Sens [public]
Revue internationale
International Web Journal
www.sens-public.org

Data.gouv.fr : de l'ouverture des données à l'ouverture des possibles ?

Yannick Maignien

Data.gouv.fr : de l'ouverture des données à l'ouverture des possibles ?

Yannick Maignien

ETALAB… Données publiques et laboratoire d'État…

Le 5 décembre 2011, le Gouvernement a ouvert le portail data.gouv.fr préparé depuis 9 mois par la mission ETALAB dirigée par Séverin Naudet. Ce portail d'ouverture des données publiques réunit à son inauguration environ 330 000 jeux de données[1], c'est-à-dire des fichiers ou base de données administratives de l'État, des collectivités territoriales[2] ou d'établissements publics, et même d'entreprises de l'État (SNCF…) ou EPIC (non inclus a priori dans cette obligation d'ouverture de publication).

Le but avancé de cette opération est la transparence de l'accès et la réutilisation des données publiques. Transparence inaugurée en 1978 par la loi sur l'accès aux documents administratifs. La directive européenne de 2003[3] portant sur la réutilisation des informations du secteur public a modifié cette loi transposée dans le droit français en 2005.

La mise en œuvre du portail data.gouv.fr intervient dans un contexte politique de campagne électorale propice aux bilans – y compris numériques ! – et dans une problématique d'innovation exacerbée par la crise économique et les interrogations sur la croissance.

Au-delà de ces tenants et aboutissants, l'open data à la française s'inscrit dans une évolution sur Internet ouverte par les États Unis : data.gov en mars 2009 (Barak Obama) et le Royaume Uni : data.gov.uk en janvier 2010 (Gordon Brown).

---

[1] De fait, une forte majorité des fichiers (281 910) provient de six fichiers édités par l'INSEE sur le recensement de la population. La page http://www.recensement.insee.fr/basesChiffresCles.action agrège les milliers de communes, arrondissements et cantons dont relèvent ces données.

[2] Cf. Guide pratique de l'*ouverture des données publiques*, édité par la Fing. Janvier 2011. www.enssib.fr/bibliotheque-numerique/document-49064

[3] Directive européenne proposée à modification ce 12 décembre 2011 : « The European Commission will adopt an Open Data Strategy – a set of measures aimed at increasing government transparency and creating a €32 billion a year market for public data. » Voir l'intervention de Mme Neelie KROES, Vice-President of the EC in charge of Digital Agenda.
http://ec.europa.eu/avservices/video/videoplayer.cfm?ref=81912 « Turning government data into gold » !

Article publié en ligne : 2012/01
http://www.sens-public.org/article.php3?id_article=883

Par ailleurs, comme le rappelle Data Publica[4] :

> « certaines collectivités locales avaient, elles aussi, déjà entamé cette révolution avec, par exemple, Rennes (septembre 2010), Paris (janvier 2011), le conseil général de Saône-et-Loire (CG71, septembre 2011), Nantes (novembre 2011), cette accumulation augmentant la pression sur le projet de l'État (voir la carte éditée par Libertic) ».

Ce phénomène d'ouverture des données publiques représente une transformation profonde des conditions d'accès et de réutilisation des contenus numériques. Nous voudrions surtout souligner ici, avec l'*Open public data*, combien est initié en même temps qu'un régime inédit du Web, une ouverture de ses possibles techniques, éditoriaux, économiques et politiques.

L'enjeu technique des données structurées au Web sémantique

Les données numériques, accumulées depuis des années, en base de données, voient non seulement leur contenu « brut » mis à disposition, mais également et graduellement les relations et structures de tables de données. Celles-ci, par leur structuration syntaxique et leur contexte des domaines décrits (ensembles statistiques, comptables, démographiques, scientifiques, etc.) faisaient jusqu'à maintenant partie du « Web profond », inaccessible par les navigateurs Web, sauf par le truchement d'outils de requêtes spécifiques et dédiés au type de base de données utilisées (souvent propriétaires).

Depuis une quinzaine d'années, sous l'impulsion de Tim Berners Lee[5], le développement d'une interopérabilité fondée sur de nouveaux standards, propose de privilégier les données, de « libérer les données », et de développer, sous l'étiquette de Web sémantique, les outils automatiques de traitement de l'information structurée à l'échelle de l'ensemble du Web.

