait à fournir, en open source, un kit d'anonymisation.

6. DIFFUSER LES NOUVEAUX USAGES DE LA DONNÉE

La gouvernance de la donnée esquissée dans ce rapport est au service d'une transformation de l'action publique. Réciproquement, elle ne se mettra en place que si les administrations bénéficient prioritairement des fruits de leurs efforts. C'est l'impact des stratégies fondées sur la donnée, en termes d'efficacité, de maîtrise des coûts, de qualité de vie au travail, qui poussera les administrations à engager les réformes de la gouvernance des données qui s'imposent.

L'Administrateur général des données a pu montrer, au cours de l'année 2015, la possibilité concrète d'appuyer des administrations dans leurs propres projets et d'obtenir des résultats vérifiables.

Le rapport du CGEIET a montré l'existence de poches de compétences au sein de l'administration, ainsi que la récurrence de demandes d'appui (à la fois techniques et juridiques, ainsi que parfois de ressources comme de la puissance de calcul ou des environnements de travail adaptés aux big data*).

L'année 2016 doit avant tout permettre de construire de nouveaux projets de datasciences apportant des résultats concrets et vérifiables. L'Administrateur général des données travaillera à soutenir toutes ces initiatives, en appuyant les initiatives des administrations qui le souhaiteront, et en organisant la mise en réseau de ces compétences au sein des différentes administrations, pour créer une communauté de pratiques, stimuler l'apprentissage auprès des pairs, susciter la mise en commun de ressources et contribuer ainsi à la montée en puissance de la qualité de l'ensemble de ces équipes, au bénéfice de l'action publique.

Le marché d'appui en Datasciences

Au cours de l'année 2015, l'Administrateur général des données a préparé un marché cadre d'appui en datasciences.

Ce marché entre dans la stratégie générale du SGMAP d'appui aux administrations.

Il référencera une dizaine de fournisseurs qui seront activables via des marchés subséquents. Par l'intermédiaire du SGMAP, les administrations auront donc la possibilité de bénéficier de ressources supplémentaires pour mener des projets de datasciences.

Différents aspects pourront être traités dans les trois composantes importantes de la datascience :

- recherche et exploration de données ;
- analyse de données : modèles prédictifs, détection de signaux faibles, classifications, etc ;
- restitution des données.

Les prestataires seront amenés à exploiter des données publiques, ouvertes ou non. Un cadre sécurisant pour les administrations et pour les citoyens a été établi à cette fin (notamment en matière de protection des données). L'Agence nationale pour la sécurité des systèmes d'information et de communication évalue les entreprises sur ces aspects dans le cadre du dépouillement et interviendra dans la rédaction et l'attribution des marchés subséquents.

L'accord-cadre insiste sur la nécessité de transférer le plus possibles les compétences aux administrations. En fonction des marchés subséquents, l'administration sera entièrement autonome à la fin de la prestation. Le dépouillement est mené sous l'égide de l'Administrateur général des données par cinq administrations de différents ministères.

Le marché sera en place et activable durant le premier trimestre 2016.

Sur saisine du Secrétaire général du ministère concerné, toute administration pourra solliciter le SGMAP qui examinera l'opportunité de l'appui en fonction de la nature du projet et du nombre de demandes reçues.

CONCLUSION

La gouvernance de la donnée prend toute son importance dans le contexte d'un ensemble de transformations importantes :

- transformations dans la nature des données produites ou détenues par l'administration ;
- transformation des outils et méthodes de traitement de ces données ;
- transformation dans les logiques d'action et dans les stratégies autorisées par ces nouvelles données et ces nouvelles méthodes.

Ce premier rapport montre que l'État, comme la plupart des grandes institutions, privées ou publiques, n'est pas encore prêt à saisir tout le potentiel de ces données. La transformation sera collective, progressive, et suivra le développement des logiques de l'action fondée sur la donnée.

