Prolégomènes à l'ingénierie des données ouvertes

Branislav Meszaros, Sitthida Samath, Sonia Guérin-Hamdi, Céline Faure

Prolégomènes à l'ingénierie des données ouvertes

Branislav Meszaros*, Sitthida Samath*,
Sonia Guérin-Hamdi* et Céline Faure*

*Institut des Sciences de l'Homme, 14 avenue Berthelot, 69363 Lyon cedex 07, France
branislav.meszaros@gmail.com, sitthida.samath@ish-lyon.cnrs.fr,
sonia.guerin-hamdi@ish-lyon.cnrs.fr, celine.faure@ish-lyon.cnrs.fr
https://www.ish-lyon.cnrs.fr/

Résumé. Les données ouvertes, ce nouvel objet de toutes les attentions, que certains n'hésitent pas à représenter comme l'« or noir » des temps modernes sont devenues très rapidement un des enjeux majeurs de l'humanité que nul ne peut prendre à la légère et encore moins ignorer. On s'efforce dorénavant de porter davantage attention à ce nouveau paradigme sociétal, culturel et économique pour mieux appréhender sa complexité et mesurer l'impact que cela peut avoir, voire a déjà, sur nos sociétés. Ainsi, ce travail propose de faire un état des lieux sur le mouvement open data et les données ouvertes. Il présente notamment les aspects clés du sujet (juridiques, techniques, économiques), afin que tous ceux qui veulent se lancer dans l'aventure de l'open data puissent mieux en évaluer la nature et les multiples facettes.

1 Introduction

Le binôme des notions *information et connaissance* conditionne, depuis très longtemps, d'une manière ou d'une autre, la vie des gens, en donnant un certain rythme à nos sociétés. Toutefois ce n'est que très récemment que l'on commence à s'intéresser de plus près également à l'objet qui les véhicule, l'objet à partir duquel on les extrait et qui constitue, en quelque sorte, leur colonne vertébrale, c'est-à-dire les données. Ce nouvel objet de toute attention que certains n'hésitent pas à représenter comme l'« or noir » des temps modernes est devenu très rapidement un des enjeux majeurs de l'humanité que nul ne peut prendre à la légère et encore moins ignorer. Ainsi, on s'efforce dorénavant de porter davantage attention à ce nouveau paradigme sociétal, culturel et économique pour mieux appréhender sa complexité et mesurer l'impact que cela peut avoir, voire a déjà, sur nos sociétés.

Les différentes facettes sous lesquelles la notion de données se présente à nous et les formes sous lesquelles elle se décline démontrent déjà à quel point son rayon d'action et d'influence, ainsi que les espérances et les craintes qu'elles véhiculent dépassent de très loin la banalité des faits sous lesquels elles ont été ensevelies jusqu'à nos jours. *Raw data*, *small data*, *big data*, *fast data*, *smart data*, *open data*, etc. sont devenues les nouvelles sources de la reconfiguration sociétale et cela à tous les niveaux. Chacune en tant que telle, ou en conjonction ou en interdépendance avec les autres, représente un nouveau terrain prolifique de l'activité humaine. De ce point de vue, l'*open data*, ou autrement dit les données ouvertes, constitue un phénomène

à part et cela non pas par sa différence mais plutôt par son aspect englobant où tous les autres types de données peuvent s'y retrouver. Tout simplement, les données, ouvertes ou non, sont tout sauf un objet banal. De ce fait, l'objectif de ces prolégomènes est de présenter quelques aspects importants, ou du moins les plus pertinents, concernant les données ouvertes, afin que tous ceux qui veulent se lancer dans l'aventure de l'*open data* puissent mieux évaluer la nature et les multiples facettes de cette problématique.

2 Historique et définitions

2.1 Genèse

La complexité de la notion de donnée, et tout particulièrement de donnée ouverte, et cela malgré son aspect tout à fait anodin, touche l'ensemble des niveaux sous lesquels ces données se présentent à nous. Cela concerne aussi bien son histoire, son statut juridique, les questions techniques ou le sens-même de la notion d'*open data*. Sans oublier que du point de vue économique, sociétal, institutionnel ou autre, les données ouvertes gardent toujours en elles une part de mystère et même de contradiction. Ainsi, afin de pouvoir mieux comprendre en quoi les données ouvertes consistent, il est alors nécessaire d'entreprendre une analyse de quelques-unes des facettes sous lesquelles elles se manifestent, en commençant par la genèse de cette tendance à l'ouverture.

2.1.1 France

L'histoire de l'*open data* a commencé bien avant que nous le laisse croire cette notion, tant en vogue à l'heure actuelle. Cela est lié au fait que la gestion de l'accès à l'information publique a été régie avant même que le monde du numérique n'ait fait son apparition. Tout a commencé, au moins en ce qui concerne la France, déjà à l'époque de la Révolution Française où plusieurs événements ont permis de mettre en place les premières briques de l'édifice de ce mouvement d'ouverture.