Les formats de données ont ainsi évolué radicalement, au point que le Web sémantique permet d'exprimer complètement le modèle logique relationnel minimal (triplet sujet-prédicat-objet) d'unités de sens clairement identifiées (URI). Depuis une dizaine d'année ce modèle RDF[6] du W3C pour le Web sémantique permet d'exprimer l'indépendance et la richesse relationnelle de grands ensembles de données, et d'interconnecter ces ensembles de « triple store », afin de les faire communiquer entre eux « automatiquement », c'est-à-dire de machine à machine, à la demande des utilisateurs et de leurs requêtes complexes. Cette rupture à l'égard du Web de documents, de sites Web HTML, permet d'aller beaucoup plus loin que les liens hypertextes, qui

[4] http://www.data-publica.com/content/2011/12/data-publica-salue-la-naissance-de-data-gouv-fr/

[5] On connaît son récent engagement pour la libération des données « raw data, now ! »

[6] RDF, Resource Description Framework.

doivent être préalablement implémentés, afin d'ouvrir aux « humains » les possibilités de navigation intertextuelles ou contextuelles. Les arborescences les plus complexes, mais cloisonnées, hétérogènes, deviennent transposables dans des graphes, déformables quantitativement et qualitativement, peuplant l'ensemble de l'Internet mondial.

En parallèle, des langages d'expression des cohésions sémantiques de domaines spécifiques (scientifiques, documentaires, conventionnels…) de plus en plus clairement identifiés dans des vocabulaires types, se standardisent en ontologies. Ceci permet de capitaliser et de pérenniser la culture propre à tel ou tel domaine de référence dans les activités humaines, pouvant ainsi contextualiser les données brutes. Là encore, cette performance informatique ouvre aux données structurées une richesse sémantique des réseaux de machines, indépendamment des usages humains.

Les données publiques doivent donc être exposées dans un format minimal (PDF, CSV) interopérable, mais surtout éditées ou transposables potentiellement en RDF. De fait, selon nos collègues d'OWNI[7], data.gouv.fr ne publie qu'un très petit nombre de données réellement ouvertes, l'immense majorité étant encore dans des formats propriétaires (doc., XLS, …). Au-delà de la montée en standardisation, c'est surtout le décloisonnement qui découle de cette publication ouverte : transparence, car les langages de requêtes doivent permettre d'interroger toute la richesse des bases, mais surtout de réutiliser des croisements issus de différents domaines hétérogènes. Ainsi la répartition territoriale (géolocalisation des données) des formes de délinquance (identification des données pénales), des lieux de délits ou d'émeutes, des adresses des émeutiers suspectés ou jugés, permettent (en Grande Bretagne, par exemple[8]) de mettre en ligne des interconnexions de données et d'offrir aux utilisateurs des croisements ou contextualisations inédits.

Les possibilités techniques, avec le Web sémantique, de développement des conditions de réutilisation ne font que commencer. L'espace public devenant numérique va s'enrichir de façon exponentielle de masses de données produites « automatiquement » (traçage de flux, de transport, de communication, de transaction des objets et des agents, captage d'alertes écologiques, etc.), en même temps que vont s'amplifier les capacités « automatiques » de traitement, de catégorisation et d'accès multi-sources.

Notons enfin que le seul fichier en RDF de data.gouv.fr est celui de la BNF, issu de data.bnf.fr, le format pivot de la BNF[9].

[7] Article de Nicolas Patte : http://owni.fr/2011/12/10/la-france-entrouverte-transparence-open-gov-open-data-etalab/

[8] http://www.guardian.co.uk/news/datablog+uk/london-riots

[9] http://www.data.gouv.fr/donnees/view/Donn%C3%A9es-compl%C3%A8tes-du-contenu-de-la-BNF-30383137?xtmc=bnf&xtcr=1, la décompression de ce fichier de 25 MO semblant d'ailleurs défaillante…

Data.gouv.fr et l'amplification des possibilités éditoriales

L'ouverture des données publiques inaugure un nouveau rapport aux données. Avec la production numérique généralisée, celles-ci sont plus riches quantitativement, mais obligent qualitativement à plus de médiation éditoriale, pour vérifier et croiser les sources, mais surtout pour les réutiliser de façon créative et contextualisée. Le data-journalisme est au cœur de cette évolution, à condition qu'il sache ou puisse s'approprier l'évolution technique du Web sémantique précédemment citée[10]. *The Guardian*[11] est exemplaire de cette possible orientation d'offre de thématiques à la fois plus consistantes en données de référence et plus riches en « rédaction » éditoriale.

