ADMINISTRATEUR
GÉNÉRAL
DES DONNÉES

LA DONNÉE COMME INFRASTRUCTURE ESSENTIELLE

RAPPORT
AU PREMIER MINISTRE
SUR LA DONNÉE
DANS LES ADMINISTRATIONS
2016-2017

La documentation Française

ADMINISTRATEUR GENERAL DES DONNEES

LA DONNÉE COMME INFRASTRUCTURE ESSENTIELLE

La **documentation** Française

ISBN : 978-2-11-145683-9

Préface

« Au cours de l'histoire, les plus grandes transformations se sont produites lorsqu'une ressource jusque-là rare est devenue abondante : le passage d'une société de chasseur-cueilleur à une société agricole (plus de nourriture), l'invention de l'imprimerie (plus d'instruction), l'émergence de nouveaux processus de fabrication et la révolution industrielle (plus de produits fabriqués en grande série), ou encore le web, qui a donné naissance à notre monde interconnecté (plus de données). Ces « révolutions » ont systématiquement généré de nouvelles structures sociales, de nouvelles formes de gouvernance, de nouvelles sources de richesse, de nouvelles opportunités et, il faut le souligner, de nouvelles inégalités. C'est ce qui s'est produit dans le passé et ce qui se produit aujourd'hui, avec la révolution des données. »

C'est ainsi que nous ouvrions, il y a bientôt deux ans, le rapport[1] commandé par le Chancelier de l'échiquier britannique et Emmanuel Macron, alors ministre de l'Économie et des Finances, à Sir Nigel Shadbolt, coprésident de l'Open data Institute, et à l'administrateur général des données.

La fonction d'administrateur général des données (AGD) a précisément été créée, en septembre 2014, en vue de préparer l'Etat français à cette révolution, en lui permettant d'utiliser ses données pour mieux piloter ses politiques publiques et concevoir de nouveaux services. Cette ambition exige à la fois la capacité à trouver et à exploiter ses données, la maîtrise des outils des datasciences, mais aussi la culture suffisante pour mobiliser des méthodes à des fins d'amélioration du service public.

Il s'agissait de la première instance de ce type à l'échelle d'un gouvernement, inspirée des *chief data officers* qui commençaient à apparaître dans certaines grandes villes américaines et de grandes entreprises privées.

Petite équipe très modeste (quatre personnes), elle devait commencer à relever un immense défi : faire entrer l'État dans l'âge de la donnée, c'est-à-dire créer les conditions pour que l'État maîtrise ses données, les utilise à bon escient, les partage – dans le respect des secrets légaux – de manière à confier à chaque donnée la valeur maximale possible, et surtout apprenne à s'en servir pour concevoir et piloter les politiques publiques.

---

1 « La révolution de la donnée au service de la croissance », rapport du groupe de travail franco-britannique sur l'économie de la donnée (2016) co-présidé par l'administrateur général des données français Henri Verdier et Sir Nigel Shadbolt, et co-fondateur de l'Open Data Institute. A télécharger sur economie.gouv.fr : bit.ly/2GeZRa6

Les années 2014 et 2015 ont conduit l'administrateur général des données à se mobiliser au service de différents projets ministériels, prouvant aisément la valeur des datasciences pour une administration ayant identifié un problème précis et disposant des données nécessaires. Plusieurs exemples sont présentés dans ce rapport. Ces premières années ont également montré la complexité de la gouvernance des données et la nécessité pour l'État d'apprendre à surmonter ses barrières internes, qu'elles proviennent d'une application, parfois rigide, des secrets légaux, des silos informatiques, ou de la faiblesse de la culture de coopération entre les administrations. Le rapport 2015 de l'AGD[2], tout en précisant l'enjeu de la politique de la donnée, accordait une large place à ces questions.

À partir de 2016, l'administrateur général des données, désormais rattaché à la DINSIC comme l'ensemble de la mission Etalab, a pu accompagner une dynamique interministérielle, à travers par exemple la nomination de plusieurs administrateurs ministériels des données menant des travaux considérables dans leurs administrations respectives, la coopération avec plusieurs startups d'État, ou encore la création du programme des Entrepreneur d'intérêt général, qui permettait à de nombreuses administrations d'accueillir des datas-cientists.

Il pouvait ainsi progressivement se consacrer à de nouveaux enjeux et en particulier au problème de la circulation des données, que ce soit par son concours à la préparation de la loi pour une République numérique ou encore en intégrant la question des données au cœur de la stratégie d'Etat plateforme avec notamment la naissance de grandes API très structurantes et très usitées, comme l'API Entreprise[3] – qui partage plus de 15 millions d'informations par an – ou l'API de géocodage[4] - 1,6 milliard d'adresses par an.

Progressivement, l'administrateur général des données en venait à penser la donnée comme une infrastructure essentielle du fonctionnement de l'économie et de l'Etat, et se voyait confier, suite à la loi pour une République numérique, la mission de bâtir un « service public de la donnée » pour les données de référence. Cette infrastructure peut se révéler très puissante. Le succès d'entreprises comme Uber ou Airbnb nous a appris en effet que la maîtrise de l'infrastructure de données permet de disposer d'une infrastructure matérielle sans même avoir besoin de la bâtir ni de la posséder réellement.

Le présent rapport consacre un important développement à cette question des infrastructures de données, qui dépasse en partie la question de l'open data, et appelle des investissements et une structure opérationnelle bien supérieurs.

Au cours des années récentes, la question de la donnée a continué à prendre de l'ampleur, avec le renforcement des exigences sociales quant à la protection de la vie privée, qui se traduit par l'entrée en vigueur, en cette année 2018, d'un nouveau règlement européen pour la protection des données personnelles. Pendant la même période a également progressé la conscience de la nécessité de construire une gouvernance de l'usage des algorithmes pour que leur

---

2 « Les données au service de la transformation de l'action publique », rapport 2015 de l'administrateur général des données au Premier ministre. Disponible sur gouvernement.fr : bit.ly/2pMUV0V
3 entreprise.api.gouv.fr
4 adresse.data.gouv.fr/api

puissance soit compatible avec les exigences d'une démocratie moderne. Toutes ces questions ont également mobilisé l'administrateur général des données et ont trouvé de premières réponses dans la loi pour une République numérique.

2018 marquera probablement une nouvelle étape. En effet, l'explosion des données disponibles et de la puissance de calcul a donné un second souffle à une discipline scientifique essentielle : l'intelligence artificielle (IA). Sous la forme communément appelée « IA faible », l'IA baigne d'ores et déjà notre quotidien. Outre la bataille industrielle qui déterminera les rares pays possédant une réelle capacité de conception des ressources en intelligence artificielle, se profile également une bataille culturelle qui tournera en grande partie autour de la question des données. La question, en effet, est simple : avec quelles données, et donc à partir de quels schémas culturels, seront éduquées les intelligences artificielles qui joueront un rôle économique et social si déterminant.

L'engagement du gouvernement, porté avec force par le président de la République, suite aux conclusions du rapport que lui a remis M. Cédric Villani le 29 mars 2018, appellera la montée en puissance au sein de l'État de la capacité à mobiliser les données et à les utiliser au service de l'action publique.

Car la révolution en cours n'est pas prête de s'arrêter, et porte, pour le service public, l'exigence d'une évolution profonde, indispensable pour qu'il réussisse à remplir ses missions au meilleur état de l'art, dans le respect de l'exigence croissante des utilisateurs, à coûts maîtrisés. Et surtout pour qu'il s'approprie ces nouvelles capacités dans le respect de ce qui fonde notre République : la démocratie, gouvernement du peuple, par le peuple, pour le peuple.

Henri Verdier,
administrateur général des données

SOMMAIRE

Introduction

Depuis la création de la fonction d'administrateur général des données en septembre 2014, le gouvernement français a pris conscience de l'importance de la révolution des données et conçu une **politique de la donnée** autour de trois axes principaux : **la fourniture** de données de qualité notamment à travers le service public de la donnée, **la circulation** de la donnée avec le principe de l'ouverture par défaut des données communicables et le développement d'API favorisant l'échange de la donnée entre administrations et avec la société civile, et enfin **l'exploitation** des données afin d'améliorer l'efficacité de l'action publique.

Pour mettre en œuvre cette politique, le gouvernement s'appuie sur l'**administrateur général des données**, fonction occupée par Henri Verdier son équipe au sein de la mission Etalab et plus généralement l'ensemble de la direction interministérielle du numérique et du système d'information et de communication de l'État (DINSIC), sur le réseau des administrateurs ministériels des données qui ont été progressivement nommés dans les différents ministères et sur une politique d'innovation à travers le programme « Entrepreneur d'intérêt général » lancé en 2016.

La donnée est aujourd'hui **au centre de l'action publique** mais aussi de l'activité économique. L'État a un rôle de catalyseur à jouer pour l'ensemble de la société. La donnée doit être aujourd'hui conçue comme **une infrastructure essentielle** au fonctionnement de l'économie au même titre qu'un réseau de transport ou de télécommunications.

La fonction administrateur général des données

Instauré par le décret n° 2014-1050 du 16 septembre 2014, l'administrateur général des données (AGD) est placé sous l'autorité du Premier ministre et rattaché au directeur interministériel du numérique et du système d'information et de communication de l'État.*

L'AGD 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.

Il encourage et soutient de ce fait le développement et l'usage des pratiques des datasciences au sein de l'administration.

***** Voir le décret n° 2014-1050 du 16 septembre 2014 modifié par le décret n° 2017-1584 du 20 novembre 2017 relatif à la direction interministérielle de la transformation publique et à la direction interministérielle du numérique et du système d'information et de communication de l'État.

Une fonction exercée de manière distribuée

La fonction d'administrateur général des données est aujourd'hui occupée par le directeur interministériel du numérique et du système d'information et de communication de l'État (DINSIC). L'équipe opérationnelle est placée au sein de la mission Etalab et travaille étroitement avec les autres composantes de la DINSIC et en réseau avec les autres administrations.

Cette nouvelle organisation a permis à l'AGD de déployer une approche conjuguant :

Actions et mesures opérationnelles :

– ouverture des données publiques ;
– développement d'API ;
– conseil au gouvernement pour la conception du service public de la donnée ;
– analyse du traitement des données dans les grands projets devant être soumis pour avis conforme au DINSIC, en application de l'article 3 du décret n° 2014-879 du 1er août 2014[1] relatif au système d'information et de communication de l'État ;
– sans négliger pour autant ses missions initiales de réponse aux saisines de l'AGD, de développement de projets de datasciences en propre, et d'appui aux premiers administrateurs ministériels.

Avis et mission de réflexion :

– saisine de la Commission d'accès aux documents administratifs (CADA) ;
– audit du fichier des titres électroniques sécurisés (TES) à la demande du ministre de l'Intérieur aux côtés de l'agence nationale de la sécurité des systèmes d'information (ANSSI) ;
– mission franco-britannique sur la donnée au service de la croissance commanditée par le ministre de l'Économie et des Finances, la secrétaire d'État au Numérique, le chancelier de l'Échiquier et le Minister for Culture, Communications and Creative Industries.

Accompagnements et appui :

– appui à la constitution du service national des données de santé ;
– accompagnement du programme « Entrepreneur d'intérêt général » ;
– accompagnement du réseau interministériel de l'État (RIE) et de l'ANSSI dans le développement de nouvelles approches de sécurité des réseaux ;
– accompagnement de nombreux ministères dans le montage puis l'exploitation de hackathons ministériels ;
– appui à certaines startups d'État.

Cette démultiplication des leviers d'action présente un grand intérêt en matière de transformation de l'action publique. L'impact des datasciences,

1 Lire l'article 3 du décret n° 2014-879 sur legifrance.gouv.fr : https://bit.ly/2GBohJY

en effet, ne peut être atteint que dans un *continuum* allant de la création de réelles infrastructures de données à la transformation complète des métiers. Mais surtout, il suppose une profonde transformation culturelle, qui appelle une maîtrise des stratégies fondées sur la donnée (*data-driven strategies)*, ainsi que la capacité à maîtriser ses flux de données dans le respect des différentes sécurités qui s'imposent (sécurité des systèmes d'information, mais aussi protection des secrets légaux).

Le rapport de l'administrateur général des données au Premier ministre

Selon le décret de création de la fonction d'administrateur général des données, **«** *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* **».**

Ce rapport a donc trois objectifs distincts : **dresser un état des lieux** des pratiques des administrations en matière de données, **projeter les évolutions de l'économie de la donnée** et enfin **proposer des améliorations** pour permettre la meilleure exploitation du potentiel.

Ces objectifs sont adressés par chacune des trois parties du rapport :

– La première partie détaille les grandes lignes de la **politique de la donnée**. Elle retrace les actions réalisées par les administrations et les leviers activés par l'administrateur général des données depuis la parution du premier rapport.

– La deuxième partie est dédiée à l'analyse d'un thème central : la donnée comme infrastructure. Elle recommande la création d'une véritable **infrastructure de données** dans lequel l'État peut et doit jouer un rôle central.

– La troisième partie trace des pistes pour **transformer l'essai** et renforcer les actions de l'État en matière de données sur l'année 2018.

Le présent rapport, présenté sous l'entière responsabilité de l'administrateur général des données, résulte d'un travail collectif associant de nombreuses contributions des équipes de la DINSIC – notamment de la mission Etalab, en particulier de Paul-Antoine Chevalier, data-scientist – et de leurs partenaires ministériels, et doit beaucoup à l'engagement et à la plume de Simon Chignard, conseiller stratégique de la mission Etalab.

Première partie

La politique de la donnée : produire, faire circuler, exploiter les données

Pourquoi se préoccuper aujourd'hui des données produites ou exploitées par les administrations ? Le premier rapport de l'administrateur général des données avait déjà souligné les enjeux de l'exploitation pleine et entière des données au regard des évolutions en cours. Les tendances identifiées alors se sont amplifiées.

La première d'entre elles est liée au passage d'une économie des données fondée sur la rareté vers **un régime d'abondance**[1]. Les données sont aujourd'hui beaucoup plus faciles – et moins coûteuses – à produire et exploiter. Les systèmes d'information et les capteurs collectent aujourd'hui des données de manière beaucoup plus systématique qu'auparavant. Les données de l'État sont aujourd'hui parfois **en concurrence** avec des données produites par des tiers, selon des modalités différentes.

Cette transformation a des conséquences majeures sur la manière dont les administrations, et au-delà la société même, peuvent **créer de la valeur économique et sociale** à partir des données.

Jusqu'à une date récente, la situation était relativement paradoxale. D'un côté, les grands producteurs de données, opérateurs ou services d'administration centrale, monétisaient leurs données (*via* des redevances) auprès d'une poignée d'acteurs économiques[2]. De l'autre, l'exploitation de la très grande majorité des données était limitée à l'administration qui les produisait ou les collectait, entraînant ainsi une perte d'opportunité.

En ce sens, l'État était auparavant **un piètre gestionnaire de l'actif** que constituent les données, en monétisant celles qui auraient intérêt à circuler et en sous-exploitant les données pour son propre usage. C'est parce que **les données circulent** largement, quand aucun secret ne s'y oppose, qu'elles permettent de créer de la valeur, d'améliorer les services publics, de susciter de nouveaux produits ou services, de donner aux acteurs économiques les données indispensables à leur activité.

La **politique de la donnée** est ainsi la réponse aux défis et opportunités que représente la révolution de la donnée pour l'État et les administrations. Elle peut être résumée en trois principes : produire les données essentielles, les faire circuler, encourager leur exploitation.