Cela débuta avec la Déclaration des Droits de l'Homme et du Citoyen de 1789 et son article 15 qui stipule clairement que « *La société a le droit de demander compte à tout agent public de son administration* » (Déclaration, 1789). C'était le premier appel à la transparence de l'action publique. Ensuite, afin de renforcer cette logique d'ouverture, une loi a été votée le 25 juin 1794, appelée la Loi du 7 messidor an II, qui portait sur l'organisation des archives établies auprès de la représentation nationale. Dans celle-ci, on peut lire (article XXXVII) que : « *Tout citoyen pourra demander dans tous les dépôts, aux jours et aux heures qui seront fixés, communication des pièces qu'ils renferment : elle leur sera donnée sans frais et sans déplacement, et avec les précautions convenables de surveillance* » (Loi, 1794). Tout cela a été réaffirmé, près de deux cents ans plus tard, par les deux lois de 1978 : la loi du 6 janvier 1978 sur la promulgation de la loi Informatique et libertés en créant en même temps la Commission Nationale de l'Informatique et des Libertés (CNIL) chargée de la protection des données personnelles et la loi du 17 juillet 1978 qui met en place la Commission d'Accès aux Documents Administratifs (CADA) créée avec pour objectif de faciliter et contrôler l'accès des particuliers aux documents administratifs. La majorité des administrations publiques se doivent dès lors de rendre leurs données accessibles, et ce pour un coût modéré ou nul.

Il faut dire que, malgré certaines critiques, sur la lenteur de la France par rapport à l'ouverture des données publiques, la France a joué le rôle d'un des précurseurs majeurs dans le monde de l'*open data* et cela bien avant l'heure. Ainsi, même si l'avancée n'est pas à chaque fois spectaculaire, elle est plutôt constante et systématique et cela même par rapport au monde numérique d'aujourd'hui car c'est déjà en 1997 que le gouvernement a décidé de la publication en ligne et gratuite des « *données publiques essentielles* » (Wikipédia, 2015). Quelques années plus tard, c'est-à-dire le 23 octobre 2000, c'est le portail de l'administration française service-public.fr qui a été mis en place par la Documentation Française (devenue Direction de l'Information Légale et Administrative DILA en 2010) et en 2002 c'est le portail vie-publique.fr qui est lancé. L'année 2005 a été marquée par la transposition dans le droit français de la directive européenne 2003/98/CE (Directive, 2003), révisée en 2013 (Directive, 2013), appelée communément Directive PSI (Public Sector Information), qui avait pour but d'encadrer et d'harmoniser la réutilisation des données publiques. Puis, en juin 2010, le premier portail « *open data* » voit le jour. C'est grâce au conseil municipal de Rennes qui a décidé de se lancer dans l'aventure de l'*open data* avec le portail data.rennes-metropole.fr [1]. Peu de temps après (c'est-à-dire le 27 janvier 2011), c'est opendata.paris.fr qui ouvre ses portes, ce qui déclenche une vague d'ouvertures de différents portails qui mettent à disposition des données ouvertes et cela aussi bien au niveau local que national. Un autre élément qui a marqué l'histoire de l'*open data* en France est la création de la mission Etalab le 21 février 2011 qui avait pour but de mettre en place des conditions nécessaires à la mise en place des données ouvertes. Cela s'est traduit par la création par Etalab, en octobre 2011, de la Licence ouverte, qui était, entre autres, destinée à régir les données du futur portail gouvernemental qui a été ouvert, à son tour, en décembre 2011. C'est également en 2011 que le principe de la gratuité du droit à réutiliser des documents et données publiques a été posé par le décret n° 2011-577. Enfin, pour mieux superviser les processus de l'ouverture des données publiques, l'État s'est doté, à travers le décret du 16 septembre 2014 (Décret, 2014), d'un nouvel instrument, à savoir du poste d'administrateur général des données qui est occupé à présent par Henri Verdier (PMAP, 2014). Celui-ci, placé sous l'autorité du premier ministre est chargé de superviser et d'améliorer l'utilisation des données par l'administration et ses nombreux opérateurs.

Voilà, en quelques lignes les moments importants de l'ouverture des données en France et même si la liste n'est pas exhaustive, elle relate quelques éléments clés de ce processus qui n'en est, malgré tout, toujours qu'à ses débuts.

2.1.2 Monde

Avant tout, il faut rappeler que la première apparition de la logique de données publiques « ouvertes » a eu lieu avant même que la Révolution Française ne se déclenche car c'est en Suède, qu'une loi datant de 1776 (Bouchoux, 2014) autorise tout citoyen à demander, auprès des autorités concernées, d'avoir accès, par exemple, aux notes de frais d'un Ministre.

1. Il faut dire que la capitale bretonne fait ses premiers pas vers les données ouvertes déjà en mars 2010 en publiant en ligne ses données de transport (à l'adresse `http://data.keolis-rennes.com/fr`). Toutefois, en raison de la licence utilisée pour ces données, on ne peut pas vraiment parler de « données ouvertes » au sens strict du terme car celle-ci limitait grandement l'usage des données libérées en imposant les clauses d'usage comme « non-commercial » et « sans-modifications ». Il s'agissait d'une licence de Creative commons du type CC BY-NC-ND. Pour plus de détails sur la licence : `http://creativecommons.org/licenses/by-nc-nd/2.0/fr/`. L'exemple est repris à l'adresse : `https://libertic.wordpress.com/2010/03/01/77/`.

