La synthèse présentée dans ce document s'appuie sur de nombreuses sources documentaires, avec de larges emprunts, notamment issus du rapport 2016-2017 « *La donnée comme infrastructure essentielle* » de l'administrateur général des données, ou d'entretiens téléphoniques (dont Simon Chignard) – nous remercions ici l'ensemble de ces auteurs.

## 1. Overview of the initiative

| | |
|---|---|
| Name of initiative | Ouverture des données publiques et de recherche. De l'administrateur général des données (*national Chief Data Officer*) au plan national pour la science ouverte. |
| Objective | Impulsion au plus haut niveau de la démarche d'ouverture de la donnée publique (inventaire, gouvernance, production, circulation et exploitation), pour la transformation de l'action publique et plus largement de l'économie, et extension de cette politique aux données de la recherche. |
| Type (strategy, policy, bill of law,…) | Le décret du 16 septembre 2014 institue, auprès du Premier ministre, la fonction d'Administrateur général des données (AGD), rattaché au secrétaire général pour la modernisation de l'action publique.<br>L'Administrateur général des données coordonne l'action des administrations en matière d'inventaire, de gouvernance, de production, de circulation et d'exploitation des données.<br>Cette action a été prolongée par la Loi pour une République numérique d'octobre 2016 et par l'annonce du Plan national pour la science ouverte le 4 juillet 2018. |
| Responsible policy making bodies | -Leading Ministry: premier ministre<br>-Relevant Ministry: tous les ministères sont concernés par l'impact législatif. A terme, tous les ministère contribueront au réseau en construction d'administrateurs ministériels des données.<br>- Tous les ministères finançant des activités de recherche, avec en premier lieu le ministère de l'enseignement supérieur, de la recherche et de l'innovation. |
| Responsible implementing bodies | L'Administrateur général des données est rattaché au Secrétaire d'État chargé du Numérique, secrétariat du Premier ministre.<br>L'Administrateur général des données est également directeur interministérielle du numérique et du système d'information et de communication de l'État (DINSIC), direction également rattachée au secrétariat d'Etat chargé du Numérique. |
| reference framework (if relevant)International | Grandes métropoles. |
| Target audience | Administrations, recherche, associations, entreprises, société civile. |
| Total duration of initiative (years) | Depuis 4 ans. |
| Total budget of initiative (in national currency) | Non pertinent. |
| Sectoral focus (if relevant) | |
| Type of data concerned (data from research, public sector information, private sector information) | Données issues du secteur public, dont données issues de la recherche publique. |
| Target audience (scientific community, business, civil society, general public) | Administrations, associations, acteurs de la recherche, entreprises, société civile. |
| Expected results | Impulsion d'une dynamique législative favorable à l'ouverture des données publiques.<br>Généralisation de mise à disposition de données (de qualité, fiables, mises à jour, disponibles).<br>Actions en faveur de l'utilisation de ces données.<br>Développement de services autour des données.<br>Développement d'une culture des données et de l'ouverture de celles-ci. |

## 2. Rationale, motives and key drivers

On trouve en France un terreau favorable dès 1978 [1] en faveur de l'accès aux données des Administrations, avec la loi relative à la liberté d'accès aux documents administratifs[2], à l'ouverture de données aux citoyens et entreprises, sur motivations de bonne gouvernance démocratique, incluant déjà un chapitre dédié au droit à la réutilisation des données publiques. Habituant les Administrations à considérer l'utilité d'accès à certaines informations, la pratique pourtant consistait pour les citoyens d'avoir recours à la loi pour obtenir des données, plutôt que de voir les Administrations partager des données publiques de leur propre initiative.

Depuis plus de vingt ans[3], la prise de conscience par l'État français de l'importance des données produites et collectées par ses services et de la nécessité de les mettre gratuitement à disposition d'utilisateurs et de réutilisateurs s'accélère, pour renforcer la démocratie et développer l'économie, ainsi que pour moderniser l'action publique. C'est notamment au tournant des années 2000 que les autorités publiques ont décidé d'affirmer un principe de gratuité de la réutilisation des informations publiques, mouvement plus largement européen avec la directive 2003/98/CE du Parlement européen et du Conseil du 17 novembre 2003 concernant la réutilisation des informations du secteur public[4].

En 2013, le rapport Trojette [5] précise l'argumentation économique, en exposant notamment que l'ouverture de données par des Administrations a un effet levier important de bénéfice social, avec un chiffrage financier d'un manque à gagner économique pouvant être 20 fois supérieur aux manques à gagner de redevances[6], mettant en exergue l'importance des infrastructures informationnelles de mise à disposition de l'information publique qui représentent des biens publics aussi importants que les infrastructures physiques et dont on peut attendre des externalités positives aussi importantes.[7]

Plus récemment, le rapport Fouilleron en 2015[8], portant sur les échanges de données tarifés entre administrations, a permis de déterminer, sur la base d'une enquête réalisée auprès de 80 administrations, le montant des transactions de ventes de données entre Administrations, de l'ordre de 20 M€ en 2014. En exemple illustratif : parmi les trente vendeurs de données identifiés, seulement cinq d'entre eux concentraient plus de 90 % des montants identifiés (la caisse nationale d'assurance retraite (CNAV), l'institut national de l'information géographique et forestière (IGN), l'Institut national de la statistique et des études économiques (INSEE), la direction générale des finances publiques (DGFiP) et l'agence centrale des organismes de sécurité sociale). A contrario, la moitié des transactions liées aux données entre administrations, opérateurs et collectivités 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.
Aussi a-t-il été démontré que cette tarification des données entre Administrations nuit à l'efficacité et la qualité de l'action publique, en engendrant des coûts de transaction et des effets pervers comme le renoncement à la donnée pour des raisons budgétaires ou des stratégies de contournement du frein

[1] Sans remonter plus avant à l'article 15 de la Déclaration des Droits de l'Homme et du Citoyen de 1789 « La Société a le droit de demander compte à tout Agent public de son administration »
[2] Loi du 17 juillet 1978 intitulée « de la liberté d'accès aux documents administratifs »
[3] Voir par exemple le Programme d'action gouvernementale pour l'entrée de la France dans la société de l'information (PAGSI) adopté lors du Comité interministériel pour la société de l'information (CISI) du 16 janvier 1998, et les études et rapports qui le suivirent, comme le rapport Mandelkern « Diffusion des données publiques et révolution numérique », Commissariat général du plan, 1999. http://www.ladocumentationfrancaise.fr/var/storage/rapports-publics/994001620.pdf
[4] https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32013L0037&from=FR
[5] M. Trojette (2013), Ouverture des données publiques Les exceptions au principe de gratuité sont-elles toutes légitimes ?, rapport au premier ministre http://www.ladocumentationfrancaise.fr/var/storage/rapports-publics/134000739.pdf
[6] Résultat corroboré par les analyses de l'IGN : le passage à la gratuité du référentiel grande échelle de l'établissement public pour les organismes chargés d'une mission de service public administratif a entraîné une multiplication par 20 des volumes de données téléchargés, soit un bénéfice social estimé à 114 M€ par an, pour un manque à gagner de 6 M€ de redevance environ.
[7] Pour un regard européen, voir G. Vickery (2010), Review of Recent Studies on PSI Re-Use and Related Market Developments, rapport à la Commission européenne (ec.europa.eu/newsroom/document.cfm?doc_id=1093)
[8] A. Fouilleron (2015), Les échanges de données réalisés à titre onéreux entre les administrations, rapport au Premier ministre http://www.modernisation.gouv.fr/sites/default/files/fichiers-attaches/rapport_echanges-donnees-entre-administrations.pdf