---

1 Les conséquences de ce passage d'un régime de rareté vers un régime d'abondance fondent le rapport franco-britannique sur "La révolution de la donnée au service de la croissance" commandité en novembre 2015 par le ministre de l'Économie, de l'Industrie et du Numérique, Emmanuel Macron, et le chancelier de l'Échiquier, Georges Osborne, à Henri Verdier, administrateur général des données, et Nigel Shaldboldt, doyen du Jesus College à Oxford et coprésident de l'open data institute. Le rapport complet (en français) est disponible à cette adresse : http://www.modernisation.gouv.fr/sites/default/files/fichiers-attaches/rapport_revolution-donnee_juillet2016_vf.pdf

2 À l'image de l'acquisition par Google, en 2012, des bases de données de l'IGN. Cette opération, d'un montant de plusieurs millions d'euros, ne s'est pas reproduite, la société Google ayant acquis la capacité à mettre à jour les données par ses propres moyens.

1. Produire les données essentielles

L'État a pour premier rôle de **produire les données essentielles** au fonctionnement des administrations et de l'économie dans son ensemble.

La puissance publique n'a (heureusement) pas attendu le *big data* pour se préoccuper des données. L'État produit de longue date **les référentiels** indispensables à son action. Qu'il s'agisse de nommer ou d'identifier un lieu, une entreprise ou une personne physique, les administrations gèrent de nombreuses bases de données essentielles au fonctionnement du pays. Le Répertoire national d'identification des personnes physiques attribue ainsi à chaque individu un numéro unique (le numéro d'inscription au répertoire, couramment désigné « numéro de sécurité sociale »). Dans le domaine économique, la base d'identification des entreprises et des établissements (base Sirene, produite par l'INSEE) joue elle aussi un rôle essentiel dans l'organisation des échanges non seulement avec les administrations mais aussi entre les entreprises elles-mêmes.

Préserver la souveraineté informationnelle

L'État prend acte des **nouvelles formes de souveraineté** liées au numérique et aux enjeux de concurrence dans la production et l'utilisation des grands référentiels.

Contrairement à la situation qui prévalait il y a encore vingt ans, avant la diffusion massive de l'Internet et des réseaux de communication, plusieurs référentiels d'origine publique ou privée sont maintenant en concurrence. Par exemple, la société Bloomberg propose son propre identifiant des acteurs économiques, y compris des entreprises françaises.

Or, dans une économie numérique ouverte, la **notion de standard de fait** domine. Dans le domaine des données, cela signifie que les références ne sont plus décrétées par un acteur de manière unilatérale. Fait aujourd'hui référence ce qui est reconnu comme tel par ses utilisateurs.

Le standard de fait proposé par un acteur privé et étranger comme Bloomberg peut donc se retrouver *de facto* en concurrence avec le référentiel étatique national qu'est la base Sirene produite par l'INSEE. Les données de l'État ne pourront rester des standards que si elles sont largement distribuées et facilement accessibles.

Identifier et donner une reconnaissance aux données de référence

Chacun pressent que toutes les données publiques ne se valent pas et que certaines présentent un **potentiel d'usage plus élevé** que d'autres. Pourtant, jusqu'à une date récente, la loi ne leur reconnaissait pas de statut particulier. C'est maintenant chose faite avec la consécration par la loi pour une République numérique [3] de la notion de **données de référence**.

3 Loi 2016-1321 du 7 octobre 2016 pour une République numérique : https://www.legifrance.gouv.fr/eli/loi/2016/10/7/ECFI1524250L/jo/texte

Les données de référence, telles que définies dans le code des relations entre le public et les administrations (CRPA) répondent à trois critères[4] :
(i) elles servent à identifier ou nommer des produits, des services, des lieux et des personnes ;

(ii) elles sont utilisées fréquemment par des acteurs publics ou privés autres que l'administration qui les détient ;
(iii) la qualité de leur mise à disposition est critique pour ces utilisations.

Le premier critère (i) correspond à la notion de **donnée-pivot** ou donnée-clé : le numéro SIRET, qui sert à identifier de manière unique l'établissement d'une entreprise ou d'une organisation en est l'illustration. Il permet de relier plusieurs bases de données entre elles, par exemple la base Sirene avec la Base des établissements de santé (FINESS) ou encore les données sociofiscales.

Le second critère (ii) insiste sur la **valeur de réutilisation** de ces données. Les données de référence sont, littéralement, les données qui *font* référence.

Le troisième et dernier critère (iii) insiste sur la **criticité de la qualité** de leur mise à disposition.

Les 9 données de référence

	Producteur	Domaine(s)
Répertoire des entreprises et de leurs établissements (Sirene)	Institut national de la statistique et des études économiques (INSEE)	Économie
Répertoire national des associations	Ministère de l'Intérieur	Associations
Base de l'organisation administrative	Direction de l'information administrative et légale (Premier ministre)	Administrations
Référentiel opérationnel des emplois et des métiers (ROME)	Pôle emploi	Économie – Emploi
Plan cadastral informatisé	Direction générale des finances publiques (Bercy)	Géographie – foncier
Code officiel géographique	Institut national de la statistique et des études économiques (INSEE)	Géographie – organisation territoriale
Registre parcellaire graphique	Agence de services et de paiement – ministère de l'Agriculture	Géographie – agriculture
Référentiel à grande échelle	Institut national de l'information géographique et financière	Géographie
Base adresse nationale	IGN, La Poste, OSM France, Etalab	Géographie

4 Articles L. 321-4 et suivants du CRPA : https://www.legifrance.gouv.fr/affichCodeArticle.do;jsessionid=99AAC7F5C97B5DE451DC16F85E6E2A10.tplgfr34s_3?idArticle=LEGIARTI000033219118&cidTexte=LEGITEXT000031366350&dateTexte=20171127

Neuf données de référence ont été identifiées à ce stade[5] : cinq d'entre elles constituent un corpus cohérent de données géographiques, et quatre autres permettent d'identifier des entreprises, des associations, des administrations, des métiers et emplois. Toutes ces bases de données sont désormais accessibles *via* la plateforme data.gouv.fr sur un espace dédié[6].

Le **Répertoire des entreprises et de leurs établissements** est produit par l'Institut national de la statistique et des études économiques (INSEE). Le répertoire Sirene (Système informatique pour le répertoire des entreprises et des établissements) enregistre l'*état civil* de toutes les entreprises et leurs établissements, quels que soient leur forme juridique et leur secteur d'activité (industriel, commerçants, artisans, professions libérales, agriculteurs, collectivités territoriales, banques, assurances, associations…). Il comprend à ce jour plus de 10 millions d'entreprises et d'établissements. Les services enregistrent quotidiennement près de 10 000 modifications.

La base Sirene est disponible librement et gratuitement[7], en open data, depuis le 4 janvier 2017, en application des dispositions de la loi pour une République numérique promulguée en octobre 2017. Une mise à jour quotidienne est diffusée.

Le **Répertoire national des associations** (base RNA) est produit par le ministère de l'Intérieur. Il permet d'identifier de manière unique chaque association déclarée en préfecture (numéro RNA, anciennement numéro Waldec) et précise pour chaque association le titre, l'objet, le siège social de l'association et adresse de ses établissements, la durée, la nature juridique de l'association, le code d'objet social. Cette base comprend aujourd'hui 1,5 million d'associations.

La **Base de l'organisation administrative** de l'État et des services publics est produite par la Direction de l'information légale et administrative (DILA, services du Premier ministre). Ce référentiel constitue la base de référence pour identifier et contacter les structures des administrations centrales.

Le **Code officiel géographique** (COG) est la nomenclature officielle des cantons, des arrondissements, des communes, des départements et des régions, il en contient les libellés et les codes. L'INSEE est officiellement en charge de sa production depuis 2003. Le COG est mis à jour annuellement, notamment pour intégrer les fusions de collectivités. Il peut être considéré comme le fichier primaire de nombreuses sources de données géographiques.

Le **Plan cadastral** informatisé est produit par la Direction générale des finances publiques (DGFiP) du ministère de l'Économie et des Finances. Les parcelles sont identifiées de manière unique par le code INSEE de la

5 Article R. 321-5 du Code des relations du public et de l'administration : https://www.legifrance.gouv.fr/affichCodeArticle.do;jsessionid=99AAC7F5C97B5DE451DC16F85E6E2A10.tplgfr34s_3?idArticle=LEGIARTI000034196073&cidTexte=LEGITEXT000031366350&dateTexte=20171127&categorieLien=id&oldAction=&nbResultRech=

6 https://www.data.gouv.fr/fr/reference

7 Voir article L. 324-6 du Code des relations du public et de l'administration.

commune, le numéro de section et le numéro de parcelle. Le plan cadastral français disponible en ligne est composé d'environ 600 000 feuilles de plan aux formats image ou vecteur.

Le **Registre parcellaire graphique** (RPG), produit par l'Agence de services et de paiement (ASP), est la base de référence concernant l'occupation des terres agricoles. Il contient 7 millions d'objets graphiques et îlots.

Le **Référentiel à grande échelle** (RGE) est produit par l'Institut national de l'Information géographique et forestière (IGN). Il comprend cinq composantes (orthophotographique – BD ortho, topographique – BD topo, altimétrique – BD Alti, parcellaire – BD parcellaire et adresse – BD adresse).

La **Base adresse nationale** (BAN) est coproduite par l'Institut national de l'information géographique et forestière, le groupe La Poste, l'association Openstreetmap France et la mission Etalab (DINSIC). Cette base a pour but de référencer l'intégralité des adresses du territoire français. Elle contient la position géographique de plus de 25 millions d'adresses.

Pôle Emploi produit le **Répertoire opérationnel des métiers et des emplois** (ROME). Cette base est l'outil de référence pour les questions d'emploi et d'évolution des compétences. À titre d'exemple, le service Bob Emploi utilise ce référentiel pour calculer des proximités entre les métiers et proposer aux demandeurs d'emploi des trajectoires d'évolution.

La liste des données de référence a vocation à s'enrichir au cours des prochains mois, en lien avec la nomination des administrateurs ministériels des données.

S'ouvrir à de nouvelles formes de collaboration et de production des données

Les modes de production des données connaissent eux aussi des évolutions que l'État ne peut ignorer. Les données d'**acteurs privés** peuvent être utiles pour la puissance publique, à l'image de l'utilisation, par l'INSEE, des données des tickets de caisse de la grande distribution, pour mesurer l'inflation.

La loi pour une République numérique a ainsi introduit la notion de **donnée d'intérêt général**, c'est-à-dire des données pertinentes pour le public, au-delà parfois d'un lien direct avec une mission de service public, dont l'ouverture présente un bénéfice social général. Outre le champ des subventions publiques et délégations de service public, cette notion a pu être mobilisée pour ouvrir, secteur par secteur, des données devant permettre au consommateur d'être mieux informé et acteur de ses décisions ou à l'usager de bénéficier d'un service plus fluide (données sur les transports, l'énergie, la couverture mobile notamment), mais il est certain qu'une telle notion pourrait s'appliquer à terme de manière plus large. Cette question figure d'ailleurs parmi les thèmes de la révision de la directive sur les informations du secteur public que la Commission européenne vient d'engager.

Les **communs numériques** sont aussi porteurs d'un autre modèle de production et de gouvernance des données, qui met l'accent sur la collaboration et le partage.

La Base adresse nationale en est un exemple. Initiée dès 2015 par l'IGN, La Poste, OpenStreetMap France et Etalab avec l'appui de l'administrateur général des données, cette base est originale et unique non seulement par son contenu (qui en fait la base la plus exhaustive à ce jour concernant les adresses en France) mais aussi par sa gouvernance qui associe des administrations, des entreprises publiques et des contributeurs d'une association.

Ce modèle, fréquemment cité en exemple à l'étranger, peut être reproduit dans d'autres secteurs : la santé, l'aménagement du territoire, etc. Il peut être envisagé dès que la disponibilité d'une donnée est essentielle pour plusieurs acteurs (État, collectivités, acteurs privés ou associatifs) mais qu'aucun d'entre eux ne peut prétendre en assurer seul la production, la maintenance plus généralement la gouvernance.

Définir de nouveaux standards de données

L'importance des *standards de fait* ne signifie pas pour autant que l'État doit renoncer à proposer de nouveaux standards, *a fortiori* quand ceux-ci viennent concrétiser des priorités de politique publique. Mais les conditions de réussite, et d'appropriation, reposent sur la capacité de la puissance publique à faire émerger et à associer un écosystème solide en lien avec ce standard [8].

L'État contribue ainsi à définir non seulement des principes généraux (« la transparence de la commande publique ») mais aussi des **standards de données** pour les mettre en œuvre concrètement. Certains travaux visent à définir des normes et obligations de format, afin de rendre comparables et agrégeables des données qui sont produites par une grande variété d'acteurs. C'est le cas, par exemple, des données des infrastructures de recharge de véhicules électriques**, des données essentielles de la commande publique** et des conventions de subvention.

L'obligation de mettre à disposition les données essentielles de la commande publique (marchés et concessions) est issue de deux directives européennes transposées par deux ordonnances en 2016. Ces textes renvoient à la formule consacrée du « *format ouvert et librement réutilisable* » [9]. Ainsi, à partir du 1er octobre 2018, les acheteurs publics devront publier les données relatives à leurs marchés publics d'un montant supérieur à 25 000 euros, ainsi qu'aux concessions, en open data.

---

8 Ce principe est aussi valable pour les standards issus du secteur privé, comme l'illustre le cas des données de transport. La participation active de Google à la création du standard d'échange *General Transit Feed Specification* (GTFS) a permis non seulement de faciliter la réutilisation des données de transports, mais a aussi assuré à Google une position privilégiée pour intégrer des données de transport public dans ses services (et notamment Google Maps).

9 Voir les articles 107 du décret n° 2016-360, 34 du décret n° 2016-86 et 94 du décret n° 2016-361.

Afin d'encourager l'**écosystème naissant autour des données** de la commande publique, il est apparu nécessaire de standardiser celles-ci. Une réflexion importante a été menée sur les formats : l'État a conduit une expérimentation avec quelques administrations pilotes afin de les élaborer, ce qui a abouti à un référentiel standard fixé par un arrêté du 14 avril 2017[10]. Le même travail de standardisation a été effectué pour les **données relatives aux subventions**[11].

Parallèlement, dans le cadre du Partenariat pour un gouvernement ouvert (ou *Open Government Partnership*), la France s'est engagée avec cinq autres pays au sein de l'***Open Contracting Partnership*** pour développer et promouvoir un format standard de données ouvertes liées aux marchés publics[12] (l'*Open Contracting Data Standard*), afin de travailler conjointement à l'ouverture et la mise à disposition des données relatives à la commande publique et faire émerger du même coup un standard international. **La France a pris la présidence de cette organisation** le 28 novembre 2017 afin de le promouvoir, tout en développant des usages concrets autour de la réutilisation des données relatives à la commande publique.

Une meilleure coordination de la production des données serait source d'économie et d'efficience

Il y a encore des **gisements d'économie et d'efficience** dans la production des données de référence. Dans la situation actuelle, la production des données de référence repose sur quelques opérateurs ou directions ministérielles. Leur savoir-faire et leur expertise dans cette production doivent être soulignés, mais des progrès restent possibles dans **la gouvernance de la production** de ces données.

En effet, la plupart de ces producteurs agissent de manière assez autonome, notamment vis-à-vis de leurs tutelles ministérielles. Les principaux référentiels, bien que liés[13], sont ainsi produits sans nécessairement partager **un cadre stratégique commun**. Une meilleure gouvernance de la production serait ainsi un facteur d'économie et d'efficience.