Toutefois, cet exemple reste, semble-t-il, bien isolé, mis à part celui de la France car il a fallu attendre jusqu'au 4 juillet 1966 et la promulgation de la Freedom of Information Act du président des États-Unis Lyndon B. Johnson qui a signé cette loi sur la liberté d'informations, suite au mécontentement des citoyens sur la gestion de la guerre du Vietnam, en obligeant les agences fédérales à mettre leurs documents administratifs à disposition de chacun qui en fait la demande. Mais ce n'est qu'en 1995 que la notion même de l'*open data* fait sa première apparition quand le Conseil national de la recherche aux États-Unis, dans le cadre du travail sur « *l'échange complet et ouvert des données scientifiques* » (NAP, 1995), a émis la volonté de décloisonner les échanges dans la recherche. Il faut toutefois dire que cette idée n'était pas vraiment nouvelle car c'est déjà en 1942 que Robert King Merton (Merton, 1973) appelait les scientifiques à abandonner la logique basée sur les droits de propriété intellectuelle et à partager librement les connaissances comme une condition sine qua non d'une recherche vraiment efficace. Mais son appel n'a porté ses premiers fruits qu'en décembre 2001 au moment du lancement du projet de l'Open Access Initiative (BOAI, 2010), connu également sous l'acronyme BOAI (Budapest Open Access Initiative), destiné à devenir la première archive ouverte en imposant en même temps les formats d'échanges ouverts comme XML et cela malgré le fait qu'à l'époque on parlait davantage de *libre accès* ou encore de *libre diffusion*. Par la suite, les choses se sont (un peu) accélérées car dès 2002 Lawrence Lessig crée la licence Creative Commons. Puis, en 2004, fait son entrée en scène l'Open Knowledge Foundation qui s'est donnée pour but à la fois de fournir des logiciels adaptés afin de mieux assister la mise à disposition des données et d'élaborer le cadre juridique que ces données demandent. Ainsi, en 2007 la licence ODbL, destinée principalement aux bases de données a vu le jour. La même année, a également eu lieu, à Sebastopol en Californie, la conférence sur le gouvernement ouvert (y ont participé, entre autres, Lawrence Lessig et Tim O'Reilly), dans le cadre de laquelle ont été élaborés les 8 critères fondamentaux (OGD, 2010) qui définissent (en quelque sorte), les données ouvertes et qui seront présentés plus loin. Toujours en 2007, c'est également le tour de l'amendement Honest Leadership and Open Government Act (US Gov, 2007a), connu sous le nom de l'OPEN Government Act of 2007 (US Gov, 2007b), puis en 2009, le jour même de son investiture, en tant que nouveau président des États-Unis, Barack Obama lance le projet Open Government Initiative comme un pas de plus dans la direction de la transparence du gouvernement qui a abouti la même année à l'ouverture du premier site destiné essentiellement aux données ouvertes – le portail data.gov. Par ailleurs, également en 2009, mais quelques mois plus tard, est lancé par le gouvernement britannique le portail data.gov.uk et complété, un an plus tard, en septembre 2010, par la licence Open Government Licence (NAUK, 2010). Afin de couronner cette logique de l'ouverture, en lui donnant davantage de légitimité et un cadre plus pratique, les chefs d'États du G8 ont signé en juin 2013, pendant le sommet de Lough Erne en Irlande du Nord [2], une charte pour l'ouverture des données publiques (G8, 2013) qui ne laisse plus aucun doute en ce qui concerne la place que les données ouvertes ont su et doivent bâtir au sein de nos sociétés.

2.1.3 Union Européenne

Les instances européennes ont été depuis longtemps conscientes de l'importance de la maîtrise du circuit informationnel. Cette volonté d'un positionnement clair et ferme, de la part des

2. Pour plus de détail, voir `http://fr.wikipedia.org/w/index.php?title=Sommet_du_G8_2013&oldid=110722253`.

membres de la Commission, vis-à-vis de ce sujet trouve son écho dans le communiqué sur les « Lignes directrices pour améliorer la synergie entre secteur public et secteur privé sur le marché de l'information » de 1989 où on peut lire : « *L'information est de plus en plus considérée comme un moteur pour le développement industriel de la Communauté dans un marché mondial hautement concurrentiel. La mise en place d'un marché des services d'information [...] est un objectif important dans la stratégie d'ensemble de la Communauté* » (CE, 1989, page 3). Cette volonté part du constat que « *les administrations et organismes publics collectent de grandes quantités de données et d'informations dans le cadre de leurs activités courantes. Ces données pourraient être mises à la disposition du secteur privé pour l'élaboration et la commercialisation de services électroniques d'information* » (CE, 1989, page 5). Depuis ce moment-là, l'ensemble des concertations et des directives émises à ce jour ne font que refléter les prémisses de ce qui se trouve déjà dans ce document. Ainsi, dans le cadre de la mise en place de cette politique de l'encadrement, de l'harmonisation et de l'uniformisation des principes de l'ouverture des données publiques, a d'abord vu le jour la directive 2003/98/CE concernant la réutilisation des informations du secteur public (qui a été transposée en France par l'ordonnance n° 2005-650 du 6 juin 2005 (Ordonnance, 2005) et le décret n° 2005-1755 du 30 décembre 2005 (Décret, 2005), tous les deux relatifs à la liberté d'accès aux documents administratifs et à la réutilisation des informations publiques), puis, en 2008, la directive INSPIRE (Directive INSPIRE, 2015) (qui imposait des standards techniques et le principe d'interopérabilité des données) et la directive 2013/37/UE du Parlement européen et du Conseil du 26 juin 2013 (Directive, 2013) (qui a modifié la directive 2003/98/CE). Par ailleurs, on y trouve de nouveau cette affirmation que « *L'un des principaux objectifs de l'établissement d'un marché intérieur est de créer les conditions qui permettront de développer des services à l'échelle de l'Union* » (Etalab, 2013). Il ne faut pas non plus oublier que depuis décembre 2012 l'Union Européenne s'est également dotée d'un portail de données ouvertes.