Le présent rapport fournit de premiers engagements et de premières recommandations qui peuvent être mises en œuvre dès l'année 2016 :

1. **Poursuivre la logique de développement de projets concrets, rapides et décloisonnants** illustrée par France Connect, Base adresse nationale, APIentreprises et les startup d'État, et résumée par le cadre stratégique de l'État plateforme. Cette logique tend vers une unification du SI de l'État à travers stratégie de plateformes et d'API, visant l'interopérabilité concrète, sécurisée, et dans le respect de l'autodétermination informationnelle ;
2. Lancer, dès 2016, une **cartographie collaborative des données disponibles dans l'État,** ouverte à toutes les administrations qui souhaiteront y participer et en bénéficier ;
3. Intégrer la **capacité à extraire et utiliser les données dans les critères d'examen des projets informatiques de l'État ;**
4. **Soutenir et développer les collaborations interministérielles,** à la fois par un appui privilégié de l'Administrateur général des données, par une priorisation des financements aux projets innovants interministériels interministériels et par le développement d'API thématiques (impôts, santé, ...) pour connecter les SI des ministères ;
5. Créer, à partir de l'AGD, **une compétence collective, à la fois technique et juridique, en matière d'anonymisation des données ;**
6. **Préciser la doctrine d'application des secrets légaux ;**
7. Solliciter la CNIL pour lancer un **pack de conformité** avec les administrations volontaires ;
8. Diffuser les nouveaux usages des données, soit par coopération directe avec l'AGD, soit en utilisant le marché d'appui aux administrations préparé par le SGMAP, et partager les résultats qui le peuvent entre administrations voire avec le public.

Ces premières mesures sont de nature à faire entrer profondément la question des données dans la palette des outils au service de la transformation de l'action publique.

Elles ne prennent leur sens que dans une logique profondément engagée vers l'action. La révolution de la donnée place l'État, comme les acteurs économiques et sociaux, à la frontière de l'innovation. Or, l'innovation n'est tirée que par les usages. Ce sont les citoyens et les administrations, ce sont les usages réels, les us et coutumes, qui sélectionnent, in fine, parmi les promesses de la technologie, celles qui deviendront des pratiques réelles, durables et utiles.

De nombreuses autres questions devront être traitées dans les années qui viennent, comme par exemple le contrôle démocratique de cette nouvelle puissance d'agir, l'éthique des données, les nouvelles stratégies d'action publique, le rôle économique des données publiques, la souveraineté nationale. Ces questions seront posées dans le monde entier. La révolution numérique, dont la révolution de la donnée est aujourd'hui la pointe avancée, est une révolution industrielle complète, qui redessine les équilibres économiques et sociaux et demande un intense travail de synthèse créative pour faire naître de nouveaux équilibres.

Dans cette révolution, la France devra trouver son propre chemin. Elle le fera avec d'autant plus de clarté, et d'autant plus de vigueur, qu'elle aura pris soin de maîtriser son destin, de s'approprier les outils et les pratiques, et de parler d'expérience.

GLOSSAIRE

A/B testing : le test A/B est une technique qui permet de tester deux versions différentes (A et B) d'un message ou d'une interface afin de déterminer la version la plus efficace du point de vue du destinataire du message ou de l'utilisateur.

AGD : Administrateur général des données. En France, la fonction a été créée par décret du Premier ministre le 16 septembre 2014. L'AGD coordonne l'action des administrations en matière d'inventaire, de gouvernance, de production, de circulation et d'exploitation des données par les administrations.

Anonymisation : l'anonymisation des données consiste à en modifier la structure afin de rendre très difficile ou impossible la « ré-identification » des personnes (physiques ou morales) ou des entités concernées (source Wikipedia).

API : les interfaces de programmation (en anglais "application programing interface") permettent à un logiciel de fournir des services ou des données à un autre logiciel de manière simple. L'API de géocodage proposée sur le site adresse data.gouv.fr permet par exemple de transformer une adresse postale en coordonnées géographiques (de type latitude, longitude).

Big data : mégadonnées (traduction officielle). Le big data désigne à la fois des données possédant certaines caractéristiques — volumineuses, variées —, mais aussi, par extension, l'usage qui peut en être fait.

Donnée : une donnée numérique est la description élémentaire de nature numérique, représentée sous forme codée, d'une réalité (chose, événement, mesure, transaction, etc.).