Par exemple, la mise en place d'un numéro unique d'identification des associations est évoquée depuis longtemps comme un facteur de simplification en faveur du secteur associatif[14]. Aujourd'hui les associations qui demandent des subventions, paient des impôts ou emploient des salariés

---

10 Voir http://www.data.gouv.fr/fr/datasets/referentiel-de-donnees-marches-publics/ et https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000034492587&categorieLien=id

11 Décret n° 2017-779 du 5 mai 2017 relatif à l'accès sous forme électronique aux données essentielles des conventions de subvention : https://www.legifrance.gouv.fr/eli/decret/2017/5/5/PRMJ1636989D/jo/texte et arrêté du 17 novembre 2017 relatif aux conditions de mises à disposition des données essentielles des conventions de subvention : https://www.legifrance.gouv.fr/eli/arrete/2017/11/17/PRMJ1713918A/jo/texte

12 https://www.open-contracting.org/2016/12/07/open-contracting-version-francaise/

13 Un exemple de lien entre les bases : la composante adresse, qui est utilisée par de multiples bases de données de référence (notamment la base Sirene ou le Répertoire national des associations).

14 Voir le rapport du député Yves Blein au Premier ministre : *50 mesures de simplification pour les associations*, octobre 2014.

doivent souvent mentionner deux identifiants, l'un dans la base Sirene et l'autre de la Répertoire national des associations.

2. Améliorer la circulation de la donnée

Il faut encourager la circulation des données, dans le respect des secrets légaux et de la vie privée des individus. **La circulation doit devenir la règle**, et la non-circulation l'exception justifiée.

Les actions entreprises depuis la création de la fonction d'administrateur général des données répondent à deux objectifs complémentaires :

– adapter et **faire évoluer le cadre juridique** pour limiter les freins techniques, économiques et juridiques à la circulation des données ;
– concevoir et **opérer les outils et les dispositifs** (plateformes, API…) permettant de faciliter la circulation des données, en cohérence avec l'approche de l'État plateforme.

Ouverture des données : des progrès significatifs

Depuis la publication du premier rapport de l'administrateur général des données, de nombreuses bases de données essentielles et très détaillées ont été publiées dans différents domaines. Si l'objectif d'open data par défaut de la loi pour une République numérique n'est pas encore atteint, il y a eu depuis deux ans un changement d'échelle dans de nombreux secteurs.

Dans le domaine des données de santé, la Caisse nationale d'Assurance Maladie a poursuivi l'effort de publication des données de santé engagé avec la publication de la base des dépenses d'assurance maladie interrégime (DAMIR) en 2015. La CNAM a ainsi publié la base sur les prescriptions hospitalières de médicaments délivrées en ville (juin 2017) et la base sur les dépenses de biologie médicale interrégimes (mars 2017). Dans le domaine économique, l'ouverture du répertoire Sirene en janvier 2017 constitue une avancée majeure qui a donné lieu à un très grand nombre de réutilisations.

Dans le domaine des données géographiques, la publication du plan cadastral informatisé (septembre 2017) est un bon exemple de publication d'un jeu de données essentiel.

Dans le domaine du logement, la publication du Répertoire des logements locatifs des bailleurs sociaux (décembre 2017) contenant des données détaillées sur 4,9 millions de logements sociaux constitue un autre exemple de la publication de données granulaires à haute valeur ajoutée.

Adapter et faire évoluer le cadre juridique

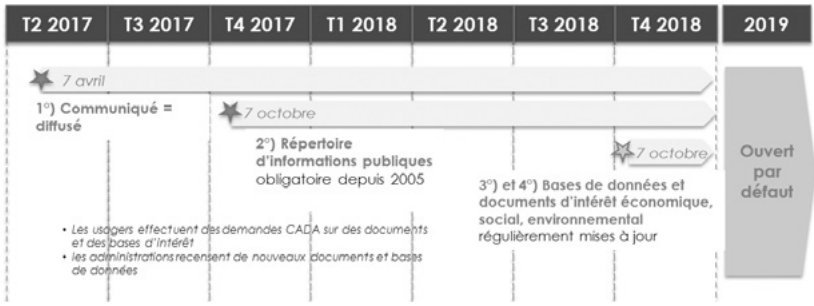
Depuis la parution du premier rapport de l'administrateur général des données, le cadre juridique et réglementaire a été profondément renouvelé pour faciliter la circulation des données, par les lois Valter (sur la gratuité des données publiques) et République numérique.

De la communication sur demande à la diffusion spontanée

La loi pour une République numérique a considérablement accru **le champ des documents administratifs mis à disposition en ligne**, en passant d'une logique de communication sur demande de l'usager (un droit d'accès) à une diffusion par défaut des données publiques.

En effet, toute administration de plus de cinquante agents (à l'exception des collectivités de moins de 3 500 habitants), est désormais dans l'obligation de diffuser, aux termes de la loi, dans un standard ouvert et aisément réutilisable [15] :

- les documents communiqués à la suite d'une demande d'accès (depuis le 7 avril 2017) ;
- les documents figurant dans les répertoires d'informations publiques (7 octobre 2017) ;
- les bases de données mises à jour de façon régulière (à partir du 7 octobre 2018) ;
- les données, mises à jour de façon régulière, présentant un intérêt économique, social, sanitaire ou environnemental (à partir du 7 octobre 2018).



Dans cette perspective, d'ailleurs, la communication d'un document administratif à un usager qui en a fait la demande peut désormais s'effectuer par publication des informations en ligne, à moins que les documents ne soient communicables qu'à l'intéressé lui-même [16].

15 Voir pour l'open data par défaut, l'article L. 312-1-1 Code des relations entre le public et les administrations (CRPA) et, pour le standard de diffusion, l'article L. 300-4 CRPA.
16 Voir l'article L. 311-9 CRPA.

Vers une uniformisation du cadre juridique pour toutes les « administrations »

La seconde évolution majeure concerne l'uniformisation du cadre juridique pour l'ensemble des administrations. En effet, la notion d'administration (au sens du code des relations entre le public et l'administration) est particulièrement large : entrent dans ce champ toutes les administrations (service public administratif comme industriel et commercial) mais aussi toute personne morale de droit privé chargée d'une mission de service public, pour les données produites ou reçues dans le cadre de cette mission. Cela concerne par exemple les exploitants de réseaux de transport qui opèrent dans le cadre d'une délégation de service public.

Toutes ces administrations, **sans distinction**, sont soumises aux obligations d'accès et de diffusion des données publiques ; elles sont également soumises aux mêmes règles de réutilisation, avec la **suppression de l'exception** faite aux données des services publics industriels et commerciaux et de la dérogation donnée aux services culturels leur permettant de définir librement les conditions de réutilisation de leurs données. On rappelle que la réutilisation n'est pas limitée et permet l'exploitation à des fins commerciales.

La loi supprime par ailleurs la possibilité pour les administrations de **se prévaloir de droits de propriété intellectuelle** pour faire obstacle à la libre réutilisation de leurs bases de données (droit *sui generis* du producteur de base), sauf pour les bases de données produites dans le cadre d'une mission de service public industriel et commercial soumise à la concurrence.

En conclusion, quelle que soit la catégorie d'administration concernée, les conditions d'ouverture, d'accès et d'exploitation des données sont similaires, à l'exception des données des SPIC en concurrence.

L'existence d'un cadre unique apporte **une clarification bienvenue** pour les usagers réutilisateurs et pour les administrations productrices elles-mêmes.

Des redevances de réutilisation à la gratuité des données publiques

La **gratuité de la réutilisation** des données publiques est maintenant devenue la règle. Les exceptions à ce principe sont maintenant très encadrées et doivent être motivées[17].

Ne peuvent recourir à des redevances de réutilisation que deux catégories d'administrations :

– celles dont l'activité principale consiste en la collecte, la production, la mise à disposition ou la diffusion d'informations publiques, dont la couverture des coûts liés à cette activité principale est assurée à moins de 75 % par des recettes fiscales, des dotations ou des subventions. Dans les faits, cela concerne le SHOM, Météo France et l'IGN ;

17 Avec la loi « Valter », transposant la directive 2013/37/UE du Parlement européen et du Conseil du 26 juin 2013, voir l'article L. 324-1 et suivants CRPA.

– les institutions culturelles pour leurs seules données issues d'opérations de numérisation.

Le montant des redevances est, par ailleurs, **plafonné** et doit correspondre aux coûts de production et de mise à la disposition du public des données concernées.

D'autre part, la loi a prévu spécifiquement la gratuité de toutes les données produites par l'INSEE et les services statistiques ministériels.

Ceci s'inscrit dans la continuité des recommandations du rapport Trojette. Ce rapport soulignait en effet un relatif affaiblissement des recettes dont une part non nulle émane d'acteurs publics (près de 15 % du montant total perçu en redevances) et un coût de gestion de la commercialisation élevé au regard de la valeur dégagée.

La gratuité lève un frein économique important à la réutilisation des données. Certaines administrations renonçaient, faute de budget disponible, à acquérir des données pourtant indispensables dans le cadre de leur mission. Il s'agit aussi d'une forme de **simplification** : la signature d'une licence ou d'une convention, l'éventuelle négociation sur les montants de redevances, l'investissement dans la commercialisation n'ont plus lieu d'être.

Une politique de licences au service de la circulation des données

Afin d'éviter la prolifération des licences et d'assurer la circulation la plus fluide des données en open data et, notamment, de permettre leur croisement, la loi pour une République numérique a prévu que, lorsque la réutilisation à titre gratuit donne lieu à l'établissement d'une licence, cette **licence est choisie parmi une liste limitative** de licences, fixée par décret et révisée tous les cinq ans[18], disponible ici : www.data.gouv.fr/fr/licences

Lorsqu'une administration souhaite recourir à une licence ne figurant pas sur cette liste, cette licence doit être préalablement homologuée par l'État, dans des conditions fixées par décret.

Par ailleurs, à l'occasion de l'évolution du corpus législatif, une nouvelle version de la « Licence ouverte/*Open Licence* » a été introduite, qui offre une liberté de réutilisation des informations la plus large en autorisant la reproduction, la redistribution, l'adaptation et l'exploitation commerciale des données et s'inscrit dans le contexte international en étant compatible avec les standards des licences open data développées à l'étranger et notamment celles du gouvernement britannique (*Open Government Licence*) ainsi que les autres standards internationaux (ODC-BY, CC-BY 2.0). La seule exigence est la mention de la paternité et de dernière mise à jour, ainsi que celle de la présence d'éventuelles données à caractère personnel ; il est également garanti que les droits de propriété intellectuelle grevant les données ne font pas obstacle à la libre réutilisation.

18 Voir l'article. D. 323-2-1 CRPA.

Vers davantage de circulation des données entre administrations

Une attention particulière est portée à la circulation des données entre administrations, dans une double logique de :

– **mutualisation** et réutilisation, notamment des grandes bases de données, pour renforcer l'efficience de l'action publique ;
– **simplification** administrative au profit des usagers, autour du principe « Dites-le-nous une fois ».

L'article 1er de la loi pour une République numérique crée un droit d'accès et de communication des administrations aux données publiques détenues par d'autres administrations, sous réserve des secrets protégés par la loi, pour l'accomplissement de leurs missions de service public : les informations peuvent être réutilisées par toute administration qui le souhaite à des fins de service public autres que celles pour les besoins de laquelle les données ont été produites ou reçues. Il s'agit pour ainsi dire d'**étendre aux administrations les droits** dont disposent d'ores et déjà les citoyens.

En outre, depuis le 1er janvier 2017, les échanges d'informations publiques entre les administrations de l'État, entre les administrations de l'État et ses établissements publics administratifs et entre les établissements publics précités, effectués dans ce cadre, ne peuvent donner lieu au versement d'une redevance (faisant exception au cadre présenté ci-avant). Cela s'inscrit dans la continuité des recommandations du rapport Fouilleron, remis au Premier ministre en décembre 2015. Ce rapport avait permis d'évaluer les effets pervers de la tarification des données entre administrations. Était notamment souligné le fait que la moitié des transactions liées aux données avait un coût unitaire inférieur à 500 euros, on imagine aisément les coûts de gestion liés à une telle transaction. En ce sens, la vente de données entre administrations n'était pas un jeu à somme nulle. Cette tarification nuit à l'efficacité et la qualité de l'action publique, en engendrant des coûts de transaction et des effets négatifs comme le renoncement à la donnée pour des raisons budgétaires ou des stratégies de contournement du frein budgétaire de la part des administrations acheteuses.

Parallèlement, dans une logique de simplification en évitant la redondance des demandes de pièces justificatives, les administrations peuvent échanger entre elles des informations, dans le cadre de la réalisation de démarches administratives par les usagers [19].

La régulation de l'open data et de la circulation des données

Les missions renouvelées de la Commission d'accès aux documents administratifs (CADA)

Dans ce nouvel environnement juridique, la commission est en effet désormais compétente pour se prononcer sur les demandes d'avis ou de conseils relatifs à la diffusion en ligne de documents administratifs, au-delà des questions de

19 Voir articles L. 114-8 à 10 et L. 113-12 et 13 du CRPA.

communication et de réutilisation, y compris le sujet des licences et redevances éventuelles. Elle s'affirme également comme un acteur déterminant de la « régulation de l'open data », dans son dernier rapport annuel.

Les saisines de l'administrateur général des données

Pour améliorer la circulation des données, le décret instituant l'administrateur général des données a créé la possibilité pour toute personne de saisir l'AGD pour toute question relative à la circulation des données et pour les collectivités territoriales, les personnes morales de droit public et les personnes morales de droit privé chargées d'une mission de service public, de le saisir pour avis pour toute question relative à l'utilisation des données de l'administration par leurs services[20].

Le dispositif des saisines de l'AGD complète le dispositif de saisine de la CADA. Lorsque la saisine porte sur l'accès à un document correctement identifié par le demandeur, il est préférable de s'adresser en priorité à l'administration concernée puis le cas échéant à la CADA. L'AGD peut notamment apporter son appui, en cas de difficultés techniques dans la mise à disposition des données.

Des saisines d'origines variées

L'administrateur général des données peut être saisi par des administrations, des entreprises et tout individu sur tout sujet en lien avec la circulation et/ou l'exploitation des données publiques. Cette diversité des demandes est illustrée par quelques demandes traitées en 2017 :

- sollicité pour avis par la Commission d'accès aux documents administratifs à propos d'une demande d'accès à un flux de données en temps réel, l'AGD a rendu un avis technique sur les conditions à remplir pour qu'un tel flux soit considéré comme publié *« dans un standard ouvert, aisément réutilisable et exploitable par un système de traitement automatisé »*, c'est-à-dire corresponde aux critères de mise à disposition fixé par la loi ;
- saisi par une entreprise sur la disponibilité d'informations sur l'état du marché immobilier, l'AGD a pu apporter immédiatement une réponse sur l'ensemble des sources de données et le cadre juridique de leur mise à disposition au public, en prenant en compte les évolutions législatives à venir ;
- saisi par un citoyen sur l'accès aux documents jurisprudentiels, l'AGD a engagé un projet d'anonymisation en concertation avec la CNIL et le ministère de la Justice ;
- saisi par le ministère du Logement sur l'accès aux données notariales, l'AGD a apporté son expertise juridique pour la préparation d'un décret en Conseil d'État.

