Comment l’ouverture des données publiques contraint l’état à focaliser ses investissements pour remplir ses nouvelles missions

par **Denis** **BERTHAULT**, Directeur du développement des contenus en ligne, LexisNexis.

# 

A

u cours de l’année 2015, Benoît Thieulin, alors président du Conseil national du numérique[[1]](#footnote-1), expliquait que le numérique a longtemps été considéré comme « une révolution médiatique », puis comme « une nouvelle économie » avant de comprendre en 2015 « que c’était un phénomène général de transformation de la société, qui n’épargnera aucun secteur ni aucun leader établi »[[2]](#footnote-2). Gilles Babinet, le digital champion français auprès de la commission européenne, relevait de son côté que la révolution digitale est souvent résumée à sa dimension technologique alors que « c’est avant tout un nouveau modèle de management qui permet l’innovation, la créativité et la prise de risque avec moins de hiérarchie. »[[3]](#footnote-3) Cette révolution s’applique aussi à l’administration et aux données qu’elle produit.

La mutation numérique de l’État progresse et s’accélère basculant sur son passage de nombreuses certitudes, hiérarchies, organisation, gouvernance… Le numérique, qui impose de la transversalité, du collaboratif et du partage, s’oppose à l’histoire même de l’administration française héritée du modèle napoléonien. Et pourtant, l’appareil réglementaire du 21 septembre 2015[[4]](#footnote-4) renforce et organisent les missions du secrétariat général pour la modernisation de l’action publique (SGMAP) et concentrent dans sa main de nombreux pouvoirs, dont certains essentiels pour les données publiques.

On pense évidemment à Etalab chargée de « coordonner les actions des administrations de l’État et leur apporte son appui pour faciliter la diffusion et la réutilisation de leurs informations publiques », à un niveau interministériel. Mais il faut aussi ajouter la création du poste d’Administrateur général des données et la nouvelle DINSIC qui regroupe en réalité 6 composantes : le RIE (Réseau interministériel de l’État), la DISIC (Direction interministérielle des systèmes d’information et de communication), le service d’administration numérique du SGMAP, Etalab (la mission chargée de l’open data), l’Administrateur général des données et l’incubateur de start-up d’État.

De même, la très récente création d’un poste de conseiller « ouverture des données et transformations numériques de l’action publique » auprès du ministre de l’Économie numérique démontre que la problématique des données publiques s’insère dans le mouvement plus large de la modernisation de l’État et de l’évolution de ses missions. Il en va de même avec le projet *France Connect* qui un exemple réussi de mutualisation et démontre que l’État, lui aussi, peut être agile.

Enfin, les dispositions de la future loi Lemaire sur le « service public de la donnée » et la mise à disposition des données par défaut vont créer un immense appel d’air et une demande phénoménale de données à laquelle le Gouvernement n’a pas les moyens budgétaires de répondre.

Dans ce nouveau contexte, il faut penser l’économie de la donnée au futur et réussir cette nouvelle révolution qui impacte les tréfonds de l’administration : investigations internes pour savoir qui fait quoi, l’audit technologique, la définition des stratégies, le choix des normes et standards des données, la coopération avec des écosystèmes endogènes (les autres administrations, les collectivités territoriales) et exogènes (associations, développeurs, start-up…). Il faut donc faire des choix, définir des priorités, bref, mécontenter…

# § 1 – L’État, les ministères, les établissements publics et leurs données

Les décideurs politiques et administratifs commencent à mesurer combien il est difficile de « sortir » les données utilisées par l’administration dans ses missions quotidiennes. Cela s’est encore vérifié lors des débats parlementaires du SNIIRAM, la gigantesque base de données préexistante de l’Assurance maladie dont l’ouverture était possible… moyennant de nombreuses adaptations. Ainsi, la conclusion devient simple : diffuser en *open data* coûte cher, pour des retours sur investissement modérés.

# § 2 – Pourquoi est-ce si cher ?

Parce que les données utilisées par l’État sont la plupart du temps des collections ou des empilages de documents, souvent non organisés, donc inexportables, *a fortiori* en open data. Au contraire : tout a été fait depuis des années pour les garder verrouillées. Aussi, tout un travail préparatoire s’impose pour décrire et organiser ces données éparses inexploitables qui, avec le « contexte métier » dans lequel elles ont été produites, leur donnent un potentiel de réutilisation infini. Ce n’est qu’une fois ce travail préalable réalisé qu’il sera possible d’apporter le niveau d’anonymisation en fonction des besoins de réutilisation.