budgétaire de la part des administrations acheteuses. A contrario, la gratuité des échanges de données au sein du secteur public est de nature à favoriser leur circulation et leur utilisation, vectrices d'externalités et d'effets de réseaux, tout en stimulant la progression de la culture de la donnée au sein de l'administration, pour une plus grande efficacité et qualité de l'action publique.
Pour réaliser le potentiel de création de valeur économique et sociale sous-jacent à la circulation des données, la mission préconisait d'instaurer dans la loi, de manière à couvrir le périmètre le plus large des administrations, un principe de gratuité de ces transmissions entre administrations réalisées dans le cadre de l'exercice des missions de service public. Afin de rendre effectif ce principe tout en préservant les équilibres financiers des administrations vendeuses et acheteuses de données, il était préconisé d'introduire parallèlement une mesure de neutralisation en base budgétaire. Ces préconisations menèrent à la loi relative à la gratuité et aux modalités de la réutilisation des informations du secteur public, dite *Loi Valter*[9].

Les freins à la circulation des données au sein du secteur public ne sont pourtant pas que budgétaires, la gratuité des échanges n'étant qu'une partie de la solution. En effet, les contraintes techniques, les restrictions juridiques, les freins culturels ou le manque d'information limitent également l'exploitation du potentiel que représentent les données. La mise en œuvre du principe de gratuité doit ainsi s'accompagner d'évolutions des systèmes d'information de l'Etat et des processus de mise à disposition des données, en cohérence avec la stratégie de modernisation de l'action publique par le numérique.

**Des démarches concrètes, comme nous le verrons, sont engagées pour faciliter et normaliser les échanges de données et accroître la communication entre administrations, appuyées par un chef d'orchestre, dont les missions *ad hoc* vont ici être précisées, l'Administrateur Général des Données.**

**Concomitamment à la création d'un Administrateur Général des Données, mais contribuant à la bonne réalisation de ces missions et qu'il nous faut ici mentionner, un ensemble d'initiatives ont été déployées, de nature législatives (Loi pour une République Numérique notamment), budgétaire (appels à projets, comme « Transition numérique de l'Etat et modernisation de l'action publique », *etc*.) ou organisationnelle (constitution de réseaux, incubation de start-ups d'Etat, *etc*.).**

**Ces éléments s'inscrivent dans une ambition récemment renouvelée par le président de la République Emmanuel Macron**, avec une nouvelle impulsion, notamment internationale[10] :

- d'une part en proposant d'ouvrir une réflexion à l'échelle européenne sur l'accès, à des fins d'intérêt général, aux bases massives de données privées, notamment celle des très grands acteurs qui se trouvent en monopole de fait sur la collecte de certaines catégories de données, en considérant que ces données ont une part de biens collectifs dont il faut que, à la fois la Recherche mais l'ensemble des conséquences et des innovation subséquentes, puissent être partagées par l'ensemble de la population européenne[11] ;
- d'autre part en évoquant la possibilité de créer un Groupe d'experts intergouvernementaux de l'intelligence artificielle (à l'instar du GIEC pour le climat), afin de créer une expertise

[9] Loi n°2015-1779 du 28 décembre 2015
*https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000031701525&categorieLien=id*

[10] Discours du Président de la République Emmanuel Macron le 29 mars 2018 au Collège de France, cf. http://www.elysee.fr/declarations/article/transcription-du-discours-du-president-de-la-republique-emmanuel-macron-sur-l-intelligence-artificielle

[11] A noter l'appel à manifestation d'intérêt de la Direction Générale des Entreprises sur la « mutualisation de données pour l'intelligence artificielle », afin de recueillir l'intérêt et l'avis des acteurs privés et publics sur les initiatives de mutualisation de données qui seraient les plus pertinentes et les modalités les plus adaptées à leur soutien (cf. https://www.entreprises.gouv.fr/numerique/mutualisation-de-donnees-pour-intelligence-artificielle)

mondiale indépendante qui puisse organiser le débat collectif et démocratique sur les évolutions scientifiques de manière totalement autonome.

**Plan National pour la Science Ouverte**

Plus récemment, le Plan national pour la science ouverte, annoncé le 4 juillet 2018 par la Ministre de l'Enseignement Supérieur, de la Recherche et de l'Innovation, Mme Frédérique Vidal, comprend 9 mesures en faveur de la Science Ouverte, selon les axes (1) généraliser l'accès ouvert aux publications, (2) structurer et ouvrir les données de la recherche et (3) s'inscrire dans une dynamique durable, européenne et internationale, qui seront détaillés plus loin.

## 3. Governance

### 3.1. Administrateur général des données

En 2014, le gouvernement crée par décret le nouvel emploi d'Administrateur général des données[12], directement rattaché au premier ministre[13], et précise ses missions :

- Coordination de l'action des administrations en matière d'inventaire, de gouvernance, de production, de circulation et d'exploitation des données par les administrations[14] ;
- Organisation, dans le respect de la protection des données personnelles et des secrets protégés par la loi, de 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 ;
- Proposition au Premier ministre toutes mesures, y compris, le cas échéant, des évolutions législatives ou réglementaires.

En concertation avec les administrations concernées, l'administrateur général des données :

- Propose au Premier ministre des stratégies d'exploitation des données produites, reçues ou collectées par les administrations dans le cadre de leurs missions de service public, y compris en s'appuyant sur des entreprises innovantes ;
- Elabore des outils, des référentiels et des méthodologies permettant une meilleure exploitation des données et un plus grand usage des sciences des données au sein des administrations ;
- Adresse, en tant que de besoin, à la direction interministérielle des systèmes d'information et de la communication de l'Etat ses recommandations en matière de cadres techniques de référence visant à accroître l'interopérabilité des systèmes d'information et des données. Il peut en outre travailler à la sémantisation des données ;
- Conduit des expérimentations sur l'utilisation des données pour renforcer l'efficacité des politiques publiques, contribuer à la bonne gestion des deniers publics et améliorer la qualité des services rendus aux usagers.

L'administrateur général des données peut également être saisi par toute personne de toute question portant sur la circulation des données.