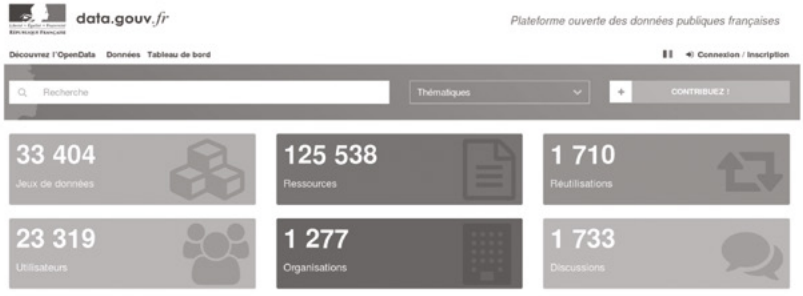
20 Le formulaire de saisine de l'AGD est disponible sur le blog de l'AGD : https://agd.data.gouv.fr/saisines-de-lagd/formulaire-de-saisine/

Au 31 décembre 2017, le portefeuille comptait une **vingtaine de saisines** pour quarante dossiers clos. Cet échantillon a permis de prouver l'utilité et l'efficacité du dispositif alors qu'il est encore assez confidentiel. La publication à venir des saisines récurrentes et le développement du réseau des administrateurs ministériels de données permettent d'envisager sereinement une augmentation du volume de saisines.

Concevoir et déployer les outils et les dispositifs pour faire circuler les données

Dans la logique de l'État plateforme, la mission Etalab au sein de la DINSIC opère un ensemble d'**outils et de dispositifs** – plateformes, API, services – qui facilitent la circulation des données.

La plateforme ouverte des données publiques françaises



La stratégie de diffusion des données ouvertes s'appuie sur la plateforme data.gouv.fr. La plateforme compte aujourd'hui plus de 33 000 jeux de données ouverts par plus de 1 200 organisations.

Parmi les 1 200 organisations, on trouve l'ensemble des ministères et la plupart des agences dont ils ont la tutelle, l'Assemblée nationale et le Sénat, des autorités administratives et judiciaires comme le Conseil constitutionnel, la Cour des comptes, la Haute Autorité pour la transparence de la vie publique, la Commission nationale informatique et libertés ou encore la Commission d'accès aux documents administratifs ainsi que des collectivités, représentant l'ensemble des échelons territoriaux, de la plus petite commune aux plus grandes régions.

Depuis deux ans, l'audience de data.gouv.fr augmente de manière régulière, pour atteindre plus de 185 000 visiteurs uniques au mois de décembre 2017.

Évolution de la fréquentation de data.gouv.fr

	Nombre de visiteurs uniques par mois	Progression année N-1
Décembre 2013	47 000	
Décembre 2014	53 000	+12 %
Décembre 2015	77 000	+47 %
Décembre 2016	127 000	+63 %
Décembre 2017	185 000	+46 %

Les verticales : des espaces thématiques dédiés

Il est apparu ces dernières années que cataloguer des milliers de jeux de données n'était pas totalement suffisant pour permettre aux administrations et plus largement à la société dans son ensemble d'en exploiter tout le potentiel. En effet, maximiser la réutilisation de ces données demande non seulement de les rendre découvrables mais aussi de lever les barrières à l'usage pour en faciliter l'appropriation ainsi que d'ouvrir et encourager un dialogue entre producteurs et réutilisateurs potentiels afin de favoriser des boucles de rétroaction vertueuses.

Les actions, méthodes et outils mis en œuvre pour ce faire sont spécifiques aux données considérées.

Ainsi, les informations nécessaires à un réutilisateur potentiel pour prendre en main un jeu de données géographique (échelle, projection, système de coordonnées, etc.) ont peu de sens pour un réutilisateur des données comptables qui à l'inverse aura besoin de connaître le plan comptable utilisé. Ces différences de nature des données considérées ont mené à l'émergence de multiples communautés d'usage rassemblées autour de données de référence thématiques, de formats, d'outils et de pratiques qui structurent verticalement l'écosystème de la donnée. L'univers géomatique, par exemple, regroupe l'ensemble des producteurs et réutilisateurs de données géographiques, sous l'égide de la directive INSPIRE.

C'est en faisant levier sur cette organisation de fait de l'écosystème de la donnée en « verticales » thématiques, qu'Etalab enrichit désormais la plateforme data.gouv.fr. Les trois premières initiatives de ce type correspondent à des écosystèmes déjà bien structurés : ce sont ceux centrés sur les données géographiques, sur les données ayant trait aux entreprises et enfin sur les données de transport. Ces verticales permettent de développer une offre de services spécifiques adaptée à ce type de données et de fédérer des communautés spécialisées.

La verticale des données géographiques

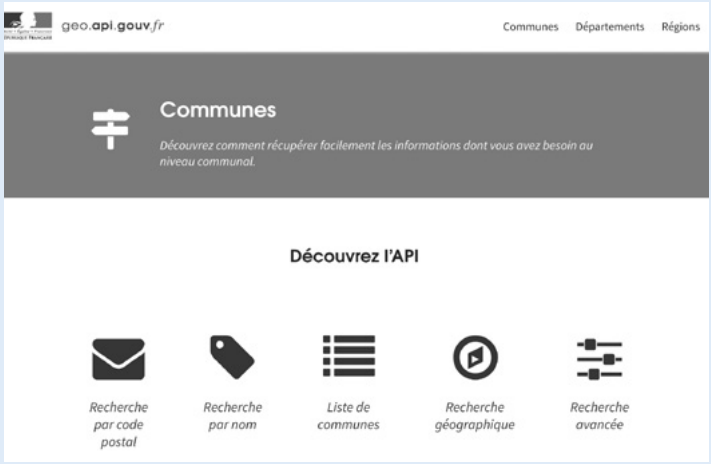
Ce qui n'était jusqu'en 2016 que la « passerelle INSPIRE » de data.gouv.fr (du nom de la directive INSPIRE), chargée de moissonner les données

géographiques ouvertes et de les cataloguer sur data.gouv.fr, est devenu en 2017 geo.data.gouv.fr, la plateforme de diffusion de données géographiques de data.gouv.fr. Cette plateforme spécialisée permet de présenter les données géographiques de manière plus détaillée que la présentation générique disponible sur data.gouv.fr, en tirant pleinement parti de la richesse des métadonnées géographiques propres à la directive INSPIRE. Elle référence aujourd'hui plus de 100 000 jeux de données géographiques dont 20 000 téléchargeables en open data issus de 118 portails locaux ou nationaux.

Trois sous-domaines ont été plus particulièrement travaillés grâce à la dynamique du lancement du service public de la donnée : la géographie administrative (régions, départements, communes), les adresses, et le plan cadastral, ces deux derniers avec leur plateforme dédiée, adresse.data.gouv.fr et cadastre.data.gouv.fr

En parallèle a été déployée une interface de programmation applicative geo.api.gouv.fr pour permettre aux réutilisateurs de s'appuyer sur ces données de référence aussi simplement que possible. Elle permet d'interroger une version toujours à jour de la géographie administrative (sur geo.api.gouv.fr), et de géocoder, vérifier et normaliser toute adresse postale (sur api-adresse.data.gouv.fr). Plus d'un milliard d'appels ont été effectués sur ces API en 2017, démontrant l'appétence des réutilisateurs pour une mise à disposition sous forme d'API en complément des données brutes en téléchargement.

Des API pour faciliter l'intégration des données géographiques



Depuis 2014, Etalab développe en partenariat avec l'Institut national de l'information géographique et forestière (IGN), La Poste et l'association OpenStreetMap France la Base Adresse Nationale (BAN), une base de données géolocalisée des adresses en France.

Au-delà de la fourniture des données, Etalab a développé depuis 2014 un moteur de géocodage, Addok, spécifiquement optimisé pour la recherche d'adresse et une API de géocodage en masse librement accessible. La volumétrie d'usage de ces outils démontre leur utilité et leur pertinence. Ainsi, en 2017, l'API de géocodage de la Base Adresse Nationale a connu une très forte progression de son audience avec plus d'un milliard de requêtes et 11,8 millions de visiteurs uniques sur les dix premiers mois de l'année contre 443 millions de requêtes et 5,5 millions de visiteurs uniques sur l'ensemble de l'année 2016.

D'importants sites d'e-commerce utilisent directement ce service pour améliorer la qualité des données de livraison.

Partie intégrante de la verticale géo, l'API Géo permet d'accéder aux données du Code officiel géographique (COG) et aux limites administratives des communes, départements et régions. Comme l'API de géocodage, l'API Géo est très utilisée et a comptabilisé près de 200 millions de requêtes au cours de l'année 2017.

La verticale des données des entreprises

Suite à l'ouverture du répertoire des entreprises et de leurs établissements de l'INSEE (répertoire Sirene), diffusé sur data.gouv.fr depuis le 4 janvier 2017, puis à l'entrée de ce jeu de données de référence sur l'identité des entreprises dans le service public de la donnée le 1er avril 2017, Etalab a initié la mise en œuvre d'une verticale de data.gouv.fr déliée aux données des entreprises sur le modèle de la verticale des données géographiques.

Une mutualisation des équipes et des efforts a été opérée avec une partie du programme « Dites-le-nous une fois », au sein duquel une API de diffusion d'information d'identité des entreprises avait déjà été développée, à destination des seules administrations.

Cette « API Entreprise », techniquement redimensionnée pour pouvoir être en partie ouverte à tout réutilisateur, est devenue en 2017 entreprise.api.gouv.fr, premier pilier de cette nouvelle verticale entreprise. Elle est à fin 2017 utilisée par une centaine d'administrations pour simplifier des démarches administratives et éviter de demander des pièces justificatives. Ce sont sur le mois de décembre 2017 plus de 1,2 million d'informations qui ont été obtenues *via* cette API et n'ont donc pas été redemandées aux entreprises.

La verticale des données de transport

Le ministère des Transports et l'incubateur de services numériques de la DINSIC ont convenu à l'été 2017 de développer ensemble, dans le cadre d'une startup d'État, une plateforme d'accès aux données ouvertes de transport prévues par l'article L. 1115-1 du Code des transports, qui préfigure le point d'accès national demandé par le règlement européen adopté le 31 mai 2017 : transport.data.gouv.fr. À fin 2017, les stations, arrêts et horaires théoriques des transports en commun des agglomérations de Brest, Toulouse et Grenoble ont été intégrées au pilote.

Le soutien de l'écosystème des données publiques

L'écosystème des données publiques regroupe à la fois les producteurs de données – État, collectivités, acteurs privés et associatifs – et les réutilisateurs.

La mission Etalab a participé en 2017 au financement de l'expérimentation Opendatalocale menée par Opendatafrance. Neuf territoires d'expérimentation ont ainsi pu bénéficier d'un accompagnement pour mettre en œuvre le principe d'ouverture par défaut. L'ensemble des collectivités de plus de 3 500 habitants et de plus de cinquante agents sont aujourd'hui concernées par cette disposition de la loi pour une République numérique.

Dans le cadre d'Opendatalocale plusieurs outils sont mis à disposition de l'écosystème :

– un socle commun de données locales : afin d'encourager l'ouverture homogène sur le territoire, des standards de données ont été définis en lien avec les réutilisateurs et les éditeurs de solutions informatiques ;

– des outils de sensibilisation et d'accompagnement sous la forme de documents, de clauses-types de marchés publics notamment ;
– un dispositif de formation des formateurs à l'open data sous la forme d'un jeu sérieux.

La préfiguration d'un réseau des administrateurs ministériels des données

Conformément aux recommandations du précédent rapport de l'administrateur général des données, plusieurs administrations centrales – et en particulier le ministère de l'Intérieur, celui de la Transition écologique et solidaire ou encore la Direction générale des finances publiques de Bercy – ont nommé en leur sein des administrateurs ministériels des données (AMD).

Le réseau se réunit sous la houlette de l'administrateur général des données.

Le **ministère de la Transition écologique et solidaire** a créé le rôle de « superviseur des données » dès 2016. La fonction a été attribuée à Laurence Monnoyer-Smith, commissaire générale au développement durable. L'attribution de cette fonction au commissariat général au développement (CGDD) durable se justifie par la capacité de son service statistique et de sa mission d'information géographique, ainsi que par sa mission d'animation du réseau scientifique et technique.

Au **ministère de l'Intérieur**, l'administrateur ministériel des données, Daniel Ansellem a pris ses fonctions le 1er juillet 2016. Le ministère de l'Intérieur a également désigné des administrateurs des données au sein de chaque direction métier et chaque opérateur sous tutelle du ministère de l'Intérieur (Agence nationale des titres sécurisés et Agence nationale de traitement automatisé des infractions). L'administrateur des données du ministère de l'Intérieur participe également au programme « Entrepreneur d'intérêt général » avec deux projets en 2017 (Carte AV et MatchID) et deux nouveaux projets en 2018

À la **direction générale des Finances publiques**, l'administrateur des données, Lionel Ploquin, a été nommé en août 2016. Il est affecté auprès du directeur du service à compétence nationale Cap Numérique en charge des projets de transformation numérique de cette administration. Ce positionnement reflète la conscience qu'a la DGFiP du rôle essentiel que joue la maîtrise des données et de leur valorisation dans la transformation numérique du secteur public.

Le **ministère de l'Agriculture** a créé le poste délégué général au numérique et à la donnée au sein du secrétariat général en décembre 2017.

3. Exploiter les données pour améliorer l'action publique

Au-delà de la production des données essentielles et de la circulation optimale des données, il est essentiel que **l'administration puisse exploiter au mieux ses données** pour améliorer l'efficacité de l'action publique.

En complément de la statistique publique, qui utilise les données pour produire de la connaissance, la **data science** dans l'administration utilise les données pour développer de nouveaux services opérationnels visant à améliorer les processus métiers.

Depuis la parution du premier rapport de l'administrateur général des données, plusieurs réalisations menées par l'équipe interne de data-scientists ont permis de **démontrer le bénéfice de cette approche** en appui des politiques publiques, qu'il s'agisse de lutter contre le chômage, de repérer au plus tôt les entreprises en difficulté ou encore d'outiller les services en charge de lutter contre les vols de voiture ou les cambriolages.

Le dispositif « Entrepreneur d'intérêt général » contribue lui aussi à généraliser les nouveaux usages de la donnée au sein des ministères.

Lutter contre le chômage en fournissant de nouveaux services aux demandeurs d'emploi

Dans le domaine des politiques de l'emploi, les projets **La Bonne Boîte**, La Bonne Formation ou encore Bob-Emploi montrent comment l'utilisation des données administratives permet de développer de nouveaux services numériques pour accompagner les demandeurs d'emploi dans leurs recherches.

Développé par Pôle Emploi en collaboration avec l'incubateur de services numériques de la DINSIC et les data-scientists de la mission Etalab, La Bonne Boîte permet aux demandeurs d'emploi d'accéder au marché caché de l'emploi et d'envoyer des candidatures spontanées ciblées auprès d'**entreprises avec une forte probabilité d'embauche**.

Le service repose sur un algorithme de prédiction de la probabilité d'embauche des entreprises dans un secteur et une région donnés. Très simple d'utilisation, il est aujourd'hui utilisé par près de 70 000 visiteurs uniques par mois et 9 000 conseillers de Pôle Emploi. **70 % des utilisateurs déclarent avoir trouvé au moins une entreprise pertinente à contacter**[21].

Bob-Emploi est un coach numérique pour accompagner les demandeurs d'emploi développé par l'association Bayes Impact. L'outil utilise des algorithmes de recommandation développés par une collaboration entre Etalab et Bayes Impact. Lancé en novembre 2016, Bob-Emploi compte

21 https://labonneboite.pole-emploi.fr/stats

un an plus tard plus de 115 000 utilisateurs inscrits et 86 % des utilisateurs considèrent que Bob-Emploi leur a été utile[22].