Une telle possibilité est bornée cependant par ce qu'on nomme « données ». Encore faut-il que celles-ci soit exploitables, c'est-à-dire suffisamment pertinentes. Prenons l'exemple d'un des fichiers de data.gouv.fr, assez intéressant en tant que tel, celui des mesures du « plan gouvernemental de Relance » de 2009. C'est en fait un agrégat à la Prévert, où les « véhicules de police » côtoient les budgets « défense, entretien de nécropoles » ou « le logement étudiant »... Les incohérences des politiques rejoignent les non-sens sémantiques ! On ne peut traiter que des données homogènes, ou du moins rendues telles par des problématiques précises de traitement. Le data-journalisme a devant lui un long et laborieux avenir !

Une des annonces de Séverin Naudet le 5 décembre est la future sortie par les partenaires techniques d'ETALAB en janvier 2012 d'un outil performant, *DataConnexions,* d'exploitation des données[12], au-delà des simples requêtes du moteur actuel. Rendez-vous en janvier donc, mais

---

Voir également : http://data.bnf.fr/semanticweb – Nous citons la notice : « La Bibliothèque nationale de France vous guide dans ses ressources patrimoniales, en publiant des fiches de référence inédites sur les auteurs et sur les œuvres. Dans data.bnf.fr, vous trouverez toutes les œuvres d'un auteur et toutes les éditions d'une œuvre, avec les liens vers les documents en ligne sur Gallica lorsqu'ils ont été numérisés. La sélection de ressources proposée est issue des catalogues documentaires, des inventaires d'archives et de manuscrits, ainsi que des collections numériques de la BnF. Elle valorise les grands classiques de la littérature, de l'histoire et du droit. C'est en fonction de vos usages et de vos préférences que l'ensemble sera progressivement élargi à un nombre croissant de références encyclopédiques. Le projet data.bnf.fr utilise les techniques du Web sémantique qui favorisent une navigation plus fluide et plus intuitive par les internautes entre les différents types de ressources. L'utilisation de formats structurés pensés pour le Web (RDF) facilite l'indexation, le partage et la réutilisation des données : la BnF met ces données librement à votre disposition à condition d'en mentionner la source. »

[10] Un des ateliers d'ETALAB était consacré à cette question, en octobre.
Cf. http://www.etalab.gouv.fr/pages/atelier-de-travail-du-13-octobre-2011-datajournalisme-5913723.html

[11] Ib. http://www.guardian.co.uk/technology/page/2009/jun/17/3 – *Facts are sacred…*

[12] Parmi ces partenaires, notons la présence des très performantes PME, EXALEAD, et surtout MONDECA dans les domaines du Web sémantique. LOGICA a semble-t-il joué un rôle majeur dans le consortium.

Article publié en ligne : 2012/01
http://www.sens-public.org/article.php3?id_article=883

pour l'heure, on est quelque peu surpris de constater la totale absence de données de l'Enseignement supérieur et de la recherche, dont les acteurs devraient jouer un rôle évident de médiation savante sur le traitement et l'éditorialisation des données !

L'absence de « transparence » n'est sans doute ici qu'un retard… Il est par exemple assez représentatif (mais peut-on tout reprocher au nouveau né ?) qu'une des seules opérations de « données ouvertes » en sciences humaines et sociales, la plateforme ISIDORE[13] ne soit pas référencée au titre de data.gouv.fr. Au-delà, c'est l'ensemble du domaine scientifique qui est absent de data.gouv.fr pour l'instant. Si ce manque d'implication, sinon cette cécité à l'un des domaines majeurs de la publication des données publiques que sont la recherche et l'université perdurait, on ne pourrait que s'inquiéter des performances possibles de data.gouv.fr.