### 3.2. Autres fonctions complémentaires

L'Administrateur Général des Données est également, actuellement, le directeur interministériel du numérique et du système d'information et de communication de l'État (DINSIC), rattaché au secrétariat général du Gouvernement, et qui pilote :

[12] Décret n° 2014-1050 du 16 septembre 2014 portant création d'un administrateur général des données (https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000029463482)
[13] Rattaché au Secrétariat d'État au Numérique du Premier ministre.
[14] Les administrations, au sens du décret cité, sont les services centraux et déconcentrés de l'Etat ainsi que les établissements publics placés sous sa tutelle.

- le Service performance des services numériques (SPSN)
- la Mission Incubateur de services numériques[15]
- la Mission Étalab : dans la logique de l'État plateforme, la mission Etalab opère un ensemble d'outils et de dispositifs – plateformes, API[16], services – qui facilitent la circulation des données
- le Réseau interministériel de l'État (RIE)

Cette configuration de cumul de responsabilités dans le même champ de compétence permet de renforcer l'efficacité d'ensemble.

### 3.3. Un cadre légal dynamique

Cette démarche de création de la fonction d'Administrateur Général des Données unique s'appuie sur un cadre juridique général renouvelé, notamment avec la loi pour une République numérique[17] (#LoiNumérique), qui renverse le cadre observé sur plusieurs aspects complémentaires, et mise en œuvre par plusieurs entités[18]. Ce projet de réforme avait d'ailleurs fait l'objet d'un processus d'élaboration ouvert qui reste en France à ce jour inédit, suivant une logique de co-construction, en procédant en amont du projet à des concertations et une consultation publique en ligne qui a rencontré un important succès de participation[19].

Nous ne retiendrons ici que trois des quinze points clefs de la #LoiNumérique :

- **Création de l'obligation pour les organisations publiques de publier sur internet leurs bases de données**, sous réserve notamment d'anonymisation et de protection de la propriété intellectuelle et du secret industriel et commercial. Ces données pourront ainsi être exploitées et réutilisées facilement par chacun, particulier comme entreprise. Certains acteurs privés (entreprises titulaires des marchés publics, bénéficiaires de subventions publiques…) seront également tenues de communiquer des données d'intérêt général, qui pourront concerner l'exploitation des services publics de l'énergie ou de l'eau, les transactions immobilières, ou encore la gestion et le recyclage des déchets ;
- **Accès sécurisé aux données pour les chercheurs et statisticiens publics :** les données produites par la sphère publique sont souvent très riches, mais tout aussi souvent très confidentielles car du niveau de chaque individu. Leur accès était jusqu'ici dans les faits quasiment impossible, même pour les besoins de la recherche. Grâce à la #LoiNumérique, un système d'accès sécurisé permettra aux seuls chercheurs et statisticiens publics habilités, dans le cadre d'un projet donné, de pouvoir étudier ces données pour mieux comprendre l'efficacité de nos politiques publiques et évaluer l'effet de futures réformes ;
- **Libre accès aux résultats des travaux de recherche publique et autorisation de la fouille de textes et de données :** les résultats de travaux de recherche financés à plus de 50 % par des fonds publics pourront être mis en ligne en libre accès par leurs auteurs, après une période d'embargo de 6 à 12 mois. Cette mesure facilitera la libre diffusion de résultats de recherche dont la diffusion était auparavant souvent restreinte et concentrée par les éditeurs. La loi autorise également la fouille de textes et de données en ligne, une pratique essentielle dans le cadre notamment de recherches en sciences humaines et sociales, pratique jusqu'alors soumise à autorisation[20].

[15] https://beta.gouv.fr/apropos

[16] https://api.gouv.fr

[17] Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique (https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000033202746&categorieLien=id) – site https://www.republique-numerique.fr

[18] Outre l'AGD, également la Direction Générale des Entreprises (DGE) et l'INSEE notamment.

[19] Le gouvernement français avait ainsi mis en place une consultation publique de septembre à octobre 2015. Près de 21 000 participants provenant d'horizons différents (société civile, milieu associatif, entreprises) ont publié 8 500 contributions. Le gouvernement a ainsi intégré 90 contributions et ajouté 5 articles directement issus de la consultation.

[20] Le projet de décret sur le *Text & Data Mining* n'a malheureusement pas été validé par le Conseil d'Etat. Nous attendons donc la révision en cours de la directive européenne sur le copyright.

### 3.4. Engagement pour une transparence publique des données dans le cadre du Partenariat pour un gouvernement ouvert

Complémentaire de la démarche d'adaptation du cadre légal et règlementaire, une consultation citoyenne, permettant d'associer citoyens et experts à la vie publique selon une approche non discriminante, est en cours pour le « Plan d'action national français 2018-2020 » dans le cadre du Partenariat pour un gouvernement ouvert[21], avec 21 engagements selon plusieurs axes, dont nous retiendrons ici :

- **transparence informationnelle**, avec une très large variété d'engagements : commande publique, aide publique au développement, bénéficiaire effectifs des personnes morales et des trusts, activités de représentants d'intérêt, informations publiques relatives aux élus et responsables publics, renforcer la transparence des algorithmes et des codes sources publics, *etc*.
- **organisation d'acteurs et de communautés** : ouverture des données publiques et enrichir le service public de la donnée, outiller les administrations pour associer les citoyens à la décision publique, ouvrir l'administration à de nouvelles compétences et accompagner les initiatives d'innovation ouverte au sein de l'Etat, mettre à disposition des données stratégiques sur l'environnement et les territoires, *etc*

Ce nouveau plan d'action 2018-2020 comprend 21 engagements :

1. Renforcer la transparence sur l'efficacité et la qualité des services publics en relation avec les usagers
2. Accroître la transparence de la commande publique
3. Poursuivre la transparence de l'aide publique au développement
4. Enrichir le service public de la donnée : vers une nouvelle liste de données de référence
5. Désigner des administrateurs ministériels des données et accompagner la mise en oeuvre du principe d'ouverture par défaut
6. Renforcer la transparence des algorithmes et des codes sources publics
7. Accompagner les territoires dans la mise en oeuvre du principe d'ouverture des données par défaut
8. Créer un laboratoire d'intelligence artificielle ouvert pour l'Etat
9. Ouvrir l'administration à de nouvelles compétences et accompagner les initiatives d'innovation ouverte au sein de l'Etat
10. Créer des incubateurs de services publics numériques dans chaque ministère
11. Améliorer la fluidité des données au sein de l'Etat avec FranceConnect plateforme
12. Développer des nouveaux formats d'échange avec la société civile : le Forum Open d'Etat
13. Mettre en place un tableau de bord ouvert et contributif des démarches sur internet
14. Organiser en France un sommet international sur les Gov Tech
15. Outiller les administrations pour associer les citoyens à la décision publique
16. Accompagner la mise en oeuvre des principes de transparence et de participation citoyenne à l'international
17. Donner les moyens aux citoyens de contrôler et s'impliquer dans les décisions publiques sur la transition écologique et le développement durable
18. Construire un écosystème de la « science ouverte »
19. Impliquer davantage les citoyens dans les travaux menés par la Cour des comptes
20. Assurer une plus grande transparence des activités des représentants d'intérêts
21. Renforcer l'accès aux informations publiques relatives aux élus et responsables publics