2.2 L'Open data comme enjeu démocratique, sociétal et économique

L'*open data*, ce nouveau phénomène des temps modernes, est avant tout une autre manière d'appréhender non seulement l'action publique (car elle n'est pas la seule à être visée) mais également, voire surtout, l'entrée de nos sociétés dans une nouvelle ère régie par les données. L'omniprésence et l'omnipotence des données ont permis de matérialiser les volontés déjà anciennes de l'accès de tout un chacun aux informations publiques et même d'étendre au-delà, cette logique de l'ouverture, à un spectre très large des données disponibles. Cette ouverture des données qui est à la base d'une importante reconfiguration sociétale et économique, et qui n'a pas encore démontré tout son potentiel, pose déjà les jalons d'une certaine démocratie de transparence et de participation. Un vieux rêve qui est en train de se transformer en réalité. Dans ce sens, comme le dit le Conseil national du numérique, « *la mise à disposition des données publiques est un impératif démocratique, un puissant vecteur de modernisation de l'administration et un composant important de l'économie numérique* » (CNN, 2012). Ces petits pas, timides, vacillants et parfois même opportunistes, qui marquent le début de la libération des données, peuvent se transformer, au fil du temps, en un grand bond historique d'un changement majeur à l'échelle de la civilisation toute entière. Il ne reste qu'à saisir l'occasion qui se présente. Les raisons en sont nombreuses mais on peut d'ores et déjà en mentionner quelques-unes, les plus importantes.

2.2.1 Transparence, participation et bien commun

Avant tout c'est la logique de transparence qui apporte son soutien à l'ouverture comme une nécessité démocratique. La méfiance envers les politiques, les structures étatiques et leurs actions, baignant très souvent dans l'obscurité, ne date pas d'hier car « *l'histoire politique est émaillée de scandales liés à l'argent public, et la défiance envers les élites politiques et administratives s'est de longue date installée dans l'esprit des citoyens. À la manière des cyclistes « tous dopés », les politiques sont « tous pourris »* » (Berthelot, 2013). Ainsi, afin de retrouver la confiance envers les élites politiques, la politique tout court ou la démocratie-même, un réel effort de transparence est de mise. De nombreuses initiatives ont vu le jour et cela même avant que l'*open data* ne devienne une réalité. C'est le cas par exemple de l'association Regards citoyens [3] qui depuis 2009 milite pour la transparence en mettant à disposition différents outils de cette « transparence » comme les sites nosdeputes.fr ou encore nossenateurs.fr. On ne peut qu'espérer que ces sites, dont l'utilité n'est plus à démontrer, et qui pour le moment se basent en majorité sur les informations récupérées d'une manière fastidieuse un peu partout, pourront un jour profiter davantage de l'ouverture des données des institutions concernées. Ceci dit, il semble que certaines instances commencent à se prêter au jeu comme c'est le cas, par exemple, de l'Assemblée nationale qui dans sa réunion du 12 novembre 2014 « *a décidé de publier les données mises en ligne sur le site de l'Assemblée dans un format exploitable pour la constitution de statistiques* » (Assemblée Nationale, 2014) parmi lesquels on trouve : des documents parlementaires, les coordonnées des députés, les mandats, les résultats des scrutins publics, etc.

On peut dire que ce n'est qu'une goutte d'eau mais c'est déjà ça car le plus difficile est de commencer. Sortir de l'obscurité et partager l'information qui constitue un levier important d'un bon fonctionnement de la démocratie est à la fois une nécessité et un devoir. Dans cette optique on ne peut pas ne pas évoquer, à titre d'exemple, l'initiative du gouvernement britannique qui a mis en place le site *Where does my money go ?* [4] avec un but clair, celui de « *promouvoir la transparence et l'engagement des citoyens à travers l'analyse et la visualisation des informations sur les dépenses publiques au Royaume-Uni* » (OKF, 2009). En France, une mission comparable est allouée, par exemple, au site vie-publique.fr, même si sa lisibilité, et encore moins son ouverture, ne peut pas vraiment égaler le portail britannique, ou encore le site nosfinanceslocales.fr de l'association Regards citoyens. Toutes ces initiatives, publiques ou citoyennes reflètent d'une manière ou d'une autre à la fois le bouleversement de nos sociétés suite à l'entrée dans l'ère du numérique, une meilleure prise en compte de la complexité du monde et une volonté à s'armer davantage vis-à-vis des immenses défis qui se dressent devant l'humanité toute entière et cela quelle que soit l'échelle sociétale.

De ce fait, la nécessité de la transparence dans l'action publique (et pas uniquement) n'est plus à démontrer et cela d'autant plus qu'il s'agit également d'un premier pas vers une meilleure compréhension du fonctionnement des instances démocratiques et, ce qui va avec, un réel gain au niveau de la confiance dont ces instances ont cruellement besoin de la part des citoyens en les incitant en même temps à davantage de participation de leur part. Ainsi, la transparence amène avec elle un renforcement du sentiment participatif des citoyens qui se voit davantage consolidé avec le mouvement de l'*open data*. C'est grâce à l'ouverture des données que de nombreuses initiatives, comme par exemple celles de Regards citoyens, ont pu voir le

3. `http://www.regardscitoyens.org`.
4. `http://wheredoesmymoneygo.org/`.

jour en perfectionnant davantage la « machine » étatique et cela aussi bien au niveau national que local.