L'Éducation nationale est au contraire extrêmement productrice de données, ce qui augure de possibilités d'études approfondies en matière de sociologie de l'éducation.

Ajoutons que la Conférence de Nelly Kroes du 12/12 d'actualiser la Directive de 2003 en « étendant considérablement le champ d'application de la directive afin d'y inclure, pour la première fois, les bibliothèques, les musées et les archives ; les règles de 2003 s'appliqueront aux données de telles institutions. » C'est là une annonce dont la transposition dans la loi française ne manquera pas d'être relevée, notamment pour ce qui est des relations actuelles entre les institutions culturelles et les éditeurs par exemple.

Enfin, un clin d'œil, on cherche vainement sur data.gouv.fr des données sur… la Mission ETALAB elle-même, son coût, ses marchés, etc.

Une « place de marché globale de la donnée »

« Place de marché globale de la donnée », l'expression est de Séverin Naudet. Sa conviction est en effet que l'inauguration de data.gouv.fr sera suivie, pour citer Vivek Kundra[14] « d'une explosion des communautés de développeurs, donnant naissance à une explosion des applications ». La refonte d'un modèle économique est au cœur du prosélytisme de l'ouverture des données. Voyons de plus près.

Au-delà de la transparence des accès, l'autre pilier de la Mission ETALAB est de favoriser la réutilisation des données. En langage bruxellois, cela veut dire élargissement sans entrave des

[13] www.recherche.isidore.fr, lancée en décembre 2010 par le TGE ADONIS a aujourd'hui environ 1 400 000 ressources, qui ne sont certes pas toutes des données publiques, et par ailleurs exclusivement approchées sur le versant « document », et très peu « jeux de données », mais dont les méta-données et données structurées sont converties et exploitées en RDF.

[14] Le Conseiller data.gov d'Obama.

accès gratuits sur les données « brutes » pour libérer une valorisation sur les produits et services (payants) dérivés des applications développées sur cette nouvelle base.

On l'a indiqué, la Commission européenne annonce le 12 décembre une stratégie renouvelée de l'ouverture des données publiques en Europe. « Turning government data into gold », les « raw data » d'aujourd'hui sont les gisements d'or noir d'hier, le pétrole ! L'étalon de la mesure est ici l'ouverture d'un grand marché de produits et de services à partir de cette libération des données :

> « D'après une étude récente[15], le marché des produits et services reposant sur les informations du secteur public représentait, en 2010, environ 32 milliards d'euros dans l'UE. La même étude indique que le fait d'ouvrir davantage le marché des informations du secteur public en y donnant un plus large accès procurerait à l'UE des avantages économiques globaux d'environ 40 milliards d'euros par an. »[16]

On reconnaît là le principe des externalités positives constamment invoqué au sujet de l'économie immatérielle. Les données publiques sont le fond d'investissement d'un nouveau marché de l'économie privée.

Nous aurons l'occasion de revenir sur ces problématiques, mais il est rien moins qu'avéré que, « naturellement », ces externalités positives soient au rendez-vous du numérique. Une économie immatérielle de ce type n'a de sens qu'au stade le plus global de son développement, dans un espace général lui même extrêmement polarisé (et non homogène comme le suppose l'approche de « marché » de la Commission ou d'ETALAB). A l'échelle mondiale de la production, des accès et de la réutilisation des données, le partage entre *soft power* et *hard power* qui structure l'économie numérique restructure les avantages comparatifs des acteurs. A cet égard, la publication de la nouvelle stratégie européenne en matière de libéralisation des données publiques devra être lue avec attention.

L'ouverture c'est d'abord la gratuité, et de fait, l'offre de data.gouv.fr se substitue à des offres de données identiques qui était il y a peu payantes, et donc d'accès restreint à un marché spécifique très ciblé. Libérée, cette gratuité doit être néanmoins cadrée par une licence ouverte de production, d'accès et de réutilisation des données publiques[17].

S. Naudet n'exclue pas d'ailleurs que les entreprises aussi aient avantage à ouvrir à l'avenir une partie de leur données. Le passage à la gratuité considéré comme une dynamique inhérente à la « destruction créatrice » de valeur à l'ère numérique.