Et d’ailleurs, disposer de données, c’est bien, mais pouvoir les utiliser/réutiliser, c’est mieux. Souvent, l’ouverture des données publiques est considérée comme une mise à disposition artisanale, « à la pièce », de données, même si ces dernières appartiennent au même corpus. Or, pour pouvoir traiter, analyser et exploiter les données à l’échelle industrielle et de manière globale, la création préalable (ou a posteriori, mais c’est plus coûteux) d’un corpus organisé sous forme d’ontologie (ou de base de données) est indispensable, que la donnée provienne d’un traitement de texte, d’un capteur ou d’une application mobile. C’est la raison pour laquelle une base de données normée est nécessaire, car elle joue, de fait, un rôle central. Face au déluge de données qui s’annonce, la nécessité de les modéliser, dès la conception des données, sous forme d’ontologie (5) est indispensable.

L’État doit donc se professionnaliser dans l’ouverture des données et mieux prendre en compte les données dont il peut entrevoir avec certitude qu’elles seront réutilisées.

Les réutilisateurs professionnels, rassemblés au sein du GFII, souhaitent qu’au-delà des données de transparence démocratique, l’État se focalise sur deux types de données : les données pivots et les données à fort potentiel de réutilisation.

# § 3 – Les données pivots

Les « données pivot » ou « données de référence » sont des données considérées comme identifiantes, par l’administration ou par l’usage, pour nommer ou identifier des produits, des entités économiques, des territoires ou des acteurs (personnes physiques et morales). Ces référentiels sont indispensables pour lier des bases de données de nature hétérogène et construire tout nouveau service intégré au web des données (Linked Open Data).   
Un format ouvert spécifique à chaque type de donnée devrait obligatoirement être utilisé, tout en laissant la possibilité de publier également selon certains formats propriétaires très usités. Ces formats ouverts devraient être documentés et stables, toute modification de format devant être précédée de procédures de concertation éliminant tout imprévu pour les réutilisateurs. Les formats de diffusion devraient s’appuyer, autant que possible, sur les normes internationales en vigueur dans les différents métiers, en particulier les normes ISO ou du W3C.   
Ces données devraient être ouvertes et mises à disposition gratuitement. Les lots de données devraient être exhaustifs par rapport au domaine décrit. La pérennité et la périodicité de la fourniture des données et de leurs éventuelles mises à jour devraient être garanties, ainsi que la traçabilité des traitements du producteur initial jusqu’à la diffusion.   
Une réflexion doit également être menée au niveau international. De nombreux référentiels sont en effet réalisés par des organismes internationaux auxquels participe activement la France. L’ouverture et la diffusion gratuite de ces données pivots doivent également être organisées et la France peut agir efficacement pour initier ce mouvement. Ces référentiels concernent de nombreux secteurs comme le domaine culturel (ISNI, ISSN, ISBN) ou le domaine de la santé (nomenclature SNOMED CT couvrant tous les champs de la médecine humaine et vétérinaire). Cette initiative mettrait une nouvelle fois la France à la pointe de l’innovation dans le domaine de la diffusion des informations publiques.  
En annexe sont reportées, une liste de ces données pivot, par secteur.

# § 4 – Élargir le périmètre des données réutilisables

À l’exception de l’intégration limitée des données détenues par les établissements culturels, le périmètre de la réutilisation visé par la Directive n’a pas subi de modification sensible. À l’occasion des textes d’application de la loi Valter (et, peut-être, de la loi Lemaire), il est nécessaire d’identifier les types d’informations qui, dans le cadre des missions de service public, sont utiles à la vie économique et doivent par conséquent entrer dans le champ de la réutilisation. Une liste de ces données a été constituée et est disponible [en ligne](http://www.gfii.fr/uploads/docs/GFII_Donneespivots.pdf).

# § 5 – Accompagner l’ouverture

Certaines de ces données demandent des traitements spécifiques préalables à leur ouverture. De nombreux jeux de données restent aujourd’hui inaccessibles, car les établissements publics ne bénéficient pas des budgets suffisants pour les mettre à disposition dans des formats lisibles par machine. Dans ces cas spécifiques, il serait souhaitable qu’un régime dérogatoire soit mis en place afin de permettre l’investissement qui rendra ces données accessibles[[5]](#footnote-5). Le périmètre de l’investissement pris en compte pour l’établissement d’un tarif devrait se limiter strictement aux traitements spécifiques et au développement de plateformes allant au-delà d’une simple mise à disposition des données produites dans le cadre de la mission de service public. Ainsi, les coûts pouvant être pris en compte pour définir l’assiette devraient être clairement définis : licence des logiciels utilisés spécifiquement pour l’anonymisation, l’exportation ou la conversion des données, adaptation de la base de données notamment pour l’exportation des données, améliorations/accroissement des capacités d’hébergement (hardware), infrastructures réseaux.