Repérer au plus tôt les entreprises qui vont rencontrer des difficultés

Il est admis que l'intervention publique auprès des entreprises en difficulté est d'autant plus efficace qu'elle est précoce. Repérer les entreprises qui vont rencontrer des difficultés est l'objectif du produit **Signaux Faibles** développé avec plusieurs partenaires publics en Bourgogne-Franche-Comté[23].

Il permet de **détecter en amont** les entreprises en difficulté potentielle pour permettre au Commissaire au redressement productif (CRP) d'aider ces entreprises avant que la situation ne devienne critique. L'algorithme développé en croisant les données des URSSAF et les données des DIRECCTE a déjà permis de déclencher vingt-cinq visites d'entreprises. **Sur ces vingt-cinq visites, seize visites ont été jugées utiles** par le commissaire au redressement productif**.**

Développer des outils d'aide à la décision pour les services de la Sécurité intérieure

Prévenir et lutter contre le vol de véhicules

Pour aider les forces de l'ordre à agir au mieux face aux vols de véhicules, Etalab a développé en collaboration avec le Service des technologies et des systèmes d'information de la Sécurité intérieure (ST(SI) ²) et la Direction centrale de la sécurité publique (DCSP), **MapVHL**, un outil d'aide à la décision permettant de connaître l'historique des vols et des découvertes de véhicules.

Initialement, l'objectif était de développer un modèle prédictif permettant d'identifier les futures zones à risque. La confrontation avec les forces de terrain a permis de montrer que le premier besoin, avant d'avoir un modèle prédictif fin des zones dangereuses, était d'avoir **une connaissance précise de l'historique des vols** à travers une interface simple et intuitive. Les retours du terrain ont aussi conduit à développer une cartographie des lieux de découvertes de véhicules[24].

Après une première démonstration avec les gendarmes de la compagnie de Compiègne (Oise), c'est finalement à la Direction départementale de la sécurité publique (DDSP) de Beauvais que l'outil a pu être testé par les forces de l'ordre. Il a été couplé en déploiement de tablettes et a permis aux policiers de la brigade anticriminalité de tester l'outil en mobilité.

---

22 https://www.bob-emploi.fr/transparence

23 En partenariat avec la DIRECCTE Bourgogne-Franche-Comté, l'URSSAF Bourgogne et l'URSSAF Franche-Comté.

24 Pour un retour d'expérience détaillé voir le billet de blog de l'AGD : https://agd.data.gouv.fr/2018/01/12/predire-les-vols-de-voitures/

Lutter contre l'insécurité routière

Afin de mieux connaître la cartographie des accidents de la route et de pouvoir mieux orienter l'action des forces de sécurité, la mission de valorisation des données au sein de la Mission de gouvernance ministérielle des systèmes d'information et de communication (MGMSIC) du ministère de l'Intérieur a développé dans le cadre du programme « Entrepreneur d'intérêt général 2017 », l'outil **CarteAV** (cartographie accidents verbalisation)[25].

Le rapprochement des données des accidents de la circulation et des données sur l'historique des procès-verbaux permet d'**identifier les zones dangereuses** pour la circulation où les forces de sécurité interviennent peu et à l'inverse les zones où il y a un **nombre important de verbalisations** alors que le danger semble moins important.

CarteAV a été développé en dix mois par une équipe de deux data-scientists/développeurs et testé en cycle court par une vingtaine de services de police.

Améliorer la qualité du système national de permis de conduire

Jusqu'à maintenant, il n'y avait pas de lien entre le système national de permis de conduire et le Répertoire national d'identification des personnes physiques (RNIPP). Par conséquent, il n'était pas possible d'identifier dans la base de données des permis de conduire les personnes décédées. Cette situation a conduit dans la pratique à retirer des points sur le permis de personnes décédées et a rendu possible un système de fraude.

Le ministère de l'Intérieur a développé un outil, **MatchID**, permettant d'apparier le RNIPP avec la base de données des permis de conduire sur la base des noms prénoms et date de naissance. L'outil utilise des techniques d'appariement flou (*fuzzy matching*) et des méthodes d'apprentissage automatique permettant à l'utilisateur d'entraîner l'algorithme pour améliorer la qualité de l'appariement.

Au-delà de l'exemple de l'appariement de la base de données des permis de conduire, MatchID peut être réutilisé dans un grand nombre de cas pour améliorer la qualité de bases de données en détectant les doublons ou apparier différentes bases de données sans identifiant commun[26].

Faciliter le travail de l'administration en rapprochant automatiquement des bases de données

L'Office central de lutte contre le travail illégal (OCLTI) a pour mission de lutter contre le travail illégal et notamment de détecter les fraudes au détachement intra-européen de travailleurs. Dans le cadre de cette mission, elle est souvent amenée à rapprocher des bases de données pour identifier des victimes de fraude transnationale. Typiquement, à partir d'une liste

25 Le Code source est disponible sur Github : https://github.com/eig-2017/cartav.
26 https://github.com/matchID-project/

de salariés, elle doit rechercher les noms de ces salariés dans les bases de données de la sécurité sociale pour vérifier qu'ils y sont bien inscrits.

En utilisant des méthodes d'appariement flou et des logiciels libres existants[27], l'équipe de data-scientists a montré qu'il était possible d'obtenir des résultats très satisfaisants avec des méthodes automatiques[28]. La mise en place de manière systématique de ces méthodes permettrait d'économiser des jours de travail humain et d'améliorer l'efficacité de la lutte contre la fraude aux travailleurs détachés.

Le programme « Entrepreneur d'intérêt général »

Lancé fin 2016, le programme « Entrepreneur d'intérêt général » constitue un programme d'innovation original. Chaque administration peut proposer un défi à relever. Pour chaque défi, une petite équipe d'une à trois personnes se donne dix mois pour relever le défi en travaillant avec un mentor dans l'administration.

Les lauréats sont sélectionnés par un jury de personnalités qualifiées du numérique et d'agents des administrations. Ils sont accueillis par les administrations, ont accès à des bases de données, et sont suivis par des mentors de haut niveau et par l'équipe d'Etalab.

Douze défis ont été sélectionnés dans le cadre de la seconde promotion :

- SocialConnect, porté par le commissariat général à l'égalité des territoires, vise à développer une plateforme collaborative pour mettre en réseau et en valeur l'écosystème de l'innovation sociale dans les territoires.

- Signaux Faibles, porté par la DIRECCTE Bourgogne-Franche-Comté, vise à développer les outils pour repérer les entreprises en difficulté.

- PréviSecours, porté par le ministère de l'Intérieur, vise à développer un modèle prédictif des secours aux personnes pour aider les pompiers à mieux allouer leurs moyens.

27 Ici la bibliothèque python dedupe https://github.com/dedupeio/dedupe
28 https://agd.data.gouv.fr/2016/11/22/rapprocher-deux-bases-donnees/

- Lab Santé, porté par le ministère des Solidarités et de la Santé, vise à analyser les données du Système national des données de santé (SNDS).
- Hopkins, porté par le ministère de l'Action et des Comptes publics, vise à détecter la fraude financière.
- Gobelins, porté par le ministère de la Culture, vise à révéler la richesse du Mobilier national.
- PrédiSauvetage, porté par le ministère de la Transition écologique et solidaire, vise à identifier l'origine des accidents en mer pour mieux les prévenir.
- dataESR, porté par le ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, vise à développer une plateforme d'analyse des données de l'enseignement supérieur et de la recherche.
- CoachÉlèves/Assistprof, porté par le ministère de l'Éducation nationale, vise à construire des coachs numériques pour accompagner les élèves et les profs dans leurs apprentissages.
- Brigade numérique, porté par le ministère de l'Intérieur, vise à construire un accueil numérique de la gendarmerie au service des citoyens.
- B@liseNAV, porté par le ministère des Armées, vise à réaliser une carte maritime augmentée pour rendre la navigation plus sûre.
- ArchiFiltre, porté par le ministère des Solidarités et de la Santé, vise à développer des méthodes pour filtrer automatiquement les données non structurées destinées à être archivées.

Un collège scientifique pour accompagner les administrations à choisir des prestataires

Lorsqu'une société de services propose une solution commerciale en datasciences à une administration, celle-ci n'a pas toujours les moyens d'évaluer la pertinence scientifique de la démarche de l'entreprise et peut ainsi avoir du mal à évaluer si la solution proposée est pertinente ou non.

Pour accompagner les administrations et faire en sorte qu'elles ne soient pas amenées à utiliser des algorithmes comme des boîtes noires qu'elles ne comprennent pas, Etalab propose de réunir un collège scientifique composé de chercheurs et d'universitaires reconnus, qui peuvent aider l'administration à mieux comprendre et mieux évaluer la solution proposée par le prestataire.

Concrètement, le collège scientifique peut interroger le prestataire sur l'algorithme utilisé, évaluer sa pertinence dans le contexte de l'application ou encore interroger le prestataire sur la base d'apprentissage choisi pour entraîner l'algorithme.

La première réunion de ce collège scientifique a lieu en décembre 2017 pour conseiller la direction centrale de la police judiciaire (DCPJ) dans le choix d'une solution commerciale d'un algorithme de ciblage géographique permettant de guider les enquêteurs spécialisés dans les crimes sériels.*

*https://agd.data.gouv.fr/2018/01/19/un-appui-scientifique-aux-administrations/

Deuxième partie

La donnée comme infrastructure essentielle

Le décret de création de la fonction d'administrateur général des données prévoit que le rapport remis au Premier ministre, au-delà d'un état d'avancement de la politique de la donnée, détaille les « **évolutions de l'économie de la donnée** » et aide à projeter l'action de l'État dans ce domaine.

La deuxième partie du rapport est ainsi consacrée à l'analyse d'un thème central : **la donnée comme infrastructure.** En proposant une approche comparée des initiatives de plusieurs pays européens, elle milite pour la création d'une véritable infrastructure de données. L'État *peut* et *doit* jouer un rôle central, dans **une logique d'État plateforme**.

1. La donnée doit être considérée comme une infrastructure

Uber, Airbnb, Booking : chacun dans leur domaine (transports, tourisme, hôtellerie), ces acteurs illustrent le **basculement des modèles d'affaires**. La détention d'un actif matériel a longtemps constitué un élément clé pour la structuration de ces marchés. Pour prétendre être un acteur de l'hôtellerie, par exemple, il fallait posséder une capacité d'investissement dans des actifs immobiliers. La place déterminante qu'occupe aujourd'hui dans ce secteur un acteur tel que Booking.com illustre bien que ce paradigme de la détention d'actifs matériels est largement remis en cause.

Dans des environnements économiques dominés par la notion de plateformes, **la possession et l'exploitation d'actifs immatériels** font la différence. Au premier rang de ces actifs, bien avant la marque, se situe les données et la capacité à les traiter. Si certains de ces acteurs possèdent aujourd'hui un pouvoir de marché important, ils le doivent en grande partie à la masse de données qu'ils ont été capables de se constituer dans leur courte existence.

On constate ainsi un **décalage croissant** entre l'immatérialité des ressources mobilisées pour rendre un service et la matérialité des effets de ce même service.

L'impact de ces activités, en ligne, sur le réel est majeur. La plupart des grandes métropoles sont aujourd'hui confrontées à ces impacts. New York, Barcelone ou encore Paris mettent en place des régulations pour limiter l'assèchement de l'offre locative dans les zones touristiques sous l'effet des plateformes de location courte durée. À San Francisco, Uber transporte autant de passagers que le réseau de transport public de la ville, et dans certaines villes, l'entreprise de véhicules de tourisme avec chauffeur se présente comme une offre de transport complémentaire de l'offre publique. Une partie importante du succès de ces initiatives privées – qui interpellent et parfois bousculent l'initiative publique – tient à l'utilisation des données.

Deux traits permettent de caractériser **le rôle que jouent les données** dans les stratégies des plateformes. Le premier tient à la collecte : *tout ce qui peut être mis en donnée l'est*. **La collecte est massive, et continue**. Aucune interaction avec l'utilisateur, aucune requête, aucun clic n'échappe à cette mise en données. L'historique de consommation, la localisation de l'utilisateur, mais aussi les notations alimentent ainsi en permanence les algorithmes. Le second trait distinctif tient à l'exploitation des données, elle aussi continue et massive. Dans une entreprise ***data-driven***, les données sont utilisées à chaque phase de la conception et de la fourniture du service. Mieux, grâce aux données, il devient difficile de séparer les phases de conception et de production, à l'image des tests A/B qui optimisent en temps réel les pages (ou les titres de presse en ligne) d'un site Web.

L'État est **concerné** à plus d'un titre par les débats autour des plateformes et de l'usage des données. Dans son rôle de régulateur tout d'abord : début

2018 le Parlement étudiera l'adaptation de la loi « Informatique et Libertés » aux dispositions introduites par le Règlement européen sur la protection des données (RGPD). De même, la France joue un rôle moteur en Europe pour impulser la discussion sur la taxation des entreprises du numérique.

L'État peut aussi apporter une réponse, **offensive et non pas défensive,** aux défis et opportunités soulevés par les plateformes en constituant une infrastructure publique de données.

L'infrastructure publique du XXIe siècle

« Data is a new class of public infrastructure for the 21st century. It is all around us and easy to miss. We need to view it as an infrastructure that is as fundamental to modern society as power and transport, and which requires investment, curation and protection[1]*. »* (Nigel Shadbolt, vice-président de l'Open Data Institute).

Le développement d'un pays est étroitement lié à la présence d'une **infrastructure performante et de qualité**, qu'il s'agisse de routes, de lignes ferroviaires, de réseaux d'énergie ou de télécommunications. L'État a d'ailleurs longtemps consacré une partie importante de ses investissements à construire et maintenir ces infrastructures. À titre d'illustration, 1 km d'autoroute représente un investissement de 6 millions d'euros, et 1 km de ligne ferroviaire à grande vitesse 16 millions d'euros. En moyenne, les pays membres de l'Union européenne y consacrent plus d'**un tiers de leurs investissements publics**[2]. Ces infrastructures contribuent à aménager le territoire et facilitent les échanges et les déplacements des biens et des personnes. Les externalités générées sont très largement positives. *A contrario* la défaillance des infrastructures est l'un des facteurs explicatifs du retard de développement de certains pays.

Il faut aujourd'hui considérer les données comme l'une de ces **infrastructures *essentielles* et *critiques***. Essentielles car, dans une économie de l'information, l'accès à la donnée de référence fiable et à jour est la **condition du développement** des services numériques. Critiques car il faudra s'assurer que la fourniture de ces données ne puisse être interrompue, qu'il s'agisse de défaillances involontaires ou d'actes malveillants. En ce sens, il serait possible de considérer que ces infrastructures font partie des activités d'importance vitale[3]. Le développement des villes intelligentes ou des *smart grids* repose en grande partie sur la sécurisation de la mise à disposition et de l'accès aux données.

---

1 « Les données constituent un nouveau type d'infrastructure publique pour le XXIe siècle. Elles sont tout autour de nous et pourtant facile à manquer. Nous devons considérer l'infrastructure de données comme une infrastructure aussi fondamentale pour la société moderne que l'énergie et les transports, et qui nécessite des investissements, des services de conservation et de protection. »

2 Infrastructure in the EU : Developments and Impact on Growth, Commission européenne, 2014.

3 Selon la définition du Secrétariat général pour la Défense et la sécurité nationale (SGDSN) : « *Parce qu'elles concourent à la production et à la distribution de biens ou de services indispensables à l'exercice de l'autorité de l'État, au fonctionnement de l'économie, au maintien du potentiel de défense ou à la sécurité de la Nation, certaines activités sont considérées comme "d'importance vitale"*. »

Plusieurs pays européens ont saisi la nécessité de considérer la donnée comme une infrastructure publique essentielle, au même rang que les infrastructures physiques.