Ce nouveau plan d'action repose sur cinq grandes parties :

- Transparence, intégrité et redevabilité de la vie publique et économique : rendre compte de la décision et de l'action publique est un principe fondamental qui contribue à renforcer la

[21] https://www.etalab.gouv.fr/opengov-openparliament-les-plans-daction-du-gouvernement-et-de-lassemblee-nationale-pour-une-action-publique-transparente-et-collaborative-ont-ete-lances

confiance entre responsables politiques et citoyens et à construire des politiques plus efficientes, plus proches des besoins des usagers. Cette première partie présente des engagements sur lesquels la communauté internationale et notamment celle du gouvernement ouvert est très engagée. Ils prolongent et renouvellent certains engagements du Plan d'action national 2015-2017 ;

- Ouverture des ressources numériques et innovation ouverte : les actions d'ouverture des données, des codes sources, des logiciels de l'Etat, d'innovation ouverte marquent la transformation des administrations et permettent notamment aux citoyens de prendre part aux processus de décision publique et de co-construction de l'action publique ;
- Des démarches de participation renforcées : la démocratie a évolué, et la participation en continu des citoyens à l'action publique doit être renforcée ;
- Un gouvernement ouvert au service des enjeux mondiaux de notre siècle : développement, environnement et science ouverte. La France soutient la mise en oeuvre des principes du gouvernement ouvert pour renforcer les politiques de développement en Afrique francophone, la protection de l'environnement et la transition écologique ainsi que l'accès aux matériaux et résultats de la recherche ;
- L'ouverture des juridictions et des autorités administratives indépendantes : la Cour des comptes et la Haute Autorité pour la transparence de la vie publique s'engagent aussi dans l'ouverture de leurs institutions.

Nous faisons ici un focus sur l'engagement 18, intitulé « Construction d'un écosystème de la « science ouverte » » qui relie pour la première fois en France la politique de science ouverte avec celle de gouvernement ouvert.

Pour ce dernier engagement, dont l'institution porteuse est le Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation, la considération des enjeux portant sur la mutation numérique de nos sociétés pour développer notamment l'accès ouvert (*open access*), va au-delà des données ouvertes (*opendata*) : elle porte plus généralement sur les processus scientifiques ouverts, comprenant une ouverture des processus d'évaluation, des indicateurs, des licences de réutilisation, des codes sources ainsi que des pratiques numériques.

Les ambitions fortes pour construire un écosystème de la « science ouverte » se retrouvent dans les engagements de la feuille de route que nous pouvons ici détailler :

- Créer un « Comité pour la science ouverte » pour un échange ouvert, à vocation nationale et internationale, sur les questions de Science ouverte (accès, données, métriques, codes, science participative)
- Mettre en place un dispositif de monitoring quantitatif de l'état d'avancement de la diffusion en accès ouvert de la littérature scientifique nationale
- Constituer un jeu de données ouvert sur les financements de projets de recherche sur AAP et leurs bénéficiaires
- Adhésion nationale à ORCID (système d'identification unique des chercheurs qui permet de connaître plus simplement et sûrement les contributions scientifiques d'un chercheur)
- Accélérer le développement de l'archive ouverte nationale, HAL avec un investissement sur la simplicité d'usage et l'interopérabilité en renforçant ses moyens
- Enrichir scanR[22], moteur de la recherche et de l'innovation et développer sa notoriété et son usage notamment pour alimenter le débat public des résultats de la recherche
- Communiquer auprès des communautés scientifiques sur les implications de la loi numérique relatives à l'ouverture des publications et des données
- Dans le cadre du soutien public aux revues, recommander l'adoption d'une politique de données ouvertes associées aux articles et le développement des *data papers*
- Généraliser progressivement via un accompagnement la mise en place de plans de gestion des données dans les appels à projets de recherche, et inciter à une ouverture des données produites par les programmes financés
- Mettre en place un dispositif de monitoring transparent (public) des dépenses relatives aux acquisitions électroniques dans les bibliothèques universitaires, et diffusion des dépenses en *open data* sur le portail du MESRI (enquête ERE)
- Mettre en place un dispositif de monitoring rapide et transparent des dépenses relatives aux « *article processing charges* » et « *book processing charges* »

[22] https://scanr.enseignementsup-recherche.gouv.fr

Notons que ce dernier engagement portant sur les frais de publication (*Article Processing Charges - APC*) a déjà été tenu à travers le travail de coordination du consortium Couperin[23].
Le consortium Couperin a lancé en effet en 2017 une enquête sur la pratique des APC dans plusieurs universités et organismes de recherche français. Ce recueil d'informations a pour but d'obtenir une première idée de l'impact global de ces dépenses sur les budgets des établissements[24].

Ces derniers développements juridique ont fait l'objet d'un effort d'explicitation, par notamment la publication de guides d'analyse et d'application[2526].

### 3.5. Plan National pour la science ouverte

Enfin, le 4 juillet 2018, le Plan national pour la Science Ouverte a été annoncé par Frédérique Vidal, Ministre de l'Enseignement Supérieur, de la Recherche et de l'innovation. Il comporte 9 engagements :

1. Rendre obligatoire la publication en accès ouvert des articles et livres issus de recherches financées par appel d'offres sur fonds publics.
2. Créer un fonds pour la science ouverte.
3. Soutenir l'archive ouverte nationale HAL et simplifier le dépôt par les chercheurs qui publient en accès ouvert sur d'autres plateformes dans le monde.
4. Rendre obligatoire la diffusion ouverte des données de recherche issues de programmes financés par appels à projets sur fonds publics.
5. Créer la fonction d'administrateur des données et le réseau associé au sein des établissements.
6. Créer les conditions et promouvoir l'adoption d'une politique de données ouvertes associées aux articles publiés par les chercheurs.
7. Développer les compétences en matière de science ouverte notamment au sein des écoles doctorales.
8. Engager les opérateurs de la recherche à se doter d'une politique de science ouverte.
9. Contribuer activement à la structuration européenne au sein du *European Open Science Cloud* et par la participation à *GO FAIR*.

## 4. Process

A partir de 2012, le secrétariat général pour la modernisation de l'action publique[27] souhaite définir une politique interministérielle, qui fasse écho à des dynamiques sectorielles déjà très fortes (secteurs de la

[23] Le consortium Couperin (Consortium unifié des établissements universitaires et de recherche pour l'accès aux publications numériques) est une association à but non lucratif financée par les cotisations de ses membres et subventionnée par le Ministère de l'Enseignement Supérieur et de la Recherche. Elle regroupe 253 membres : 108 universités et établissements assimilés, 29 organismes de recherche, 87 écoles, 3 bibliothèques dotées de la personnalité morale et 26 autres organismes ayant une mission d'enseignement supérieur ou de recherche.
https://www.couperin.org/breves/1339-couperin-org-fournit-les-premieres-donnees-apc-pour-la-france

[24] Depuis le 8 juin 2018, une première série de données, provenant de 10 institutions, est publiée sur le site d'OpenAPC.L'ensemble de données initial couvre les frais de publication pour 695 articles publiés en 2015. Les dépenses totales s'élèvent à 1 173 258 € auprès de 55 fournisseurs et *la redevance moyenne à 1 688 €*.