# Conclusions

La bataille des données est une bataille stratégique et risquée : elle concerne la position de la France vis-à-vis des autres pays, de l’administration nationale ou locale face à d’autres administrations nationales ou locales, de l’administration vers les collectivités territoriales (et réciproquement) et de cet ensemble vers le monde de l’entreprise ou associatif. L’État doit focaliser ses investissements en fonction de nombreux critères, dont ses besoins propres ne doivent pas être exclus. Il doit aussi se dépêcher : si le secteur public ne réussit pas à créer ou diffuser certaines bases de données, ou qu’elles n’ont pas le niveau de qualité attendu, le risque est grand que le secteur privé trouve des alternatives (3). En effet, le besoin est si prégnant pour tout un ensemble de professionnels (banques, assurances, industrie pharmaceutique, transport…) qu’ils se contenteront de bases de données de moindre qualité, non officielles, mais disponibles. C’est en train de devenir le cas pour les statistiques publiques en matière de chômage avec le « baromètre de l’emploi en France » créé par la société américaine Automatic Data Processing (ADP) qui concurrence l’INSEE, la Dares (le service statistique du ministère du Travail) et Pôle emploi (4). Il est enfin indispensable, quand les données font l’objet d’une tarification, de garder à l’esprit qu’il est possible, comme c’est le cas des données météorologiques concernant la France, de les obtenir via d’autres offices météorologiques européens à des conditions beaucoup plus avantageuses. Le dumping de la donnée existe aussi…

# Annexe : Exemples de données pivots

## Données géographiques

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Libellé** | **Nature des données** | | **Organisme producteur** | **Lot de données** | **Argumentaire pour l’ouverture** | **Méta données** | **Formats souhaités** | **précision de localisation** | **Services souhaités** | **Actualisation** |
| **Adresses** | | Géoloc | Communes, La Poste, INSEE, IGN | Unifier différentes sources concurrentes et hétérogènes | Référence pour les services à domicile, la sécurité et toutes activités | Type  INSPIRE | Format unique (INSPIRE) | décamétrique | Téléchargement service de type OpenLS | quotidienne |
| **Noms de lieux** | | Géoloc | Communes, IGN | Référentiel géographique à grande échelle (RGE) | Référence pour les services à domicile, la sécurité et toutes activités | type INSPIRE | Format unique (INSPIRE) | décamétrique | Téléchargement WMS, WFS, WMTS | quotidienne |
| **Limites de communes et limites corrélées (EPCI, arrondissements, départements, régions…)** | | Géoloc | Communes, cadastre, IGN | Unifier les sources concurrentes | Référence pour les zonages administratifs | type INSPIRE | Format unique (INSPIRE) | métrique et décimétrique dans le domaine urbanisé | Téléchargement WMS, WFS, WMTS | cohérente avec l’information légale (décrets) |
| **Zones de compétences administratives (juridictions, carte scolaire...)** | | Géoloc | Ministères | Formater les informations existantes | Informations administratives de référence | type INSPIRE | Format unique (INSPIRE) | Selon la nature des informations | Téléchargement WMS, WFS, WMTS | annuelle |
| **IRIS** | | Géoloc | INSEE — IGN | Diffusion payante existante | Indispensable à l’utilisation cohérente des informations statistiques | type INSPIRE | Format unique (INSPIRE) | métrique | Téléchargement WMS, WFS | annuelle |

## Données entreprises et organismes

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Libellé** | **Nature des données** | **Organisme producteur** | **Lot de données** | **Argumentaire pour l’ouverture** | **Formats souhaité** | **Services souhaités** | **Actualisation** |
| **Combinaison de : SIREN,  adresse, raison sociale, forme juridique** | Personnes morales, Entrepreneurs individuels. | INSEE | Base SIRENE, JO-ASSOCIATION | Données pivots pour la gestion des informations sur les personnes morales et/ou entrepreneurs individuels. | XML | Téléchargement OpenL | Quotidienne |
| **Siret (Siren + NIC)** | Personnes morales, Entrepreneurs individuels. | INSEE | Base SIRENE, JO-ASSOCIATION | Données pivots pour la gestion des informations sur les personnes morales et/ou entrepreneurs individuels. | XML | Téléchargement OpenL | Quotidienne |
| **RNA** | Personnes morales associatives. | Préfectures | Registre national des associations à constituer par défaut JO ASSOCIATION (pas complet, mais unique) | Données pivots pour la gestion des informations sur les personnes morales associatives loi 1901. | XML | Téléchargement OpenL | Quotidienne |
| **TVA Intracommunautaire  et SIREN** | Personnes morales, Entrepreneurs individuels. | Douanes (DGDDI) | Base SIRENE/N° de TVA Intracommunautaire | Données pivots pour la gestion des informations sur les personnes morales et/ou entrepreneurs individuels. | XML | Téléchargement OpenL | Quotidienne |
| **Numéro CHORUS** | comptabilité publique, pour identifier une administration |  |  |  |  |  |  |