Depuis 2013, le gouvernement fédéral allemand intègre un ministère en charge des transports et des infrastructures numériques. Les investissements en matière d'infrastructures physiques et informationnelles sont gérés de manière intégrée, par exemple sur le sujet de la voiture connecté[4]. Outre-manche, la Commission nationale des infrastructures (*national infrastructure commission*), qui a pour objet de formaliser la stratégie nationale à long terme et d'orienter les investissements dans le domaine des infrastructures critiques pour le pays, a intégré la problématique des données à son périmètre d'action. De même, l'OCDE a souligné en 2015 (« Data-driven innovation report ») l'importance de telles infrastructures pour le développement économique et social.

2. Les objectifs d'une infrastructure de données

Quelles que soient les différences d'approche au niveau européen (nous détaillerons ci-après les démarches initiées au Royaume-Uni, au Danemark et en Estonie pour les comparer aux initiatives françaises), il est frappant de constater que **les constats, les objectifs** sont partagés. Le manque de circulation et d'exploitation des données, leur indisponibilité représentent une perte nette pour l'ensemble de la société.

Des données qui ne circulent pas, ou qui sont sous-utilisées sont des données dont on ne tire pas la totalité de **la valeur d'usage**. Elles peuvent même tendre à perdre de leur valeur au fil du temps, faute d'être confrontées à leurs utilisateurs et de bénéficier d'un retour qui permet leur amélioration. Leur qualité se **détériore** d'autant plus rapidement.

Les conséquences financières de l'absence d'une infrastructure de données de qualité sont bien réelles. Le non-acheminement de courriers postaux dus à une erreur d'adressage représente un surcoût de l'ordre de 300 millions d'euros par an.

Le rapport Fouilleron a identifié plusieurs cas où des administrations dupliquent ou recopient des bases de données existantes, voire même **créent leurs propres bases** faute de pouvoir accéder librement, gratuitement et de manière sécurisé aux données produites par d'autres administrations.

Les coûts de la mauvaise circulation et de la non-qualité des données peuvent se détailler de la manière suivante :

– les pertes directes et indirectes liées à l'utilisation de **données inexactes** ;
– le maintien de **bases de données redondantes** et le coût de double-saisie quand la base de référence n'est pas diffusée à tous ceux, acteurs publics et privés, qui ont en besoin, c'est par exemple le cas des collectivités

4 « Policies for mobility and modernity », Federal Ministry of Transport and Digital Infrastructure (BMVI), 2014.

qui n'avaient pas accès, jusqu'à une date récente, à la base officielle des associations alors qu'elles sont le premier financeur du secteur associatif ;
– le fonctionnement en **mode dégradé** faute d'accès à la donnée la plus récente et/ou la plus précise ;
– les **coûts de transaction** liés à la recherche, l'acquisition et le traitement de données publiques, qui peuvent être très élevés pour des industries ayant un besoin critique de données.

Une infrastructure pour permettre la meilleure exploitation des données

Une route sert à faciliter les échanges et le transport des biens et des personnes. Un réseau de télécommunications vise à faire circuler l'information et à faciliter la coordination des individus. Une infrastructure de données a elle aussi une finalité : permettre la **meilleure exploitation possible des données**.

Elle s'adresse bien sûr en premier lieu à l'administration elle-même, mais concerne *in fine* l'ensemble des acteurs (entreprises, associations, société civile) qui utilisent des données publiques. Voilà **l'objectif qui est partagé au niveau de l'ensemble des pays européens** qui se sont engagés dans des démarches de construction d'infrastructure de données.

Comment faire en sorte que l'ensemble des acteurs s'empare des données mises à disposition ? Quels sont **les engagements** que la puissance publique doit prendre ? Pour le comprendre, Etalab a mené une consultation publique [5] auprès des utilisateurs potentiels des données de référence. Cette consultation, à laquelle 160 répondants (acteurs publics et privés, associations) ont contribué, a notamment permis d'identifier précisément les attentes, en particulier sur les critères de qualité des données de référence.

5 Consultation menée du 29 septembre au 20 octobre 2016, synthèse des résultats disponible en ligne : www.etalab.gouv.fr/consultation-spd

Des données à jour, disponibles et facilement réutilisables

La **fraîcheur** apparaît ainsi très nettement comme la principale dimension attendue (mise à jour des données, délai entre la survenance d'un fait, par exemple l'enregistrement d'une association, et son apparition dans la base diffusée). La **haute disponibilité** de l'infrastructure (de l'ordre de 99,5 % mensuels) est un autre élément attendu. L'utilisation de **standards ouverts** (second critère le plus fréquemment cité) a été introduite dans le Code des relations entre le public et les administrations à l'occasion de la loi pour une République numérique.

La complétude, l'exactitude et la fraîcheur constituent des dimensions classiques de **la qualité des données**. Mais les utilisateurs attendent aussi de la traçabilité sur le processus de production et de mise à disposition, la possibilité d'interagir avec le producteur (pour signaler une erreur ou proposer une amélioration d'une donnée) ou encore la transparence sur les indicateurs de qualité des données et de leur mise à disposition.

« Des données sur lesquelles on peut compter »

L'une des caractéristiques d'une infrastructure c'est d'**être invisible**.

N'importe quel individu qui tourne un robinet s'attend à ce que l'eau potable y coule instantanément. Seule l'interruption temporaire du service d'eau (ou d'électricité/télécommunications/transports) nous permet de prendre conscience de l'ensemble des efforts et des moyens qui ont dû être mis en œuvre pour fournir le service de manière instantanée.

La fourniture de données publiques n'a pas pour le moment atteint ce niveau de qualité. Certaines données ne sont pas suffisamment mises à jour. D'autres sont peu ou mal documentées. Le schéma des données peut être modifié par leur producteur pour ses besoins internes et sans que les réutilisateurs aient été consultés ni même informés des changements à venir.

Cela est d'autant plus dommageable que l'infrastructure de données partage pourtant les mêmes caractéristiques que les autres infrastructures. Ses utilisateurs expriment le même **besoin de confiance dans l'infrastructure**[6] : je dois connaître *a priori* la qualité de la mise à disposition des données pour être en mesure d'appuyer mon service ou mon analyse sur ces éléments.

Si le service fourni est interrompu ou dégradé, si la qualité se dégrade alors la **confiance** des utilisateurs est rompue et l'infrastructure n'aura pas atteint ses objectifs.

6 Un utilisateur compare notamment la mise à disposition des données de référence avec la standardisation existant dans le domaine hôtelier (« *quand je choisis un hôtel de marque Novotel, je m'assure que mon expérience client sera identique dans tous les pays. Il n'y a pas de surprises et je sais précisément ce que je suis en droit d'attendre* »).

3. *Benchmark* des initiatives européennes

Les **objectifs** de l'infrastructure de données et les attentes sont partagées par l'ensemble des pays engagés sur cette voie. Mais les **manières d'atteindre** les objectifs, et *in fine* de **construire** cette infrastructure diffèrent parfois. Afin de mettre en perspective les différents modèles existants nous proposons d'analyser et comparer les initiatives menées au **Royaume-Uni**, au **Danemark** et en **Estonie** avec les initiatives nationales.

GOV.UK Registers (Royaume-Uni)

En 2013, le Gouvernement britannique a lancé le projet **National Information Infrastructure** (NII), en réponse aux recommandations de la *Shakespeare Review*. De 2013 à 2015, le *Cabinet Office* a posé les principes généraux de la NII[7] et a commencé à identifier les jeux de données concernés. En 2015, la présidente du open data *User Group* (ODUG-UK, la structure représentant les utilisateurs des données ouvertes[8]) a rendu un rapport critique sur la mise en œuvre du NII, soulignant notamment le manque de priorisation des efforts. La première liste des données publiées sur data.gov.uk comprenait 233 jeux de données et le suivi de leur qualité n'était pas satisfaisant (une partie d'entre eux n'était plus mis à jour quelques mois après leur première publication).

Parmi les recommandations de l'open data *User Group* figurait l'idée que le Gouvernement devrait dans un premier temps se concentrer sur quelques jeux-clés et notamment les **nomenclatures**.

GOV.UK Registers, dont l'ambition est moindre que celle du NII, s'inscrit donc dans cette lignée[9].

GOV.UK Registers est porté par le *Government Digital Services* (GDS) du Royaume-Uni, qui le présente comme l'un des composants de l'approche *Government-as-a-platform*, au même niveau de GOV.UK Notify (gestion des notifications), *Pay* (gestion des paiements) ou *Verify* (brique d'authentification).

« La source de la donnée la plus fiable dans son domaine »

Quatorze registres sont aujourd'hui proposés en ligne, et quarante-cinq autres sont déjà identifiés comme des candidats potentiels au statut de registres. Chacun doit être « **la source de données la plus fiable**[10] » dans son domaine.

---

7 Voir https://data.gov.uk/consultation/national-information-infrastructure-prototype-document/what-national-information

8 La structure open data *User Group* a depuis été dissoute.

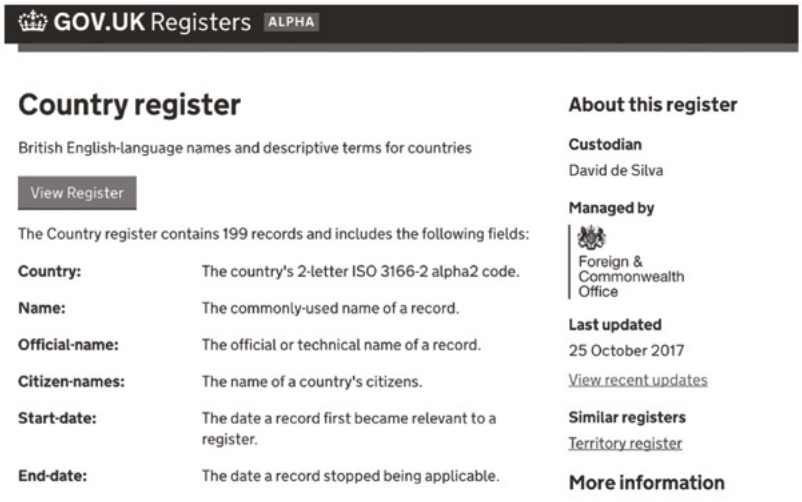
9 À ce jour, et d'après les différents contacts que nous avons pu interroger au Royaume-Uni, le projet de *National Information Infrastructure* est en *stand-by*.

10 "each register is the most reliable list of its kind" – What registers are ? https://registers.cloudapps.digital/

Ces jeux de données contiennent en moyenne *quelques dizaines* ou *centaines* d'enregistrements chacun et concernent principalement des **nomenclatures** concernant l'organisation administrative ou territoriale (des pays aux comtés), ainsi que certains équipements (prisons) ou infrastructures publiques (organisations en charge des réseaux de drainage).

Les registres sont fournis sous **plusieurs formats** en téléchargement et *via* des API. Le schéma d'API de la plateforme Registers est commun à l'ensemble des registres ce qui en facilite l'intégration par des développeurs tiers.

Exemple : Country Register, la liste officielle du Foreign and Commonwealth Office



Une gouvernance centralisée, avec une responsabilisation forte des producteurs

Le ***Government Digital Service*** (GDS) est chargé de maintenir la plateforme des registres. À ce titre, il a défini une démarche unique pour faire reconnaître un nouveau registre et l'intégrer. Chaque ministère peut **solliciter** auprès du GDS la reconnaissance de l'une de ses données comme un registre officiel. Le GDS a fixé les **critères d'éligibilité** : le registre ne contient aucune donnée à caractère personnel, il s'agit de données brutes (*Raw Data*) et non de données dérivées ou de statistiques, rien ne fait obstacle à sa publication en open data.

La **décision finale d'intégrer** ou non un nouveau registre dans la liste officielle revient au **GDS**, qui tient notamment compte de la demande des utilisateurs mais aussi du profil du producteur[11].

Le producteur du registre doit désigner nominativement un conservateur ou gardien (*custodian*) parmi ses agents. **L'engagement demandé est personnel** : son nom figurera sur la page du registre. Il est notamment responsable de la mise à jour des données, il répond aux questions des utilisateurs et constitue le point de contact du GDS. En cas d'absence momentanée ou durable, le *custodian* doit désigner son remplaçant. L'ensemble des *custodians* désignés fait l'objet d'une formation par l'équipe du GDS en charge des registres.

Basic Data – Grunddata (Danemark)

Le programme Grunddata est inscrit dans la **stratégie numérique du Danemark**. Il a été initié en 2012 par le gouvernement et les communes danoises, les régions l'ont rejoint dès 2013. Il est aujourd'hui porté par l'Agence pour la numérisation rattachée au ministère des Finances.

Le Danemark a une longue tradition de numérisation des registres et bases administratives, et dispose d'une législation qui **facilite le croisement** de bases de données notamment ***via* le numéro unique d'identification des individus**.

À titre de comparaison : en France l'utilisation du numéro d'inscription au Répertoire national d'identification des personnes physiques (le « numéro de sécurité sociale ») pour le croisement de bases de données est très fortement encadrée par la loi.

Grunddata est un **programme** très large qui couvre l'ensemble du cycle de vie des données de référence, de leur production à leur diffusion et leur financement. Les principales actions témoignent de cette ambition :

- l'identification et la **montée en qualité** des principaux registres ;
- la **convergence** des principales bases de données de référence, avec un important travail de modélisation des données et des liens entre elles ;
- la mise en place d'un **Distributeur de données** (*Data Distributor*) qui a vocation à remplacer les systèmes de diffusion actuels ;
- la mise en place d'une gouvernance forte, pilotée par le ministère des Finances et qui dispose d'**un levier budgétaire propre**.

11 Le GDS se garde notamment le droit de juger si une administration est la plus pertinente pour maintenir un registre, dans le cas où plusieurs administrations sont susceptibles de se déclarer compétentes.

Une organisation par thématique

Grunddata est organisé selon six zones qui correspondent chacune à des thématiques de données :

– l'**immobilier** et le foncier : le cadastre, mais aussi les titres de propriété foncière ;
– les **individus** : les données du Registre civil (identifiant personnel) c'est-à-dire le nom, l'adresse, l'état civil, la filiation (enfants et parents), les droits liés à la citoyenneté ;
– les **entreprises** : les données d'enregistrement des entreprises, mais aussi des données sur leurs résultats, leurs effectifs ;
– les **adresses**, les routes et les divisions administratives ;
– les **cartes** et les données géographiques ;
– l'**eau et le climat** : cartes et modèles hydrographiques, données météorologiques.

Chacune de ces zones a été **confiée à un ou plusieurs producteurs**, sous le contrôle du ***Basic Data Board*** (*cf. infra* sur la gouvernance). Au sein de chaque zone, un travail d'identification de la situation initiale et des axes d'amélioration ont été menés depuis 2013.

Par exemple, dans le domaine des **adresses**, plusieurs actions ont été jugées indispensables : la numérotation d'espaces publics ou privés qui n'avaient pas d'adresse (les jardins, certaines zones industrielles) mais aussi la création d'un identifiant unique et pérenne pour identifier les lieux, même en cas de changement de dénomination de rue ou de nouvelle numérotation des adresses.