[25] Ouverture des données de recherche – Guide d'analyse du cadre juridique en France http://www.bibliothequescientifiquenumerique.fr/wp-content/uploads/2018/03/Guide_analyse_Cadre_Juridique_Ouverture_donnees_Recherche_V2_licenceOuverte_prefaceDGRI.pdf

[26] Guide d'application de la loi pour une République numérique (article 30) Écrits scientifiques http://www.bibliothequescientifiquenumerique.fr/wp-content/uploads/2018/05/180221_Guide-d%E2%80%99application-de-la-loi-article30-version-courte.pdf

[27] Le SGMAP est une administration française ayant existé entre 2012 et 2017, placée sous l'autorité du Premier ministre et rattachée au secrétaire général du Gouvernement. Le SGMAP a laissé la place en novembre 2017 à la direction interministérielle de la transformation publique (DITP) et à la direction interministérielle du numérique et du système d'information et de communication de l'Etat (DINSIC).

santé, des transports notamment) – cette réflexion a mené à la création de l'AGD. Pour les actions de l'AGD, certaines études et consultations (4.1) ont permis de dessiner le cadre de l'action de l'AGD, avant de mettre en place des actions directement relevant de l'AGD (4.2) ou en lien avec son action (4.3)

### 4.1. Etudes et consultations

La grande problématique dans l'approche consistant à développer de l'*offre de données (la mise à disposition de données)* est de faire en sorte que l'ensemble des acteurs s'empare des données mises à disposition ? En regard, quels sont les engagements que la puissance publique doit prendre ? Pour le comprendre, Etalab[28] a mené une consultation publique auprès des utilisateurs potentiels des données de référence. Cette consultation[29], auprès d'acteurs publics et privés ainsi que des associations, a notamment permis d'identifier précisément les attentes, en particulier sur les critères de qualité des données de référence.
**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.
Pour le critère relatif à la fraîcheur des données publiques, la fourniture de données publiques n'a pas pour le moment atteint un niveau de qualité tel que l'infrastructure elle-même devienne invisible, comme un service physique de fourniture d'eau potable ou d'électricité par exemple. 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 : « 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.

### 4.2. Actions de l'AGD

#### Action - 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

[28] Etalab, voir section 4.3, est une mission créée en 2011, chargée de la politique d'ouverture et de partage des données publiques du gouvernement français, pour tendre vers un gouvernement ouvert. Etalab développe et maintient notamment le portail des données ouvertes du gouvernement français data.gouv.fr. La mission Etalab est rattachée à la Direction interministérielle du numérique et du système d'information et de communication de l'État (DINSIC), direction du Secrétariat d'État au Numérique du Premier ministre.
[29] https://www.etalab.gouv.fr/consultation-spd

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.
Pour illustrer les activités relevant de collège scientifique, la réunion de ce collège scientifique de décembre 2017 a permis de 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 (rôle d'explicitation de la « boîte noire » algorithmique).

**Action – Saisie de l'Administrateur Général des Données**

Par ses missions, l'administrateur général des données se place en facilitateur de la relation entre citoyen et administration, ou entre deux administrations. Son but est d'améliorer la circulation des données afin de favoriser la transparence et l'innovation et de permettre des politiques publiques plus efficaces car orientées par la donnée. Il intervient notamment en proposant son expertise aux administrations afin de les aider sur les questions juridiques et techniques. Les saisines, par simple formulaire numérique [30], sont effectuées à ce stade par des particuliers ou par des agents des administrations publiques.
Un exemple de saisine a été de clarifier les contours des nouvelles zones touristiques internationales sur Paris[31], explicitant les adresses à partir des zones définies par arrêtés[32].

**Action - Adapter et faire évoluer le cadre juridique**

Pour ces actions qui trouvent parfois leurs limites, et 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, comme vu en sections 4.2 et 4.3, en particulier par les lois *Valter* (sur la gratuité des données publiques) et *La loi pour une 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 :

- les documents communiqués à la suite d'une demande d'accès
- les documents figurant dans les répertoires d'informations publiques
- les bases de données mises à jour de façon régulière
- les données, mises à jour de façon régulière, présentant un intérêt économique, social, sanitaire ou environnemental

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

Une évolution majeure du cadre juridique 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

[30] https://agd.data.gouv.fr/saisines-de-lagd/formulaire-de-saisine

[31] https://agd.data.gouv.fr/2016/10/31/de-larrete-ministeriel-au-fichier-geojson-2

[32] Ce dispositif est à distinguer de ceux de la Commission d'accès aux documents administratifs (CADA, www.cada.fr), autorité administrative indépendante créée en 1978 qui a pour objectif de faciliter et contrôler l'accès des particuliers aux documents administratifs.

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.
Afin d'encourager l'écosystème naissant autour des données de la commande publique, il est apparu nécessaire de standardiser celles-ci[33]. 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 relatif aux données essentielles dans la commande publique[34], ainsi que celui relatif aux fonctionnalités et exigences minimales des profils d'acheteurs[35].

### 4.3 Actions en lien avec l'AGD

**Action - programme « Entrepreneur.e d'intérêt général »**

Le Programme Entrepreneur.e d'Intérêt Général a été lancé en 2016 par la Présidence de la République. Ce programme a vocation à constituer une promotion de talents extérieurs à l'administration, recrutés pour 10 mois, pour résoudre, par leurs compétences numériques d'exception et grâce aux données, des défis d'intérêt général, **au sein des ministères**.
Ce programme est financé via le programme d'investissement d'avenir (PIA), dans le cadre du fonds « Transition numérique de l'Etat et modernisation de l'action publique », doté de 126 M€ dès 2014. Plusieurs appels à projet ont pu être lancés, pour des cofinancements, typiquement entre 500 k€ et 3 M€ par projet (pour un budget total double), dont l'appel à projets « Industrialisation de la mise à disposition des données ouvertes », permettant de soutenir des projets d'amorçage en laissant sa place à l'expérimentation et en assumant les risques de l'innovation.

Lancé fin 2016, le programme « Entrepreneur d'intérêt général » constitue un programme d'innovation original aussi, surtout, sur son volet opérationnel. En effet, 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.

Les lauréats portent des projets dans des secteurs variés, dont nous ne retiendrons ici quelques résultats pour exemple[36] :

- *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

[33] 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 ». 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.
[34] https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000034492587&categorieLien=id
[35] https://www.legifrance.gouv.fr/affichTexte.do%3Bjsessionid=00B73A5DA9B3A710ABD6B312CD109476.tpdila16v_3?cidTexte=JORFTEXT000034492557&dateTexte=&oldAction=rechJO&categorieLien=id&idJO=JORFCONT000034491769
[36] https://entrepreneur-interet-general.etalab.gouv.fr

- *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)
- *Signaux Faibles*, porté par la DIRECCTE Bourgogne-Franche-Comté[37], vise à développer les outils pour repérer les entreprises en difficulté
- 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

## 5. Adoption and implementation of the initiative

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.