---

[15] http://ec.europa.eu/information_society/policy/psi/docs/pdfs/report/final_version_study_psi.docx

[16] http://europa.eu/eucalendar/event/id/279090-the-european-commission-proposes-an-open-data-strategy/mode/standalone

[17] http://www.etalab.gouv.fr/pages/licence-ouverte-open-licence-5899923.html

Le temps de gestation technique de data.gouv.fr a également été un temps d'élaboration de cette licence libre relativement mouvementé, selon « Regards citoyens »[18].

Ce cadre juridique doit en retour protéger positivement l'émergence d'un marché de réutilisation des données par l'émergence d'applications nouvelles, surtout dans le croisement inédit d'informations telle que la géolocalisation de services.

Les applications et innovations attendues sont fortement sollicitées par ETALAB, et de fait, ce sera un volet intéressant de la suite de la Mission confiée à S. Naudet, son équipe et ses partenaires.

A court terme, le véritable enjeu économique de data.gouv.fr est de savoir si cette plateforme va servir de socle au développement du Web sémantique en France, et surtout des compétences et d'expertises françaises en matière de Web de données, de Linked data, dans le monde.

Mais n'est-ce pas une autre façon de poser la question de la polarisation de l'économie immatérielle potentiellement générée par l'ouverture des données publiques, et pas seulement supposer un effet mécanique du « marché européen » ?

Rien en l'occurrence n'est « économiquement écrit », en matière de possible et d'avenir. L'ouverture des données publiques, l'attente économique qui en est potentiellement faite, ouvrent surtout à une recomposition des sphères traditionnelles « publiques /privées », donc en appelle à une stratégie politique renouvelée.

Data.gouv.fr, medium d'« ouverture » politique ?

Nul ne sera assez naïf pour ne pas voir dans le forcing de cette sortie de data.gouv.fr au terme du quinquennat de N. Sarkozy une simple opération de campagne. Le contexte est politicien et la crise économique d'interrogation sur la destruction de valeur (elle, bien en adéquation avec la polarisation mondiale de l'économie immatérielle) est l'évidence quotidienne.

Plus sérieusement, de nombreuses questions sont encore dans le flou en ce qui concerne le périmètre de définition des données publiques ou de l'exploitation (publique ou non) des traitements des données : données de la recherche et stratégie de protection des avantages comparatifs à l'exemple des brevets, droits d'auteur, données culturelles, données personnelles, secret d'État… autant de points de vue où la « donnée publique » peut être substantiellement limitée, sinon obérée.

Nicolas Patte a raison de poser la question : les données sont-elles « données » de par leur production *top down*, de l'administration de l'État vers le citoyen-contribuable passif, ou *bottom-up*, en fonction des intérêts, des réactions et participation de celui-ci en tant qu'administré actif et

[18] http://www.regardscitoyens.org/opendata-etalab-la-guerre-francaise-des-licences-sacheve/

responsable ? L'innovation des possibles devra rejoindre cet aspect participatif ou collaboratif indispensable.

Owni, en juillet 2011, écrivait déjà :

> « L'État devra donc réfléchir à la possibilité de passer d'un modèle "à sens unique" (diffusion des données du secteur public vers la société civile) à un modèle d'écosystème où les données de l'État et des collectivités, ouvertes à la société civile, pourraient être enrichies en retour de façon collaborative ("crowdsourcing") ».

Aussi le discours de présentation du 5 décembre en rajoutait sur les promesses citoyennes et participatives.

Pour l'instant, rien du data.gov.fr n'infère d'une « ouverture de Gouvernance» au-delà des belles paroles invoquant la modernité (technique ?) de l'innovation.

Pour reprendre notre exemple du fichier du Plan de relance gouvernemental, les « données » considérées ne prendront vraiment leur sens que lorsque chaque administration concernée (par exemple les lignes 34 à 51 relevant de l'enseignement supérieur et de la recherche), et encore plus les acteurs concernés, produiront *bottom up* aussi leurs données de projet… Le croisement avec l'actuel *Top down* risquerait d'être éloquent !

Nicolas Patte, s'appuyant sur la non participation de la France à l'*Open Government Partnership* (OGP), a par ailleurs quelques raisons de s'interroger sur les recouvrements ou oppositions de compréhension des notions de *transparence* en langage anglo-saxon ou en français[19] !