Dans ce sens, on ne se situe plus uniquement au niveau des producteurs et distributeurs de données ouvertes mais également de ceux qui se chargent de leur réutilisation. Tout un chacun peut accéder aux données libérées et en développer des usages et des services nouveaux. La donnée et, ce qui va avec, l'information, deviennent un bien commun des temps modernes. On n'est plus uniquement dans une exclusivité et rivalité mais au contraire dans une ouverture à la participation, partage et collaboration qui redéfinissent nos appréhensions des paradigmes économiques dominants pour en entrevoir de nouvelles possibilités de gestion et de gouvernance [5].

2.2.2 Potentiel économique

Les données ouvertes ne riment pas uniquement avec les aspects démocratiques et sociétaux mais également avec celui de l'économie. La « valeur » des données ouvertes semble constituer une richesse et cela à la fois via les économies qui peuvent être réalisées grâce à la transparence de la gestion de l'action publique ou privée que via la création de divers services pouvant se traduire par de nouvelles initiatives entrepreneuriales. Par exemple, selon le rapport fait pour le Ministère de l'industrie, de l'énergie et de l'économie numérique « *chaque année dans l'Union Européenne, la réutilisation de données issues du secteur public générerait un chiffre d'affaires de plus de 27 milliards d'euros* » (Lacombe et al., 2011). Selon Neelie Kroes, ancienne commissaire européenne chargée de la société numérique, « *le marché de l'information publique représente 30 milliards d'euros par an. Mais en rendant le marché plus ouvert, les données plus accessibles, les retombées économiques globales à attendre du marché de l'information publique sont estimées à 40 milliards d'euros par an* » (Nicolas, 2012). Dans cette surenchère, on peut aller encore plus loin car « *en additionnant les bénéfices indirects de l'ouverture des données publiques, le marché pourrait atteindre un volume de 140 milliards d'euros par an* » (Nicolas, 2012). Mais certaines autres estimations dépassent de loin ces chiffres déjà impressionnants car le rapport de McKinsey de 2013 évalue l'impact de l'*open data* sur l'économie mondiale à plus de 3000 milliards de dollars (McKinsey Global Institute et al., 2013). Tout simplement, des estimations vertigineuses pour le « simple » fait de libérer des données.

Il ne faut toutefois pas oublier que « *la donnée ouverte acquiert une valeur seulement à travers sa (ré)utilisation et l'interprétation qu'en fait son utilisateur, puisque, par définition, son exclusivité et sa singularité ne peuvent être monnayées* » (Salaün, 2014). Dans ce sens c'est justement à travers leur utilisation que la valeur est créée ou peut être chiffrée et cela, entre autres, via le montant des économies réalisées ou sur les estimations des valeurs marchandes des applications et des services développés. Ainsi, par exemple, dans le cadre d'une *dev'party* de 2008 [6], qui a coûté 50000 dollars à la ville de Washington, les 47 applications créées en 30 jours ont été estimées à 2 millions de dollars. Ou encore la Catalogne qui, pour un coût de 1,5 million d'euros, a mis en place un dispositif destiné à des données ouvertes en générant au

5. Ceci dit, il faut également admettre que les notions telles que la transparence, la participation et le bien commun ne sont pas si univoques que ça et qu'il faut être extrêmement vigilant sur le contenu qu'ils sont censés véhiculer et la réalité qu'ils prétendent refléter. Pour s'en convaincre il suffit de lire l'ouvrage « L'Etat en mode start-up : le nouvel âge de l'action publique » de Yann Algan et Thomas Cazenave (Algan et al., 2016).

6. Pour plus de détail, voir `http://istrategylabs.com/work/apps-for-democracy-contest/`.

passage 2,6 millions d'euros d'économies par an [7]. Comme on peut alors le voir, les données ouvertes et l'écosystème qu'elles sont censées générer reflètent avant tout les possibilités et les opportunités qui se cachent derrière elles. Il ne reste qu'à matérialiser cet immense potentiel.

2.3 La quête de définition

2.3.1 Donnée ouverte vs donnée publique

Il faut dire qu'une grande confusion règne en ce qui concerne la notion de la donnée ouverte qui est très souvent confondue avec la donnée publique. Comme on l'a dit précédemment il s'agit en réalité avant tout d'un certain nombre de principes (exposés plus loin) qui doivent être réunis pour qu'on puisse parler d'une donnée en tant que donnée ouverte. À cet égard, de nombreuses données publiques ne peuvent pas être prises d'une manière automatique pour des données ouvertes. Par ailleurs, c'est un des challenges de la mission d'Etalab de rendre le maximum de données publiques ouvertes en gardant en tête les quelques principes qui doivent être pris en compte à cet effet. De plus, la notion des données ouvertes n'est pas réservée strictement aux données publiques mais c'est l'ensemble des secteurs et des acteurs qui produisent ou travaillent avec les données qui sont concernés. De nombreuses sociétés publiques et privées se prêtent également au jeu de l'ouverture pour des raisons qui peuvent aller d'une simple volonté de transparence, comme par exemple le groupe ENEL [8] en Italie, jusqu'à intégrer cette logique de l'ouverture, en tant que partie intégrante du business plan, comme c'est le cas pour Keolis [9]. Comme on peut alors le voir, les données publiques n'ont pas le monopole de l'ouverture même si ce sont elles qui représentent pour le moment la majorité des données ouvertes mises à disposition.