Deux chantiers transverses à l'ensemble des zones ont été lancés :

– la définition et la mise en place d'un **modèle de données interregistres**, qui vise notamment à faire le lien entre les différentes bases de manière stable et pérenne (par exemple : que l'identifiant des adresses soit utilisé dans l'ensemble des autres registres) ;
– la mise en place d'un **Distributeur de données** unique au niveau national [12], qui a vocation à remplacer les distributeurs existants et à offrir un niveau élevé de qualité de mise à disposition (par exemple avec une mise à jour plusieurs fois par jour pour les registres les plus dynamiques).

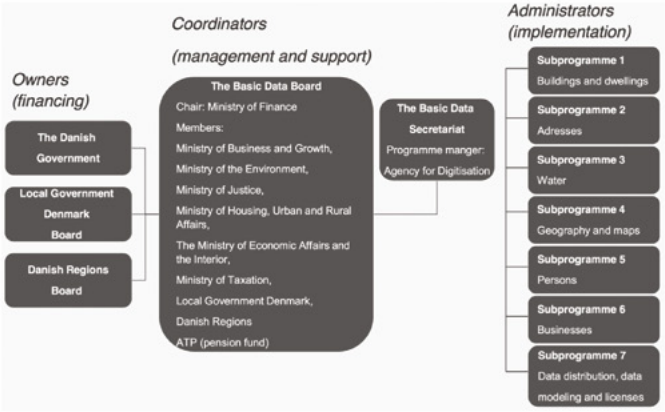
Une gouvernance forte qui intègre aussi le financement des données de référence

Le **financement des producteurs** des données de base a été rapidement identifié comme une clé pour la réussite du programme. Au lancement de Grunddata, certains producteurs étaient très dépendants des redevances, qui pouvaient représenter jusqu'à 80 % de leur budget. De plus,

12 *Cf.* http://datafordeler.dk/

les producteurs de données concentraient les fonctions de maître d'œuvre et de maître d'ouvrage.

Le Gouvernement danois a décidé de mettre en place une gouvernance dédiée aux données de base et de doter cette gouvernance d'un levier budgétaire. Le ***Grunddata Board*** est dirigé par le ministre des Finances et comprend l'ensemble des producteurs de données concernés, ainsi que les principaux financeurs (l'État danois, les communes et les régions).



Cette structure gère l'ensemble des budgets liés à **la production et à la diffusion des données de base**. Ainsi, les producteurs ont perdu la relative autonomie dont ils disposaient auparavant sur une partie de leur budget (de l'ordre de 20%) au profit d'une **gestion centralisée et unique**. Chaque ministère ou agence de l'État en charge d'une zone dispose ainsi d'un budget pour mettre en place les actions décidées en commun avec le *Board* et l'Agence pour la numérisation (qui constitue l'opérateur du programme Grunddata).

Une mise en place plus longue et compliquée que prévue

Très ambitieux, impliquant de nombreux partenaires – État, collectivités, secteur privé et couvrant de multiples facettes des enjeux de la donnée – de la production à la diffusion, le programme Grunddata a connu un certain nombre de **retards**.

Cinq ans après le lancement du programme Grunddata, le Distributeur de données vient de mettre en ligne en novembre 2017 son tout premier registre, en l'occurrence la base des noms de lieux danois. Le calendrier révisé prévoit une mise en ligne progressive qui doit s'étendre jusqu'en 2020 pour certains registres. Le Parlement danois a ainsi été sollicité pour **dégager des fonds supplémentaires** pour la finalisation du programme. La mise en place de la gouvernance centralisée et les travaux de convergence des bases de données se sont avérés plus compliqués à mettre en œuvre qu'initialement prévu.

X-Road (Estonie)

X-Road est présentée comme l'épine dorsale (*backbone*) de la stratégie d'e-gouvernement de l'Estonie, un cas d'école au niveau européen.

Initiée dès 2000, X-Road vise à **interconnecter les bases de données** des administrations dans une optique de simplification administrative. Le système s'appuie sur une identité numérique, obligatoire pour tout estonien de plus de 15 ans et qui concerne aujourd'hui plus de 95 % de la population du pays.

Les producteurs de données raccordés à X-Road proposent des services accessibles, sous forme de requêtes, à des **utilisateurs dûment enregistrés**. Ils décident des conditions d'accès à leurs services. Ainsi, le service des impôts détermine à qui il donne accès au service de délivrance d'une attestation fiscale. L'accès à X-Road est régulé par l'**Autorité des systèmes d'information** (RIA). C'est elle qui étudie les demandes de raccordement au système, tant au niveau des producteurs de données que des utilisateurs du service (administrations et entreprises).

À ce jour, X-Road permet d'accéder à plus de **1 200 services** proposés par près de 150 producteurs de données. 950 « consommateurs » de services ont accès à X-Road. Il s'agit essentiellement d'administrations ou d'entreprises qui ont besoin de données d'identité présentes dans les bases des administrations estoniennes.

En 2017, les principaux services utilisés (en nombre de requêtes) sont[13] :

- l'attestation de **situation fiscale** qui permet d'assurer qu'un individu est à jour de ses obligations fiscales ;
- l'attestation de **droits d'assurance maladie** qui permet de s'assurer qu'un individu est assuré ou, dans le cas contraire, de lui ouvrir de nouveaux droits ;
- la **situation professionnelle** d'un individu (actif, inactif, demandeur d'emploi, etc.) ;
- l'accès au **dossier médical** d'un patient (et notamment la liste des prescriptions).

Les principaux « consommateurs » des services fournis *via* X-Road (toujours en nombre de requêtes sur l'année 2017) sont le **secteur bancaire et financier** (notamment pour l'attribution de crédits aux particuliers), le **secteur médical**, l'administration fiscale estonienne, les services de police et de défense du territoire.

Elle se distingue des autres initiatives étudiées (Danemark et Royaume-Uni) sur plusieurs critères :

13 Source : http://x-road.eu/xtee-stats/

– la **finalité** : X-Road est conçue dans une optique de **e-administration** et de simplification administrative et non dans une optique d'exploitation des données dans un sens plus large ;
– la **régulation** de l'accès : contrairement aux exemples danois et britannique, l'accès aux données est limité à la finalité du service ;
– le **type de données transmises** : la majorité des bases interconnectées *via* X-Road sont des bases comprenant des données à caractère personnel : registre des contribuables, des propriétaires fonciers, dossiers patients, etc.

En ce sens, l'approche estonienne se rapproche davantage des programmes de simplification administrative (dont « Dites-le-nous une fois ») que d'une infrastructure de données ouverte et partagée telle qu'envisagée au Royaume-Uni et au Danemark.

L'accès aux données de référence (cadastre, registre des entreprises) est payant pour le secteur privé

Parallèlement au système X-Road de nombreux registres – dont le Cadastre ou le registre des entreprises – sont accessibles en ligne.

L'accès à ces registres et le téléchargement de données sont payants pour les acteurs du secteur privé. À titre d'information, les revenus générés par ces redevances s'élevaient à plus de **3 millions d'euros** pour l'année 2015[14].

La situation comparée en France

La notion de **« service public de la donnée énergétique »** a été évoquée dès 2014 à l'occasion de la discussion sur la loi de transition énergétique[15]. Il s'agissait alors d'encourager la circulation des données produites par les gestionnaires de réseaux d'énergie.

L'année suivante, la préparation de la loi pour une République numérique a été l'occasion de poser les bases du service public de la donnée. La loi pour une République numérique, promulguée en octobre 2016 a ainsi créé un nouveau service public « *chargé de mettre à disposition, en vue de faciliter leur réutilisation, les données de référence*[16] ».

Un focus sur la diffusion des données, une gouvernance par les objectifs

Le législateur a souhaité que le **service public de la donnée** se concentre sur la **diffusion** des données de référence et non sur les conditions de leur production. Ainsi, la plupart des critères définis par les textes juridiques concernent d'abord la qualité de la mise à disposition des données : disponibilité, fraîcheur, qualité des métadonnées, etc.

14 Ces revenus couvrent environ 70 % du coût de production des registres concernés. Source : http://www.rik.ee/en/e-land-register/service-fee-rates
15 http://www.assemblee-nationale.fr/14/pdf/cr-cstransenerg/13-14/c1314005.pdf
16 Pour une présentation détaillée de la notion de données de référence, *cf.* la partie 1 de ce rapport.

D'autre part, il a été fait le choix d'une gouvernance basée sur des **objectifs**, mais qui ne déterminent pas un modèle unique de répartition des rôles entre producteur et diffuseur. Contrairement au modèle danois où la diffusion sera réalisée exclusivement à terme par le Distributeur de données national, en France les producteurs sont **libres de diffuser eux-mêmes** les données de référence, ou d'en **confier la diffusion à un tiers**. Dans les deux cas, les règles techniques et d'organisation de mise à disposition de référence résumées ci-dessous doivent être respectées.

Les règles techniques et d'organisation de mise à disposition des données de référence

Les principales règles établies par l'arrêté du 14 juin 2017 concernent :

- la documentation des données (métadonnées) ;
- l'information des utilisateurs sur le processus de création des données de référence ;
- l'engagement des producteurs sur la fréquence de mise à jour de chaque base de données de référence (de l'annuel au quotidien selon les producteurs concernés) ;
- le taux de disponibilité : 99 % mensuel pour le téléchargement, 99,5 % pour les interfaces de programmation ;
- les modalités de mise à disposition qui doivent notamment garantir l'authenticité des données ;
- la procédure de signalement au producteur de données de référence en cas d'erreur ou d'incomplétude relevée dans ces données ou dans les informations associées ;
- le délai, qui ne peut être inférieur à trois mois, d'information des usagers de toute modification substantielle des caractéristiques des données de référence, de leurs modalités de mise à disposition, et de la structure de la base de données.

Une construction progressive

Le choix de construire le service public de la donnée **par la diffusion** et non par la production des données de référence, comme dans l'exemple danois, a plusieurs impacts. Tout d'abord, cela a permis, un an à peine après la promulgation de la loi pour une République numérique de commencer la diffusion en open data de neuf jeux de données de référence.

Cette **approche agile**, qui se construit progressivement en interaction avec les utilisateurs, nous semble être la plus efficace pour commencer à délivrer rapidement des services.

Mais **construire l'infrastructure de données par l'aval**, contrairement à l'approche danoise par l'amont (la diffusion est précédée d'un long travail sur les conditions de la production) présente aussi un **ensemble de défis**.

Le premier d'entre eux concerne la mise en place d'une gouvernance adaptée. La répartition des engagements entre producteur et diffuseur[17] est un point critique. De même, il faut associer les réutilisateurs aux évolutions du service public de la donnée mais aussi à la gouvernance. Leur contribution est essentielle, notamment pour faire progresser la **montée progressive en qualité** des jeux de données de référence.

Tableau de synthèse

	UK Registers	Danish Grunddata	X-Road Estonia	Service public de la donnée FR
Données	Nomenclatures	Registres	Registres d'identité (état civil, fiscalité, santé)	Registres et bases administratives
Volumétrie des données (ordre de grandeur)	Quelques dizaines à quelques milliers d'enregistrements	Quelques millions d'enregistrements	Quelques millions d'enregistrements	Quelques millions d'enregistrements
Données à caractère personnel	Explicement exclues	Incluses	Incluses et prépondérantes	Vocation à être incluses
Approche	Par la diffusion	Par la production et la diffusion	Par l'interconnexion de bases *via* l'identité numérique	Par la diffusion
Centralisation de la gouvernance	Faible	Forte	Mixte : autorité centrale (raccordement) et producteurs (droits d'accès)	Faible
Mode de coordination avec les producteurs	Par la labellisation, la responsabilisation individuelle des conservateurs	Par la dotation budgétaire	Par une autorité centrale	Par les objectifs, avec le suivi public des engagements

Les leçons à tirer des initiatives européennes

L'analyse des initiatives du Danemark, du Royaume-Uni et de l'Estonie est riche d'enseignements pour la construction d'une infrastructure de données dont le service public de la donnée constitue l'ébauche pour le cas de la France.

Le premier enseignement est que construire une véritable infrastructure de données à la hauteur des enjeux demande du temps, des investissements

17 Le diffuseur ne peut s'engager sur la fréquence de mise à jour, par exemple. À l'inverse, une donnée de référence mise à jour mais indisponible en raison de l'interruption de l'interface de programmation (API) ne présente pas d'intérêt.

mais aussi **un engagement politique fort et constant sur plusieurs années**. On ne construit pas une infrastructure, informationnelle – et encore moins physique – en deux ou même cinq ans.

La France peut s'appuyer sur l'expertise des grands producteurs de données publiques (et notamment l'INSEE, IGN, Météo France, la DGFiP). Mais la construction d'une infrastructure de données doit être considérée comme un **investissement public** à part entière. Son financement doit être pérennisé.

Le second enseignement est que la construction d'une infrastructure de données repose sur **plusieurs leviers**, dont certains ne sont pas techniques :

– le **levier budgétaire** : le Danemark a ainsi fait le choix de centraliser le financement de la production des données de référence dans une structure interministérielle, au détriment d'une partie de l'autonomie financière des producteurs ;
– le **levier contractuel** : les objectifs fixés aux ministères, les contrats d'objectifs et de moyens des opérateurs doivent intégrer la contribution de chaque producteur à l'infrastructure de données ;
– le **levier juridique** : les efforts menés en faveur de l'ouverture des données publiques (sur le principe de gratuité, les licences de réutilisation, les standards ouverts) constituent des accélérateurs pour construire l'infrastructure de données.

Enfin, **le choix d'un modèle de gouvernance** apparaît bien comme un élément structurant d'une infrastructure de données. Un certain degré de centralisation est nécessaire, ne serait-ce que pour fixer *a minima* des règles et des standards communs à l'ensemble des bases de données de référence.

Plan de route de l'infrastructure de données

Une infrastructure de données est composée :

- de données de qualité, et en particulier de celles qui répondent aux critères des données de référence ;
- d'infrastructures de mise à disposition, en API et en téléchargement ;
- de mécanismes d'identification, de sécurisation et de contrôle ;
- de mécanismes de participation des utilisateurs à la montée en qualité des données.

Certaines **briques de cette infrastructure** sont déjà en place : le portail data.gouv.fr offre ainsi un service de téléchargement des données, ainsi que la possibilité d'interagir avec les producteurs des données (via le signalement et un forum de discussion pour chaque jeu de données). De même, certaines données de référence sont aujourd'hui exposées par des API, en particulièrement les données géographiques ou d'entreprises (API entreprise). Ces API sont référencées dans un catalogue : **api.gouv.fr** où les consommateurs peuvent les découvrir puis prendre contact avec les producteurs.

Un financement dédié dans le cadre du **Programme d'investissements d'avenir** (PIA) permet de renforcer les dispositifs existants (notamment en termes de performance et de sécurisation) et de mettre en place progressivement les éléments manquants. L'infrastructure de données doit permettre d'assurer la plus large circulation possible des données. Cela signifie concrètement que lorsque rien ne s'oppose à la diffusion à tous de ces données, elles doivent être proposées selon les principes des données ouvertes.

Cependant, quand certaines bases de données – ou certaines parties d'entre elles, par exemple les éléments d'identification d'un responsable associatif – contiennent des données à caractère personnel ou sont couvertes par d'autres secrets légaux, l'infrastructure de données doit permettre de s'assurer que seuls ceux qui ont le droit d'en connaître y accèdent. Pour ce faire, des **mécanismes d'identification et de contrôle d'accès** seront progressivement développés, en s'appuyant notamment sur les développements de FranceConnect Identité et de FranceConnect Agents.