## Données de santé

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Libellé** | **Nature des données** | **Organisme producteur** | **Lot de données** | **Argumentaire pour l’ouverture** | **Métadonnées** | **Formats souhaités** | **Précision de localisation** | **Actualisation** |
| **Professionnels paramédicaux**  **ADELI** | Identifiant, caractéristiques professionnelles, mode d’exercice, structure… | ARS-DREES | complet par profession | Nécessaire pour services d’orientation des citoyens et des professionnels, analyses... Des données sont ouvertes, mais revoir conditions et limitations | Profession, région, mode d’exercice — voir documents, nomenclatures (programme RASS ASIP) | XML | adresse | En permanence |
| **Professionnels médicaux FNPS->RPPS** | Identifiant, caractéristiques professionnelles, mode d’exercice, structure… | Ordres, CNAMTS, ASIP | complet par profession | Nécessaire pour services d’orientation des citoyens et des professionnels, analyses... Des données sont ouvertes, mais revoir conditions et limitations | Profession, région, mode d’exercice — voir documents, nomenclatures (programme RASS ASIP) | XML | adresse | En permanence |
| **Répertoire des établissements pharmaceutiques** | Entreprises pharmacie | ANSM |  | tous services étude du secteur, marché, transparence citoyens | Département, activité | XML | adresse | En permanence |
| **Nomenclature médecine humaine et vétérinaire SNOMED CT** | Tentative de terminologie clinique totale de la santé | IHTSDO |  | Terminologie intermédiaire pivot et pour codage  Demande ouverture international |  |  |  |  |

## Données des collectivités

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Libellé** | **Nature des données** | **Organisme producteur** | **Lot de données** | **Argumentaire pour l’ouverture** | **Formats souhaités** | **Services souhaités** | **Actualisation** |
| Combinaison de : SIREN,  adresse, raison sociale, forme juridique | Personnes morales, Entrepreneurs individuels. | INSEE | Base SIREN, JO-ASSOCIATION | Données pivots pour la gestion des informations sur les personnes morales et/ou entrepreneurs individuels. | XML | Téléchargement OpenL | Quotidienne |
| Siret (Siren + NIC) | Personnes morales, Entrepreneurs individuels. | INSEE | Base SIREN, JO-ASSOCIATION | Données pivots pour la gestion des informations sur les personnes morales et/ou entrepreneurs individuels. | XML | Téléchargement OpenL | Quotidienne |
| TVA Intracommunautaire  et SIREN | Personnes morales, Entrepreneurs individuels. | Douanes (DGDDI) | SIREN/N° de TVA Intracommunautaire | Données pivots pour la gestion des informations sur les personnes morales et/ou entrepreneurs individuels. | XML | Téléchargement OpenL | Quotidienne |
| Numéro HELIOS | Identifiant d’une collectivité territoriale dans le système informatique de l’État dédié au secteur public local. | Ministère des Finances (DGFIP) |  | Données pivots pour l’intégration des données comptables des collectivités territoriales | XML | Téléchargement OpenL | Quotidienne |

## Données environnementales

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Libellé** | **Organisme producteur** | **Lot de données** | **Argumentaire pour l’ouverture** | **Méta données** | **Formats souhaités** | **précision de localisation** | **Services souhaités** | **Actualisation** |
| Zones protégées environnementales | Muséum national d’histoire naturelle (MNHN) | Base des sites protégés de l’INPN (inventaire national du patrimoine naturel) | Connaissance partagée des zones naturelles à protéger | type INSPIRE | Format unique (INSPIRE) | métrique ou décamétrique selon la nature des informations | Téléchargement, WMS, WFS, WMTS | cohérente avec l’information légale (décrets et arrêtés) |