**Produire les données essentielles**

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 (voir *focus* infra).

Les données de référence, telles que définies dans le code des relations entre le public et les administrations (CRPA) répondent à trois critères :

(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.

[37] DIRECCTE : Direction régionale des Entreprises, de la Concurrence, de la Consommation, du Travail et de l'Emploi, service déconcentré du Ministère de l'Économie et des Finances

| | 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 |

Les neuf données de référence : des producteurs diversifiés.

*Focus Système informatique pour le répertoire des entreprises et des établissements*

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, en open data, depuis le 4 janvier 2017, en application des dispositions de la loi pour une République numérique

**Exemple original de coconstruction : la *base adresse nationale***

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.

**Généralisation de mise à disposition de données (de qualité, fiables, mises à jour, disponibles) & conception et déploiement d'outils et de dispositifs pour faire circuler les données**

La stratégie de diffusion des données ouvertes s'appuie sur la plateforme *data.gouv.fr*, exploitée par Etalab, qui contribue aux missions de l'Administrateur Général des Données. La plateforme compte aujourd'hui plus de 33 000 jeux de données ouverts par plus de 1 200 organisations.

Ce sont maintenant près de 200 000 visiteurs uniques mensuels sur la plateforme, pour plus de 33 000 jeux de données ouverts par plus de 1 200 organisations). Le flux de visites (et de requêtes) impose parfois un renvoi vers d'autres sites dédiés, comme le succès de l'API de géocodage de la Base Adresse Nationale, qui pour l'année 2017 a totalisé plus d'un milliard de requêtes, et près de 15 millions de visiteurs uniques – au-delà des chiffres, c'est bien des caractéristiques d'un mouvement automatisé et massif qui se dessinent.

Au-delà de la mise à disposition « simple » des données, il y a aussi le développement de modèles permettant l'interprétation, comme par exemple l'application *OpenFisca*[38], qui a pour objectif de

[38] https://openfisca.org/fr/index.html

« transformer le code législatif en code logiciel », qui utilise un moteur de calcul libre et ouvert qui permet des simulations transparentes et collaboratives.

**Cas des données issues d'infrastructures de recherche : la stratégie nationale des infrastructures de recherche**

La production, le stockage et la mise à disposition de données sont des paramètres essentiels de la recherche d'aujourd'hui ; c'est particulièrement vrai pour les infrastructures de recherche. Certaines sont dédiées au numérique, pour développer les outils de calcul intensif, de transmission ou de stockage des données. Dans certains domaines, cette mise à disposition est immédiate et entièrement publique. Dans d'autres, une période d'embargo est la pratique courante avant dissémination.
La Stratégie Nationale des Infrastructures de Recherche[39] caractérise les données, infrastructure par infrastructure, et explicite les modalités d'accès (par exemple, immédiatement disponible, sous embargo, etc.).

Ces données scientifiques, comme vu pour le cadre plus général de la donnée publique en section 4.4, ont d'avantage d'intérêt si elles sont produites de manière harmonisée. Cet impératif d'harmonisation impacte déjà les pratiques et les cultures de gestion et de partage des données qui varient entre les domaines, les communautés, les pays et les organisations. À mesure que ces derniers deviennent plus exigeants en matière de données, il convient d'optimiser leur conservation et leur réutilisation pour dynamiser les développements technologiques et sociétaux, ainsi que le partage des résultats de la recherche entre ses disciplines.
L'observation, la mesure, le calcul intensif, le stockage et le partage de données supposent de grands instruments portant les capacités techniques au-delà de l'existant et intégrant la porosité disciplinaire, source d'innovation. Ces outils sont les conditions des futures découvertes tout autant que le produit des dernières avancées scientifiques et technologiques. De grands équipements ont ainsi été créés, pilotés par des organisations nationales, européennes ou internationales, nécessitant une instrumentation de premier plan mais aussi des ressources humaines et financières conséquentes, grâce au soutien de la puissance publique.

Parallèlement à ces grands programmes se sont développés, ces dernières années, une quantité d'instruments partagés entre de nombreux acteurs sur le territoire : nouveaux modes de microscopie et d'imagerie, nouveaux dispositifs de criblage à haut débit, expériences virtuelles, bases de données sociales, environnementales et de santé, corpus de textes numérisés enrichis d'outils d'exploitation... En France, le soutien du Programme d'Investissements d'Avenir a été essentiel à ce succès.

Elle est aussi l'occasion de mener une enquête sur l'ampleur de la production de données scientifiques, actuelle et envisagée à cinq ans ainsi que leur management. La mise à disposition des données de la recherche, dont les infrastructures de recherche sont un pourvoyeur très important, est une exigence qui s'impose dorénavant à toute la communauté scientifique. Elle soulève néanmoins un défi énorme de dimensionnement de nos e-infrastructures impactées par les besoins considérables qui s'annoncent en matière de stockage, de flux de données et de moyens de calcul qu'il faut anticiper.

> La très grande majorité des infrastructures de recherche produisent, manipulent, traitent et/ou échangent des données. Les infrastructures les plus grandes consommatrices de moyens de stockage ont déclaré un total de 540 Pétaoctets (Po). À l'horizon de 5 ans, ce chiffre devrait être multiplié par 5.

Notre stratégie nationale se construit naturellement en relation avec celle de l'Europe. Au niveau des e-infrastructures, plusieurs projets ambitieux vont marquer le paysage pour les années à venir. La Commission Européenne a ainsi lancé une *European Cloud Initiative*, avec notamment le volet *European Open Science Cloud*.

[39] Stratégie Nationale des Infrastructures de Recherche, édition 2018 http://cache.media.enseignementsup-recherche.gouv.fr/file/Infrastructures_de_recherche/70/3/Brochure_Infrastructures_2018_948703.pdf

*Focus infrastructures de recherche en SHS*

Le développement du numérique combiné à celui de l'internet a donné accès à des données massives et aux moyens de calculs qui autorisent leur traitement, ainsi qu'à de nouveaux modes d'analyse de sources non numériques. Les disciplines des sciences humaines et sociales (SHS) se trouvent ainsi confrontées à une dynamique qui transforme le métier même du chercheur. Ainsi les infrastructures de recherche en SHS doivent permettre de constituer et de manipuler des corpus volumineux et très hétérogènes, de nature qualitative ou quantitative, susceptibles d'ouvrir de nouvelles voies de recherche et de favoriser l'interdisciplinarité. Inscrites dans un espace social largement ouvert au monde, les infrastructures contribuent à une meilleure valorisation d'un patrimoine scientifique et culturel et intéressent tous les établissements regroupés au sein de l'Alliance Athéna.