Données de référence : les données de référence sont des données fréquemment utilisées par de multiples acteurs publics et privés, et dont la qualité et la disponibilité sont critiques pour ces utilisations, comme, par exemple, les données des référentiels géographiques de l'État.

Données pivot : une donnée pivot (ou donnée-clé) est une donnée qui permet de relier plusieurs jeux de données, comme, par exemple, le numéro SIRET d'une entreprise.

Gouvernance de la donnée : ensemble de principes et de pratiques qui visent à assurer la meilleure exploitation du potentiel des données.

Registre : en administration, un registre est un livre dans lequel sont inscrites des informations administratives. Exemple : le registre du commerce et des sociétés géré par les greffes des tribunaux (source Wikipedia).

Machine learning : traduction : apprentissage automatique. Issu de l'intelligence artificielle, le machine learning est un ensemble de techniques où les algorithmes sont dits apprenants. C'est-à-dire qu'ils se perfectionnent et s'améliorent d'eux-mêmes en traitant de nouvelles données.

BIBLIOGRAPHIE

Administrateur général des données (2015) : Analyser les consommations énergétiques des bâtiments publics, disponible sur agd.data.gouv.fr

Anderson C. (2008) : The end of theory, the data deluge makes the scientific method obsolete, Wired Magazine

Andreessen M. (2011) : Why software is eating the world, The Wall Street Journal

Bracken M. (2013) : On Strategy : The strategy is delivery. Again., disponible sur mickebracken.com

Conseil général de l'Economie, de l'Industrie, de l'Energie et des Télécommunications (2015) : Meilleures pratiques pour le « big data » et l'analytique dans l'administration : une nouvelle étape, rapport au Ministre de l'Économie, de l'Industrie et du Numérique, la Secrétaire d'Etat chargé de la réforme de l'Etat et de la simplification et la Secrétaire d'Etat chargée du numérique

Conseil national de l'habitat, Etalab (2015) : ouverture des données publiques dans le champ du logement, synthèse des débats

Conseil d'orientation de l'édition publique et de l'information administrative (2013) : Mémento sur la protection des informations à caractère personnel dans le cadre de l'ouverture et du partage des données publiques

Davenport T., Patil DJ (2012) : Data Scientist, the sexiest job of the 21st century, Harvard Business Review

De Soto H. (2005) : Le mystère du capital : « Le mystère du capital : pourquoi le capitalisme triomphe en Occident et échoue partout ailleurs ?, trad. française Flammarion

De Vries M. (2012) : Re-use of public sector information, report for Danish Ministry for Housing, Urban and Rural Affairs

Desrosières A. (2000) : La Politique des grands nombres : histoire de la raison statistique, Editions La Découverte (2nde édition)

Flowers M. (2013) : NYC by the numbers, annual report to the mayor of New York

Fouilleron A. (2015) : Les échanges de données réalisés à titre onéreux entre les administrations, rapport au Premier ministre

Grossman N. (2015) : White Paper : Regulation, the Internet Way. A Data-First Model for Establishing Trust, Safety, and Security | Regulatory Reform for the 21st Century, Mimeo

Jetzek T.,Avital, M. (2013) : The Generative Mechanisms Of Open Government Data, ECIS 2013 Proceedings

Ministère des finances du Danemark (2012) : Good Basic Data for Everyone – A Driver for Growth and Efficiency

Morin-Desailly C. (2013) : L'Union européenne, colonie du monde numérique ?, rapport fait au nom de la commission des affaires européennes du Sénat

Press G. (2013) : A very short history of data sciences, Forbes.com

Trojette A. (2013) : Ouverture des données publiques : les exceptions au principe de gratuité sont-elles toutes légitimes ?, rapport au Premier ministre

Vickery G. (2010) : Review of Recent Studies on PSI Re-Use and Related Market Developments

Volle M. (2006) : De l'Informatique, savoir vivre avec les automates, Economica


Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
Premier ministre
SECRÉTARIAT
GÉNÉRAL POUR
LA MODERNISATION
DE L'ACTION
PUBLIQUE