## Données de transport

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Libellé** | **Nature des données** | **Organisme producteur** | **Lot de données** | **Argumentaire pour l’ouverture** | **Formats souhaités** | **précision de localisation** | **Services souhaités** | **Actualisation** |
| **Point d’accès aux réseaux de transport** | Ponctuels ou objets complexes géoréférencés | 400 autorités organisatrices de transport (AOT) et 3000 opérateurs de transport | régionaux et nationaux ou par grands opérateurs (RFF, SNCF, RATP) | Nécessaires au calcul d’itinéraires multimodaux |  | NETEX, INSPIRE | Métrique | Téléchargementwms, wfs, georss |
| **Zones tarifaires de transport et identifiant trajet** | géolocalisé | 400 autorités organisatrices de transport (AOT) et 3000 opérateurs de transport | Données hétérogènes à intégrer dans un service mutualisé. | Information des usagers — favoriser les parcours intermodaux et la substitution des transports collectifs à l’automobile. |  | NETEX | Selon la nature des informations | Téléchargementwms, wfs, georss |
| **Périmètres de transport urbain** | Périmètres des agglomérations desservies | Environ 200 agglomérations urbaines | Lot de données France entière | Analyse des dessertes d’agglomérations urbaines |  | INSPIRE |  |  |
| **Accessibilité PMR** | Information PMR pour les transports | 400 AOT | Par région et France entière | Contrôle de légalité ; services aux PMR |  | NETEX |  |  |

## Données culturelles

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Libellé** | **Nature des données** | | **Organisme producteur** | **Lot de données** | **Argumentaire pour l’ouverture** | | **Métadonnées** | **Formats souhaités** | **Services souhaités** |
| **Bases vocabulaires de la base Patrimoine (Mérimée, Palissy, Mémoire, Auteur)** | Vocabulaire — auteur et thesaurus | | Ministère de la Culture |  | Jeu de données de référence vers les données patrimoniales :  la base Archi-XX | | Répertoire des auteurs -- Thesaurus (types d’édifices et types d’objets) communs aux bases Mérimée, Palissy et Mémoire | XML, RDF | Accès via une URI pérenne ; récupération dans des formats structurés |
| **Thésaurus iconographique Garnier** | Thesaurus iconographique des musées | | Ministère de la Culture |  | Thesaurus utilisé pour l’indexation des documents iconographiques des musées français, en particulier dans la base Joconde. | | Thésaurus des sujets représentés | XML, RDF | Accès via une URI pérenne ; récupération dans des formats structurés |
| **Référentiels auteur de la base Joconde** | Auteurs (personnes, organisations) | | Ministère de la Culture | Plus de  30 000 auteurs | Référentiel d’auteurs permettant d’accéder aux ressources des musées de France. | | Répertoire des auteurs | XML, RDF | Récupération dans des formats structurés. |
| **ISNI** | Identifiant international normalisé (ISO) | | Agence internationale ISNI — BnF Agence d’enregistrement | 7 millions d’identités publiques | Les données sont déjà accessibles en RDF et XML par négociation de contenu et sous Open License | | Identifiants des identités publiques (personnes et organisations). | XML, RDF | Ouvrir un dump complet de données Objectif : ouvrir un dump complet de données en RDF ou XML et en Open data. |
| **ISSN** | Identifiant international normalisé (ISO) | ISSN International Center — Attribution par le Centre ISSN France pour les publications en séries éditées en France. | |  | Permettre l’échange des publications périodiques, **pour l’ensemble des publications (françaises et étrangères)**. Favoriser l’interopérabilité dans le secteur du livre (bibliothèques, éditeurs, diffuseurs). | Identifiant des publications en séries (périodiques, collections éditoriales). | | XML, RDF |  |

1. Depuis le 8 février 2016, il a été remplacé par Mounir Mahjoubi comme président du CNNum. [↑](#footnote-ref-1)
2. http://www.lesechos.fr/tech-medias/hightech/021579824441-numerique-pourquoi-2015-est-un-grand-cru-pour-la-france-1187704.php. [↑](#footnote-ref-2)
3. http://www.lesechos.fr/tech-medias/hightech/021579823941-gilles-babinet-les-entreprises-francaises-sont-encore-trop-timorees-1187703.php?xxDp4EeL3r3Pu8k4.99 [↑](#footnote-ref-3)
4. Décret n° 2015-1165 du 21 septembre 2015 relatif au secrétariat général pour la modernisation de l’action publique – Arrêté du 21 septembre 2015 portant organisation du secrétariat général pour la modernisation de l’action publique. [↑](#footnote-ref-4)
5. Dir. 2013/37/UE du Parlement européen et du Conseil du 26 juin 2013 modifiant la directive 2003/98/CE concernant la réutilisation des informations du secteur public, art. 6-2b – transposé par L. n° 2015-1779 du 28 décembre 2015 relative à la gratuité et aux modalités de la réutilisation des informations du secteur public, art. 5. [↑](#footnote-ref-5)