Les utilisateurs de cette infrastructure participent de manière active à sa réussite et à sa gouvernance. Ils participent notamment à la montée en qualité des données en signalant des anomalies et en proposant des mises à jour. Cette brique de l'infrastructure de données est essentielle. A contrario, une plus large circulation des données sans possibilité d'interagir avec les producteurs se traduirait par une propagation des défauts des données et par une multiplication des coûts de contrôle et de correction des erreurs.

Troisième partie

Transformer l’essai

Depuis sa création en 2014, l'administrateur général des données a accompagné la mise en place d'une véritable **politique de la donnée** dans ses multiples dimensions : techniques, économiques et juridiques. Les lois ont évolué. Des outils, des plateformes, des API ont été développés et sont maintenant utilisés par de très nombreux utilisateurs. Des données essentielles sont mises à disposition en open data. Des réalisations très concrètes ont permis de valider le **bénéfice des datasciences** au service de l'action publique. Les mentalités aussi ont évolué, et plusieurs ministères commencent aujourd'hui à intégrer la donnée dans leur stratégie et leurs actions.

Il faut maintenant transformer l'essai. L'heure n'est plus aujourd'hui à démontrer l'intérêt d'une meilleure exploitation des données. Il faut à présent **mettre en ordre de marche** les administrations, en commençant par les ministères et les grands opérateurs. De par son rattachement aux services du Premier ministre et à la DINSIC, l'AGD a pour rôle d'accompagner l'appropriation de la révolution de la donnée par les administrations.

La feuille de route 2018 de l'administrateur général des données comporte cinq volets :

- mettre à disposition les **données et les infrastructures** mutualisées, les faire changer d'échelle ;
- développer une **doctrine de la circulation** des données au sein de la sphère publique ;
- renforcer le **réseau des administrateurs ministériels** des données, en faire un levier pour la politique de la donnée ;
- développer une expertise en matière d'**intelligence artificielle** au service de l'action publique pour faire de l'État l'un des premiers utilisateurs de ces outils ;
- **soutenir l'écosystème** des utilisateurs des données produites par l'administration, mesurer l'impact en termes sociaux, économiques et de transformation de l'action publique.

1. Mettre à disposition les données, les ressources et les infrastructures

Les données à fort impact économique ou social

Les deux dernières années ont été marquées par la mise à disposition de jeux de données à fort impact : base des entreprises, cadastre, base adresse nationale… Parallèlement, des approches sectorielles ont été développées pour faciliter la réutilisation des données géographiques ou des données des entreprises.

En 2018, et en lien avec les administrateurs ministériels des données (*cf. infra*), nous identifierons, dans chaque ministère et chaque domaine de l'action publique, les données qui peuvent être qualifiées de **données de**

référence, au sens du Code des relations entre le public et l'administration. Certaines d'entre elles sont déjà disponibles en ligne sans toutefois atteindre le niveau de qualité de mise à disposition attendu du service public de la donnée. D'autres ne sont pas encore « ouvertes ». Dans les deux cas, il s'agira de travailler, en lien avec les producteurs, pour les rendre encore plus facilement *découvrables* et réutilisables. Cela pourra notamment inclure le développement d'API ou de toute autre ressource nécessaire.

Une attention particulière sera accordée à **la qualité des données** mises en ligne, en lien avec d'autres initiatives comme le projet Qualidata, lauréat du Programme d'investissements d'avenir. Les utilisateurs des données seront mobilisés pour participer à la montée en qualité des données, notamment par le développement d'un outil de signalement d'erreurs et de proposition d'amélioration.

Les standards de données et les infrastructures

En 2018, le premier chantier consistera à **mettre en œuvre les standards** définis sur la commande publique et les conventions de subventions. Cela passera tout d'abord par la mobilisation des écosystèmes concernés : les administrations (État et collectivités) mais aussi les éditeurs de solutions logicielles et l'ensemble des réutilisateurs de ces données. Cet effort sera réalisé au niveau international dans le cadre de l'***Open Contracting Partnership*** dont la France a pris la présidence fin 2017 (*cf. supra*). La lutte contre la corruption *via* l'analyse des données de la commande publique figure parmi les premiers cas d'usage étudié dans le cadre de ce partenariat.

Par ailleurs, nous poursuivrons notre effort pour **définir de nouveaux standards de données**. Traduire en standard de données des obligations légales et réglementaires cela permet non seulement de rendre l'application des règles plus simples, mais aussi de faciliter l'émergence d'écosystèmes de réutilisation de ces données.

Concernant les **infrastructures**, la DINSIC va poursuivre la mise à disposition d'outils mutualisés qui facilitent la circulation de la donnée : plateforme ouverte des données publiques, verticales thématiques (géographie, entreprise, transports), API, dispositif France Connect Identité. Ces outils facilitent la circulation et l'exploitation des données. Ils contribuent à la mise en place d'une véritable infrastructure de données telle que nous l'avons exposé dans la seconde partie de ce rapport.

2. Développer la doctrine de la circulation des données au sein de la sphère publique

Le principe d'ouverture des données publiques par défaut figure maintenant dans la loi. Les outils – plateformes, API – existent pour permettre la circulation la plus large des données qui ne sont aujourd'hui couvertes par aucun secret et peuvent donc librement être partagées avec le plus

grand nombre. Cela n'épuise pas totalement cependant la question de la circulation de la donnée dans son ensemble.

Fournir la bonne donnée à la bonne personne, gérer le droit d'en connaître

En 2018, l'administrateur général des données contribuera à faire **évoluer la doctrine sur la circulation des données**, y compris des données protégées par des secrets et en premier lieu la vie privée.

Comme nous l'avions déjà souligné dans le précédent rapport de l'administrateur général des données, 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. Le plus important, dans le cas d'un secret légal est donc de bien savoir à qui il s'oppose, et dans quelles conditions. Il convient donc, pour permettre de faire circuler la donnée, de savoir qui a **le droit d'en connaître** et d'être capable de gérer ce droit.

Gérer le droit d'en connaître, cela passe notamment par une phase d'instruction : qui est le demandeur ? quelle est sa mission ? à quelles données veut-il accéder ? En fonction de cette analyse, nous devons être capables d'offrir l'accès à des données, y compris celles protégées par des secrets. Gérer le droit d'en connaître cela signifie aussi prendre la responsabilité du traitement et en assurer la traçabilité.

Notre ambition, réaffirmée, est d'être capable de fournir **la bonne donnée à la bonne personne**, dans le respect du droit d'en connaître.

Pour rendre cette doctrine pleinement opérationnelle, nos efforts porteront sur deux niveaux. Pour les données dont la circulation doit être contrôlée : nous accompagnerons le passage à l'échelle de l'API Particulier et de FranceConnect Identité. Pour les données qui doivent être anonymisées ou pseudonymisées : nous développerons une expertise et des outils afin d'être en mesure d'accompagner les administrations dans la publication des données, en s'appuyant sur l'équipe interne de data-scientists d'Etalab.

3. Renforcer le réseau des administrateurs ministériels des données

Certains ministères ou directions ministérielles ont désigné un administrateur des données, sur le modèle de l'administrateur général des données. En 2018, l'une des priorités consiste à **renforcer ce réseau** naissant pour en faire un **véritable outil** au service de la politique de la donnée.

Le renforcement du réseau passe tout d'abord par la nomination d'un administrateur ministériel des données au sein de chaque ministère. Cet administrateur aura pour tâche de porter la déclinaison de la politique de la donnée au sein de son ministère sur les quatre dimensions : **inventaire**

et cartographie des données existantes, **production des données** essentielles, **circulation maximale** des données, **exploitation** des données, notamment par les datasciences.

Il a aussi vocation à s'assurer de la bonne maîtrise des enjeux juridiques, en particulier en cette année 2018 où l'administration devra gérer en même temps le nouveau règlement européen sur la protection des données personnelles et la généralisation du principe d'open data par défaut.

Compte tenu du **rôle essentiel joué par les opérateurs** dans la production des données essentielles, il est souhaitable qu'ils désignent eux aussi un responsable des données en leur sein, en cohérence avec les administrateurs de leurs ministères de tutelle.

L'**administrateur général des données**, de par sa position interministérielle, est chargé d'assurer l'animation et la montée en puissance de ce réseau. La mise en commun des pratiques, des difficultés rencontrées et des solutions doit permettre de faire monter en compétence l'ensemble des administrations. L'administrateur général des données veillera aussi à s'assurer de la cohérence des actions entreprises par les ministères dans une optique de mutualisation des moyens.

4. Développer un pôle de compétences en matière d'intelligence artificielle

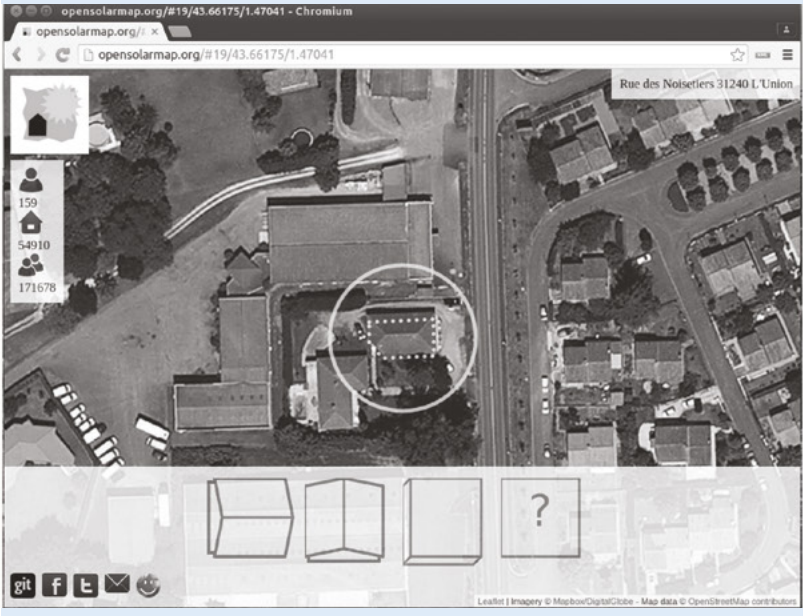
Le champ technologique autour des données et de leurs usages est en **constante et rapide évolution**, à l'image de la diffusion rapide de l'intelligence artificielle notamment sous la forme d'algorithmes apprenants (et de l'apprentissage profond – *deep learning*). L'intelligence artificielle, son potentiel et ses risques interrogent bien sûr l'État tant dans son rôle de régulateur que d'opérateur de politiques publiques.

Sous l'égide de l'administrateur général des données, la DINSIC renforcera en 2018 son expertise et ses capacités en matière d'intelligence artificielle. Les data-scientists de la mission Etalab ont déjà mis en œuvre avec succès de telles approches, notamment dans le cadre du projet OpenSolarMap (*cf. infra*).

OpenSolarMap : combiner intelligence humaine et intelligence artificielle

Et si nous étions capables d'indiquer, rapidement et simplement, le potentiel photovoltaïque d'un bâtiment ? Cela permettrait d'évaluer l'opportunité d'installer des panneaux solaires. C'est l'objectif poursuivi par OpenSolarMap, qui combine la contribution par la foule (*crowdsourcing*) et l'apprentissage par des algorithmes (*machine learning*) pour produire de nouvelles données.

L'approche suivie par OpenSolarMap est de déduire la forme des toits par analyse d'imagerie satellitaire, grâce à la mise en open data des images des satellites Spot. L'orientation de la pente d'un toit et sa surface ainsi obtenues sont les éléments les plus importants pour estimer l'opportunité d'installation de panneaux solaire. Une première interface graphique a été développée : elle se présente sous la forme d'un jeu où chaque internaute est invité à indiquer l'orientation de la pente du toit. Ainsi, en moins d'un mois, la plateforme a collecté près de 100 000 analyses de qualité. En recoupant les différentes analyses pour un même bâtiment, environ 10 000 toits ont ainsi pu être caractérisés avec certitude.



Cet échantillon de toits déjà classifiés a permis de développer un classifieur automatique en utilisant des techniques classiques en traitement d'image (régression logistique et deep learning). L'algorithme obtenu ne se trompe que dans 20 % des cas, ce qui est suffisamment peu pour l'application envisagée*.

* Les données calculées par l'algorithme sont publiées sur data.gouv.fr. Elles sont aussi publiées sous la forme d'une carte à l'adresse cadastre.opensolarmap.org

L'ambition est double : d'une part être capable d'orienter les ministères dans le recours aux technologies de l'intelligence artificielle, d'évaluer les outils et les offres existantes, de **mener des projets** sur quelques cas d'usages emblématiques. D'autre part, de rester vigilant sur la dimension d'éthique et de responsabilité de la mise en œuvre de tels traitements, *a fortiori* pour des systèmes apprenants qui ne sont pas toujours entièrement explicables.

Définir les conditions d'une utilisation éthique et responsable

Les enjeux de responsabilité, de transparence mais aussi de cohérence entre le droit et l'informatique (*Code is Law*) ont déjà pu être abordés en 2017 à propos d'admission post-bac et de son successeur Parcoursup [1]. La loi pour une République numérique a introduit des dispositions relatives à la transparence des algorithmes et à l'ouverture des codes sources.

En 2018, nous définirons, dans le cadre des engagements de la France au sein du Partenariat pour un gouvernement ouvert de l'*Open Government Partnership*, les **conditions d'une utilisation éthique et responsable des algorithmes** (apprenants ou « classiques ») pour l'action publique.

5. Soutenir l'écosystème des utilisateurs de données publiques

L'écosystème des utilisateurs de données publiques est un écosystème dynamique et riche. On dénombre ainsi plus de **185 000 visiteurs uniques mensuels** sur la plateforme data.gouv.fr. Un nombre bien supérieur d'entreprises, d'associations ou de particuliers accède à des services qui sont rendus possibles grâce aux données proposées en open data et, pour certaines d'entre elles, *via* un accès contrôlé (API Entreprise, API Particulier).

Le soutien de cet écosystème est l'une des conditions de la pleine exploitation du potentiel des données. Cela passe notamment par la participation ou l'organisation d'événements publics et la reconnaissance des initiatives du secteur public, associatif ou privé les plus impactantes, voire le soutien financier à certaines d'entre elles.

En 2018, l'administrateur général des données s'efforcera de **documenter les impacts sociaux et économiques** d'une meilleure circulation des données, dans la continuité du premier rapport de l'AGD qui analysait les mécanismes de création de valeur par les données.

---

1 En 2017, la mission Etalab a été mandatée pour étudier les conditions d'ouverture d'admission post-bac.

GLOSSAIRE

A/B testing : le test A/B est une technique qui permet de tester deux versions différentes (A e 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 : trad. 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.

La donnée est devenue un outil indispensable à la transformation de l'action publique et plus largement de l'économie. Pour en assurer la meilleure exploitation possible, il est nécessaire d'en garantir la qualité et la circulation. Au même titre que les réseaux de transport, d'énergie ou de télécommunications, une nouvelle forme d'infrastructure devient ainsi essentielle dans notre société moderne : celle de la donnée.

Dans son deuxième rapport, remis au Premier ministre, l'administrateur général des données prône la construction de cette infrastructure, en s'appuyant notamment sur une analyse comparée des approches de plusieurs pays européens. Il y propose des pistes d'amélioration afin que l'État joue pleinement son rôle dans ce nouveau paysage des données.

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
Premier ministre
Direction interministérielle du numérique et du système d'information et de communication de l'État

Direction de l'information légale et administrative

Imprimé en France
ISBN : 978-2-11-145683-9
DF : HC47150