| CATÉGORIE | NOM | NOM COMPLET | ESFRI |
|---|---|---|---|
| *TGIR* | *Huma-Num*[1] | *Humanités Numériques* | *DARIAH (2006)*<br>*CLARIN (2006)* |
| TGIR | Progedo | PROduction et GEstion de DOnnées | ESS (2006)<br>CESSDA (2006)<br>SHARE (2006)<br>GGP (2016) |
| *IR* | *ERIHS-FR*[2] | *European Research Infrastructure for Heritage Science* | *ERIHS (2016)* |
| *IR* | *METOPES*[3] | *Méthodes et outils pour l'édition structurée* | |
| *IR* | *OpenEdition*[4] | *Edition électronique ouverte en Sciences humaines et sociales* | |
| *IR* | *RnMSH*[5] | *Réseau national Maison des Sciences de l'Homme* | |

*Liste des infrastructures françaises de recherche dans le domaine des sciences humaines et sociales*

**Plan National pour la Science Ouverte**

La diversité des données de recherche est extrême, et à côté des « big sciences », aux données très structurées et partageant des pratiques communément admises par les communautés disciplinaires ou thématiques, environ 80% des communautés appartiennent à la « longue traine » (« long tail »)[40], qui travaillent avec des données de façon moins coordonnée et normalisée.

L'ambition est de faire en sorte que les données produites par la recherche publique française soient progressivement structurées en conformité avec les principes FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable), préservées et, quand cela est possible, ouvertes. Lors de l'annonce du plan « Intelligence artificielle » au Collège de France, le 29 mars 2018, le Président de la République a annoncé la mise en place d'un principe d'ouverture par défaut pour toutes les données publiées dans le cadre d'appels à projet sur fonds publics. Cette obligation sera limitée par les exceptions légitimes encadrées par la loi, par exemple en ce qui concerne le secret professionnel, les secrets industriels et commerciaux, les données personnelles ou les contenus protégés par le droit d'auteur. Elle sera par ailleurs encadrée par les bonnes pratiques définies par chaque communauté scientifique, par exemple pour définir des durées d'embargo.

D'autre part, le ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation se dotera d'un administrateur des données de la recherche. Il travaillera avec l'administrateur général des données et animera un réseau des administrateurs de données de la recherche dans les établissements concernés. Un appel « flash » de l'ANR permettra d'accélérer la structuration de la communauté scientifique afin de promouvoir les principes « FAIR » et de développer l'ouverture des données. D'une façon générale, les dépenses de traitement des données seront éligibles dans les appels à projets.

Les chercheurs seront invités à déposer les données dans des entrepôts de données certifiés, dont la gouvernance et les règles de propriété intellectuelle seront conformes aux bonnes pratiques. À ce titre, les infrastructures nationales et européennes de recherche seront privilégiées, notamment via des centres de données thématiques et disciplinaires.

[40] Christine Borgman, *Big Data, Little Data, No Data: Scholarship in the Networked World*, MIT Press, 2015.

Les plans de gestion des données, instrument de définition des règles de construction, conservation et diffusion des données, seront généralisés.
La France apportera son soutien à la Research Data Alliance (RDA), réseau international définissant les bonnes pratiques dans le domaine des données de la recherche. Elle soutiendra aussi le développement et la conservation des logiciels, support indissociable des connaissances techniques et scientifiques de l'humanité. Dans cette optique, elle apportera son soutien à Software heritage.
Dans le cadre du soutien public aux revues, la France recommandera l'adoption d'une politique de données ouvertes associées aux articles et le développement des data papers. Une politique similaire concernant les thèses sera également mise en place.

Ce plan est doté d'un budget de 5,4 M€ la première année et de 3,4 M€ les années suivantes, avec une mise à jour prévue tous les deux ans, en s'appuyant notamment sur le Comité pour la Science Ouverte, qui rassemble plus de 200 experts du domaine.

## 6. International aspects

La science ouverte est un mouvement mondial qui, bien qu'il dispose de caractéristiques régionales, ne pourra se développer que dans le cadre d'une forte coordination internationale. La France souhaite y prendre sa part en défendant l'idée d'un écosystème efficace, régulé, transparent et résilient, se plaçant au service de la communauté scientifique et de la société. Elle contribuera à la structuration de ce paysage international, dans le domaine des services, des standards et des bonnes pratiques, à la fois à travers un renforcement de sa participation dans les infrastructures européennes et internationales de la science ouverte.

### Engagements internationaux

Dans le prolongement de la loi pour une République numérique qui incite les établissements publics à rendre leur données ouvertes et réutilisables, la France a inscrit son action en faveur de plusieurs mouvements internationaux :
*Données publiques :*

- elle s'engage à construire un écosystème de la science ouverte dans le cadre du partenariat pour un **gouvernement ouvert**, qui s'est concrétisé notamment par l'adhésion de la France à la dynamique *Open Government Partnership* (OGP)[41], moteur en particulier pour les standards de données sur la Commande Publique Ouverte[42], en s'engageant avec cinq autres pays pour développer et promouvoir un format standard de données ouvertes liées aux marchés publics (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 ;

*Données de la recherche :*

- elle participe à la définition et à la construction de l'**EOSC** (European Open Science Cloud)[43] ;
- elle a rejoint l'initiative **Go FAIR**[44] comme membre fondateur en 2017[45], avec les Pays-Bas et l'Allemagne – initiative qui propose le développement d'un environnement international de recherche enrichi par les données ;
- elle va adhérer au niveau national à ORCID, système d'identification unique des chercheurs qui permet de connaitre plus simplement et sûrement les contributions scientifiques d'un chercheur ;

[41] https://www.opengovpartnership.org/about/ogp-steering-committee/membership
[42] https://www.open-contracting.org/2016/12/07/open-contracting-version-francaise
[43] http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
[44] https://www.go-fair.org/go-fair-initiative/mission
[45] http://www.enseignementsup-recherche.gouv.fr/cid124728/science-ouverte-la-france-rejoint-go-fair-en-tant-que-co-fondatrice.html

• elle crée la Fondation franco-néerlandaise DOAB (Directory of open access books), afin de mettre au point une certification internationale de la qualité des ouvrages scientifiques en accès ouvert ;
• elle souhaite contribuer aux infrastructures de la science ouverte comme le DOAJ, OpenAIRE, SCOSS, OPERAS, Crossref et DataCite.

*Focus Research Data Alliance*

La *Research Data Alliance* regroupe avec près de 7 000 personnes physiques adhérentes, réparties dans 137 pays. Les membres de la *RDA* collaborent à travers le monde pour traiter de multiples enjeux en matière de structuration et de partage de données (Reproductibilité, Conservation des données, Interopérabilité juridique, Citation des données, Registre de types de données, Métadonnées, Bonnes pratiques pour les entrepôts disciplinaires, etc.) En 2018, la France soutient cette initiative avec la création du volet national *Research Data Alliance France*[46].

## 7. Monitoring and evaluation

### Benchmark d'initiatives européennes sur les *données*

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, circonscrits ici sur les seules *données*, nous proposons d'analyser et comparer les initiatives menées dans certains pays européens : au Royaume-Uni, au Danemark et en Estonie avec les initiatives nationales[47].

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, 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 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.