D'autre part, cette confusion vient du fait que les premières à être visées étaient justement les données publiques. Comme on l'a vu, au début de ce document, l'ouverture était, et demeure, connotée principalement avec le principe de la transparence. Les États et les institutions publiques, en voulant gagner la confiance des citoyens en faisant preuve de transparence de l'action publique, comme signe indirect que tout un chacun peut participer, en bon citoyen, à la construction et la gestion des instances de l'État, ont été les premiers à parler de l'ouverture. De ce fait, il est normal de voir que les données ouvertes se confondent très souvent avec les données publiques ou informations publiques ou encore documents administratifs. Bref, le poids de l'histoire laisse des traces.

2.3.2 Donnée ouverte, quèsaco ?

Il semble qu'à la base, la donnée ouverte, en tant que concept juridique, n'existe pas vraiment. Certains préfèrent même parler uniquement du mouvement d'open data afin de mettre l'accent avant tout sur l'action qui mène vers cette ouverture. Par ailleurs, il n'est déjà pas facile de définir une donnée en tant que telle, et cela surtout si on ne parle pas uniquement des cas comme des mesures, des relevés statistiques, des données d'inventaires ou encore des coordonnées géographiques et environnementales. En SHS, par exemple, les « données » sont

7. Voir `http://fr.slideshare.net/libertic/lopendata-5128072`.
8. `http://data.enel.com/dataresults?language=en`.
9. `http://data.keolis-rennes.com/fr/les-donnees/donnees-et-api.html`.

(trop) riches pour en établir sans difficulté une définition universelle. Par exemple, on peut regarder cette liste qui constitue en quelque sorte une base de ce qui peut être pris comme une donnée ouverte le cas échéant :

— « *des données de description du territoire (cartes, cadastre...) ;*
— *des fonds documentaires (études, réglementation, statistiques...) ;*
— *les données de la décision publique (projets, enquêtes, délibérations, subventions, budgets...) ;*
— *le fonctionnement des réseaux urbains (eau, énergie, transports, logistique, télécoms...) ;*
— *la localisation et les horaires d'ouverture des services et des commerces ;*
— *l'occupation des ressources et des capacités (voirie, bâtiments, espaces, parkings...) ;*
— *des mesures (environnement, trafic...) ;*
— *des événements (culture, sports...) ;*
— *des informations touristiques, culturelles, des données d'archives ;*
— *les flux urbains (circulation...) ;*
— *des données de surveillance ;*
— *des données électorales*. » (Rennes DGIC, 2015)

Et évidemment bien d'autres qui ne sont pas mentionnées dans cette liste. Bref, les données ouvertes, c'est l'ensemble des données dont on décide l'ouverture dans le sens que « *Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness)* » ou autrement dit que « *Open data and content can be freely used, modified, and shared by anyone for any purpose* » (OKF, 2015).

2.3.3 8 principes fondamentaux

Afin de donner du corps à cette « *définition* », quelques principes (OGD, 2010) ont été posés en décembre 2007 [10] à Sebastopol, au nord de San Francisco, par un groupe de travail de 30 personnes, parmi lesquelles Tim O'Reilly – auquel on doit la définition du web 2.0 – et Lawrence Lessig – professeur à l'Université de Stanford et promoteur des licences Creative Commons. Ce groupe de travail a posé ainsi les bases de l'*open data* dont les 8 principes fondamentaux sont :

1. Les données sont complètes. Toutes les données publiques doivent être rendues disponibles sauf les données pouvant porter par exemple atteinte à la vie privée des citoyens ou à la sécurité.
2. Les données sont primaires. Les données doivent être brutes, telles qu'elles ont été collectées à la source, non agrégées, non modifiées.
3. Les données sont tenues à jour. Elles doivent être rendues disponible aussi vite que possible afin de préserver leur valeur.
4. Les données sont accessibles. Les données sont disponibles au plus large spectre d'utilisateurs pour l'usage le plus large

10. Ces 8 principes ont été repris et enrichis par Sunlight Foundation en les amenant à 10 sous la forme de « Ten Principles for Opening Up Government Information » (`http://sunlightfoundation.com/policy/documents/ten-open-data-principles`). Les deux éléments qui n'ont pas étés mentionnés ici mais qui font, en même temps, partie des 7 principes auxiliaires, sont la permanence et la gratuité des données. Il s'agit ici, en parlant de permanence et de gratuité, d'une difficulté à part nécessitant une approche spécifique dont l'évidence est difficile à démontrer.

5. Les données doivent permettre un traitement automatisé. Elles doivent être structurées et documentées.
6. L'accès aux données est non discriminatoire. Elles sont disponibles à tout le monde de façon anonyme ne nécessitant pas d'enregistrement.
7. Les données sont disponibles dans un format non propriétaire. Elles doivent être rendues disponibles au moins dans un format sur lequel aucune entité ne détient le monopole (ex : non PDF, non Excel).
8. Les données sont libres de droits. Elles ne doivent pas être l'objet de droits d'auteurs, marques déposées, brevets, etc.