Sans le vouloir, data.gouv.fr et la réussite à marche forcée de son lancement sont un puissant révélateur de la verticalité hiérarchique de l'État jacobin[20] ! Les neuf mois d'ETALAB, c'est d'abord l'efficacité quasi militaire de l'implication de tous les ministères par le Premier d'entre eux, puis de façon descendante des administrations concernées, puis les établissements publiques

[19] Selon Owni : « invitée par Barack Obama et Dilma Rousseff, la présidente du Brésil, à prendre siège autour de la table du projet *Open Government Partnership* (OGP), la France ne fait aujourd'hui pas partie de la cinquantaine de pays [en] s'étant engagés fermement à suivre les intentions vertueuses de cette initiative promue par l'Onu – dont le programme est pourtant alléchant : "engagement à la disponibilité accrue d'informations relatives aux activités gouvernementales", "engagement à promouvoir la participation civique", "engagement à faire appliquer par les administrations les normes les plus strictes d'intégrité professionnelle", ou encore "engagement à intensifier l'accès aux nouvelles technologies à des fins de transparence et de responsabilisation". Le reste est à l'avenant. Des pays européens comme la Grande-Bretagne, la Grèce, l'Espagne ou l'Italie, la Suède, la Norvège ou encore le Danemark ont franchi le pas vers l'avant que le couple franco-allemand aura décidé de ne pas faire.

[20] L'opposition de la verticalité des institutions et de l'horizontalité du réseau est aussi de Séverin Naudet.

Article publié en ligne : 2012/01
http://www.sens-public.org/article.php3?id_article=883

dépendants, etc. Bref, du *Top down* exactement à l'inverse des dynamiques du Web… On connaît les justifications non moins colbertistes : il faut bien que l'initiative de l'État supplée aux manques de la Société civile… Ce sont bien sûr les lendemains qui seront juges, lendemains qui se traduisent souvent en plates promesses électorales… qui n'engagent que les *bottom up* consentants…

Soyons juste, la Mission ETALAB le rappelle :

> « chaque producteur de données publiques a vocation à enrichir et à actualiser ses données de façon autonome, après le premier recensement de données réalisés pour l'ouverture de la plateforme "data.gouv.fr" ».

Il n'en reste pas moins que, si cette perspective participative reste asymptotique, data.gouv.fr peut n'être qu'un outil de contrôle social… Et dont il n'est pas dit que ce soit de l'auto-contrôle ou de l'autogestion !… mais au contraire (et au pire, connaissant la consistance de la technocratie française) de la surveillance de tous par tous… meilleur outil cybernétique possible d'un méta-contrôle social.

Pour le présent, la Mission ETALAB, menée en équipe projet performante par Séverin Naudet, a pleinement joué son rôle d'impulsion. De façon opportune, La Commission européenne, vient d'amplifier et de relayer l'initiative française, comme celles du Royaume Uni et du Danemark avant elle. Mme Nelly Kroes annonce que l'ensemble des données de la commission seront raidement ouverte dans un nouveau portail de données ; que des mesures d'interopérabilité seront promues entre les 27 pays, et qu'un fonds de 100 millions d'euros viendra soutenir la recherche dans la gestion des données pour la période 2011-2013.

Les mois qui viennent nous permettront de donner corps à tous les possibles évoqués ici. Techniquement, éditorialement, économiquement et surtout politiquement, les acteurs innovants, informaticiens du Web, journalistes, chercheurs, universitaires, fonctionnaires, collectivités locales, acteurs de la société civile, entrepreneurs ou simples citoyens auront-ils intérêt à amplifier la production et l'exploitation des données publiques pour leur ajouter toute la pertinence requise, pour en retirer toute l'innovation sémantique potentielle ? Comment la société civile (cette « donnée publique »…) s'appropriera-t-elle data.gouv.fr ? Question (qui rejoint les premières analyses de « Regards Citoyens ») et qui ne manquera pas de modifier profondément le projet lancé le 5 décembre.

Article publié en ligne : 2012/01
http://www.sens-public.org/article.php3?id_article=883