[46] https://www.rd-alliance.org/groups/rda-france
[47] Administrateur général des données, rapport au premier ministre « La donnée comme infrastructure essentielle » (pp 47-56), 2017, éd. La Documentation française

| | 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 |

*Tableau de synthèse de comparaisons internationales*[47]

**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 ainsi que des investissements. 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. La donnée comme infrastructure essentielle 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 National pour la Science Ouverte**

Pour le suivi des mesures annoncées dans le cadre du Plan national pour la Science Ouverte, est décidée la création d'un baromètre national de la science ouverte, qui se concentre dans un premier temps sur l'estimation de la part de la production de publication scientifiques qui sont diffusées en accès ouvert. Le pilotage de l'ouverture des publications et des données de la recherche impose en effet la constitution de tableaux de bords. L'*open science monitor* de l'union européenne[48] répond à la même nécessité. Il faut dire que beaucoup des données disponibles historiquement sont produites dans des environnements propriétaires.

La CURIF (Coordination des universités de recherche intensive françaises) qui regroupe 18 universités françaises parmi les plus importantes en terme de recherche, a décidé de s'engager fortement pour la science ouverte en juillet 2018. Elle contribue ainsi à l'implémentation concrète des principes de la science ouverte dans les universités françaises.[49]
L'outil de référence pour superviser les politiques dans de la domaine de la science ouverte est ROARMap (Registry of Open Access Repository Mandates and Policies[50]), composante du projet *EPrints* de l'Université de Southampton.

**Autres éléments de *benchmarking***

Des éléments de comparaison internationale sont parus en septembre 2018, l' « Open Government Data » de l'OCDE[51], ainsi que l'*OpenData Barometer*, de la World Wide Web Foundation[52]. Nous pouvons voir que la France est systématiquement présente dans les 5 premiers des différents classements, ce qui rend compte d'un modèle de déploiement équilibré.

## 8. Lessons and Challenges ahead

### 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. 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 ou encore des conventions de subvention – cf. supra en section 4.4 le travail de standardisation de la commande publique ; le même travail de standardisation a été effectué pour les données relatives aux subventions.

### Accélérer la dynamique par un travail en réseau

Pour accélérer la dynamique impulsée par l'Administrateur Général des Données au sein du gouvernement et des Administrations, un réseau national d'administrateurs ministériels des données est en cours de consolidation (déjà plusieurs ministères se sont dotés d'un tel administrateur dédié, dont le ministère de la Transition écologique et solidaire, le ministère de l'Intérieur, la direction générale des Finances publiques, le ministère de l'Agriculture – et annoncé pour le ministère de

[48] https://ec.europa.eu/info/open-science/open-science-monitor_en
[49] http://www.curif.org/fr/la-curif-sengage-pour-la-science-ouverte
[50] roarmap.eprints.org
[51] https://one.oecd.org/document/42201839/en/pdf
[52] https://opendatabarometer.org/doc/leadersEdition/ODB-leadersEdition-Report.pdf

l'Education [53], ainsi que le ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation[54]).

Les initiatives de la société civile sont également nombreuses. Nous pouvons présenter brièvement ici l'une d'entre elles. La Fing (Fondation Internet Nouvelle Génération) est une association créée il y a 17 ans, dont les ressources proviennent des cotisations de ses 330 adhérents, des contributions des partenaires et du soutien des entreprises, des collectivités locales et des acteurs publics à ses projets, et se définit comme un « *think & do-tank* ». Leur démarche *Open Data Impact*[55], avec des ateliers notamment sur le thème « Programmer/déployer » pour formuler les 10 pistes d'innovation ou la feuille de route de l'open data ces 10 prochaines années, propose de se projeter et retravailler les objectifs et les imaginaires à 10 ans, avec élaboration d'une feuille de route sous forme de scénarios. Une autre démarche propose des Conférence internationale annuelle sur la culture de la données (#DLC Data Literacy Conference) pour proposer au plus grand nombre les clés d'une « culture de la donnée » enfin accessible aux non-spécialistes – d'en comprendre les enjeux, d'en discuter les sources et les usages, et d'en tirer parti dans sa propre activité[56].

A l'international, comme vu en section 4.6, c'est l'implication volontariste de la France dans des organisations telles que l'*Open Government Partnership* (OGP), l'European Open Science Cloud (EOSC) ou bien encore l'alliance *Research Data Alliance*.

**Conforter le leadership sur les considérations éthiques**

La France avait déjà adoptée la loi Informatique et libertés le 6 janvier 1978, en vue notamment d'encadrer le traitement automatique de données à caractère personnel, et créant une autorité administrative indépendante, la Commission nationale de l'informatique et des libertés[57] (CNIL), ayant pour mission d'informer toutes les personnes concernées et tous les responsables de traitements de leurs droits et obligations, et veiller à ce que les traitements de données à caractère personnel soient mis en œuvre conformément aux dispositions de la loi précitée.

Le rapport 2018 de la mission parlementaire menée par Cédric Villani, « Donner un sens à l'intelligence artificielle - *Pour une stratégie nationale et européenne* »[58], préconise la création de centres nationaux experts sur IA est préconisée, comme par exemple DATAIA[59], situé au cœur du cluster d'innovation Paris-Saclay, en intégrant recherches technologiques, mais également sciences juridiques, sciences sociales, *etc*.

***Data litteracy* et longue durée**

Sans personne pour exploiter les données, il est possible que l'intérêt pour l'*Open data* s'essouffle, amenuisant de fait l'intérêt des producteurs et du public pour les données ouvertes. Pour éviter cette situation, et faire en sorte que le pouvoir des données soit partagé avec le plus grand nombre, plusieurs acteurs ont avancé le concept de *data literacy*[60]. D'une façon générale, la question du développement d'une culture des données apparaît comme la condition *sine qua non* au décollement d'une politique

[53] Annonce de Jean-Michel Blanquer, ministre de l'éducation nationale, inauguration du « *110 bis* » laboratoire d'innovation de l'éducation nationale le 5 Juin 2018

[54] Plan national pour la Science Ouverte, annoncé le 4 juillet 2018 par Frédérique Vidal, Ministre de l'Enseignement Supérieur, de la Recherche et de l'Innovation.

[55] fing.org/campagne-Open-Data-Impact http://fing.org/IMG/pdf/Open_Data_2025_officiel-3.pdf

[56] http://dataliteracyconference.net/2018

[57] Voir par exemple les rapports produits par cette commission, dont en 2017 son « Rapport sur les enjeux éthiques des algorithmes et de l'intelligence artificielle : Comment permettre à l'Homme de garder la main ? »

[58] Mission Cédric Villani « Donner un sens à l'intelligence artificielle », remis le 28 mars 2018, https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf

[59] https://dataia.eu

[60] Samuel Goëta, co-fondateur Datactivist, Livre en cours de publication sur l'*Open Data*

d’ouverture des données qui dépasse lois et règlements et devienne un usage par défaut. Ce changement de paradigme et de culture ne pourra se produire que sur la durée.