2.3.4 7 principes auxiliaires

Ces 8 principes fondamentaux peuvent être également complétés par 7 autres qui précisent davantage les conditions d'une ouverture efficace et durable (OGD, 2010) :

1. En ligne et gratuites – l'information n'est pas significative pour un public si elle n'est pas disponible sur l'Internet, sans frais, ou du moins pas plus que le coût marginal de reproduction. Elle devrait également être facilement trouvable ;
2. Permanentes – les données devraient être disponibles à un emplacement stable et dans un format de données stables aussi longtemps que possible ;
3. Vérifiés – le contenu publié doit être signé numériquement ou inclure une attestation de publication / date de création, l'authenticité et l'intégrité ;
4. Présomption d'ouverture – la présomption d'ouverture repose sur des lois comme la CADA en France ou Freedom of Information Act aux Etats-Unis, y compris les procédures de gestion des documents et des outils tels que des catalogues de données ;
5. Documentées – les données sont bien documentées à la fois sur les données en elles-mêmes que sur leurs formats et les métadonnées qui les accompagnent ;
6. Sures – ceux qui publient les données en ligne doivent toujours chercher à éviter les formats exécutables ;
7. Prend en compte le public – Le public est sans doute le mieux placé pour déterminer ce qui est pour lui nécessaire et suffisant.

2.3.5 72 bonnes pratiques

Dans le cadre d'une politique de bonne lisibilité et compréhension de ces quelques principes fondamentaux et auxiliaires, une liste de 72 « bonnes pratiques » (Temesis, 2015) a également été mise en place pour mener à bien les projets de l'ouverture des données. Ces considérations supplémentaires précisent les conditions nécessaires et suffisantes pour chaque étape de la libération des données et peuvent s'avérer fort utiles.

3 Les données ouvertes à la loupe

3.1 Aspect juridique

3.1.1 Encadrement juridique

Comme les données peuvent avoir différentes provenances, c'est-à-dire être issues du secteur public ou privé, les cadres législatifs qui encadrent ces deux sphères ne sont pas non plus les mêmes, et cela malgré le fait que les chevauchements et les interdépendances des lois restent en vigueur et cela particulièrement en France. Ainsi, là où les données publiques sont principalement régies par la loi CADA de 1978 (Loi, 1978a,b) et les lois connexes, les données issues du secteur privé dépendent du code de la propriété intellectuelle (CPI, 2015) et de ses diverses lois connexes.

En ce qui concerne les données dites publiques, comme le prévoit la loi nn^{o} 78-753 du 17 juillet 1978, dans le cadre de la liberté d'accès aux documents administratifs, toute personne a le droit à l'information tant que les conditions stipulées par cette loi, et les lois connexes, sont réunies. La loi prévoit non seulement la nature des documents et des informations qui peuvent être mises à disposition mais également les conditions dans lesquelles ces données peuvent être réutilisées. Bien évidemment, tous les documents ne sont pas accessibles, ou du moins d'une manière directe, comme par exemple ceux qui concernent la sécurité de l'État ou les différents secrets protégés par la loi. Ainsi, ne sont pas communicables les données qui peuvent porter « *atteinte au secret de la vie privée et des dossiers personnels, au secret médical et au secret en matière commerciale et industrielle* » (Loi, 1978b, article 6) sans oublier les données relevant de la sécurité nationale et les données sur lesquelles des tiers détiennent des droits de propriété intellectuelle. La loi CADA ne s'applique, bien évidemment, qu'aux données publiques, les données du secteur privé sont régies par le code de la propriété intellectuelle.

3.1.2 Opposition des cadres juridiques

Afin de pouvoir analyser la situation concernant l'utilisation des diverses licences mises à disposition dans le cadre de l'ouverture des données (non seulement publiques) il faut, au préalable, lever le voile sur quelques lois fondamentales qui régissent cet écosystème. En même temps, il faut également mentionner que l'ouverture des données ne dépend pas d'un seul cadre juridique.

À l'heure de la globalisation des échanges et des coopérations transnationaux, la question de la clarté sur qui encadre quoi et dans quelles conditions est plus que d'actualité. Cela est surtout vrai si à l'échelle planétaire, au moins du point de vue qui nous intéresse, ce sont deux conceptions majeures qui s'affrontent : celle basée sur le copyright et celle basée sur le droit d'auteur tel que cela est régi en France par exemple. C'est une question fondamentale qui malgré les apparences provoque de nombreuses répercussions et cela même à l'échelle nationale. En simplifiant, on peut dire que « *cette opposition des modèles n'est, elle-même, que l'expression de philosophies opposées. Le droit d'auteur français est une propriété intellectuelle engendrée par l'acte créatif. Le copyright américain est un monopole légal accordé à un investisseur afin qu'il prospère à l'abri de la concurrence. Le droit d'auteur est donc nécessairement reconnu au créateur, alors que le copyright est octroyé à l'investisseur : normalement le producteur, parfois le créateur, quand il s'est autoproduit* » (Gaudrat, 2006). Comme on va le voir plus loin, cette distinction est très importante, surtout concernant la question des licences.

Droits d'auteurs Une des spécificités du système juridique français (pas uniquement, mais celle-ci est particulièrement prononcée en France) est que le droit d'auteur se décompose en deux volets de base – les droits moraux et les droits patrimoniaux. Là où les droits patrimoniaux peuvent être cédés dans le respect des conditions prévues à cet effet, tels que la loi sur les droits d'auteur le prévoit, les droits moraux ne sont pas séparables de la personne à laquelle ils s'appliquent. De ce fait, la transposition directe des données qui dépendent du code sur la propriété intellectuelle vers une licence, par exemple, du type CC-0 [11], semble plutôt difficile, pour ne pas dire irréalisable, au moins dans le cadre de la législation française. Bien évidemment, ce genre de spécificité a été pris en compte au moment de l'élaboration de la licence CC-0, en y ajoutant la clause « *Dans la mesure du possible et sans enfreindre la loi en vigueur* » (Creative Commons, 2015) concernant la possibilité de « *céder, abandonner, et renoncer ouvertement, pleinement, définitivement, irrévocablement et sans conditions à tous ses Droits d'Auteur et Droits Voisins* » (Creative Commons, 2015). Toutefois, à l'égard du droit d'auteur, on ne peut réellement renoncer à ces droits que dans la limite stipulée par ceux-ci, c'est-à-dire que vu que le droit moral « *est perpétuel, inaliénable et imprescriptible* » il peut être transmissible uniquement « *à cause de mort* » et cela uniquement « *aux héritiers de l'auteur* » (CPI, 2015, article L121-1). À première vue, il semble qu'il s'agit ici d'un vide juridique qui ne demande qu'à être comblé et éclairci. La situation est, de ce fait, très intéressante car cela pose de nombreuses questions sur la possibilité de créer un environnement effectivement ouvert et non contraignant de quelque manière que ce soit en ce qui concerne l'aspect juridique des possibilités d'une mise à disposition définitive et irrévocable des données (ouvertes) et des structures qui les accompagnent dans le cadre d'une renonciation absolue de ces droits sur ce qui est cédé. Par exemple, afin de mieux illustrer les propos, lisons le fragment de la licence SNCF : « *Le Réutilisateur s'engage à respecter, le cas échéant, les droits moraux que l'auteur conserve sur la Base de données et/ou sur le Contenu, conformément aux dispositions du Code de la Propriété intellectuelle. Les droits moraux comprennent le droit d'être identifié en qualité d'auteur de la Base de données ou du Contenu ou de s'opposer à tout traitement susceptible de porter atteinte à son honneur et à sa réputation, ainsi que tout autre traitement dérogatoire* » (SNCF, 2015). Cela concerne aussi bien des logiciels (article L112-2 du code de la propriété intellectuelle) que des bases de données (article L112-3 du code de la propriété intellectuelle). Par ailleurs, la portée du droit d'auteur est vraiment très large et comme certains le disent, peut-être même trop large (voir à titre d'exemple Tricoire (2006), et aussi, Gaudrat et Massé (2000)), et cela surtout à la fois par rapport à la nature même de la logique pour laquelle cette loi a été créée, qu'à la réalité actuelle dans laquelle elle tente de s'inscrire.

Droit des bases de données La situation autour des bases de données est également très intéressante car elle reflète en quelque sorte la complexité même de la réalité à la fois technique, juridique et historique. À la base, au moins d'une manière générale, les bases de données sont constituées de faits représentés, par exemple, par des relevés, mesures, dates, etc., qu'on peut appeler, le cas échéant, les « *données brutes* ». Ces « *données brutes* », comme elles ne représentent pas toujours des « créations » aux yeux du législateur, ne tombent pas directement sous le régime du droit d'auteur. De ce fait, afin de protéger les investissements réalisés pour obtenir ces données, il fallait trouver un moyen législatif pour encadrer l'accès à ces données par des tiers.

11. `https://creativecommons.org/publicdomain/zero/1.0/legalcode.fr`.

Aspect créatif – originalité Comme on ne pouvait pas accorder l'aspect « créatif et original » aux données, on l'a alors accordé aux structures (c'est à dire aux bases) qui les contiennent. Ainsi, aux termes de l'article L112-3 du code de la propriété intellectuelle, « *les auteurs de traductions, d'adaptations, transformations ou arrangements des œuvres de l'esprit jouissent de la protection instituée par le présent code sans préjudice des droits de l'auteur de l'œuvre originale. Il en est de même des auteurs d'anthologies ou de recueils d'œuvres ou de données diverses, tels que les bases de données, qui, par le choix ou la disposition des matières, constituent des créations intellectuelles* » (CPI, 2015, article L112-3). De ce fait, les bases de données sont protégées en vertu de la loi sur les droits d'auteur car elles-mêmes deviennent les fruits d'un acte de création.

Sui generis – retour sur investissement En parallèle, là où cette « originalité et aspect créatif » de la base de donnée ne pouvait pas être démontrée, on a pris en compte les investissements réalisés pour constituer ces bases de données. De ce fait, aux termes de l'article L. 341-1 du code de la propriété intellectuelle « *le producteur d'une base de données, entendu comme la personne qui prend l'initiative et le risque des investissements correspondants, bénéficie d'une protection du contenu de la base lorsque la constitution, la vérification ou la présentation de celui-ci atteste d'un investissement financier, matériel ou humain substantiel* » (Bouchoux, 2014). On remarque qu'on s'approche davantage de la logique du copyright.

Dans le cadre de ce qui est stipulé par la loi, le producteur de la base de données peut aux termes de l'article L342-1 interdire :

— « *L'extraction, par transfert permanent ou temporaire de la totalité ou d'une partie qualitativement ou quantitativement substantielle du contenu d'une base de données sur un autre support, par tout moyen et sous toute forme que ce soit* » ;
— « *la réutilisation, par la mise à la disposition du public de la totalité ou d'une partie qualitativement ou quantitativement substantielle du contenu de la base, quelle qu'en soit la forme* » .

Cet interdit peut également s'appliquer si les conditions des « *opérations [d'utilisation] excèdent manifestement les conditions d'utilisation normale de la base de données* », comme le stipule l'article L342-2. Tout cela, bien sûr, à condition que ces droits n'aient pas été cédés ou n'aient pas fait l'objet d'une licence.

Voilà le cadre d'accès aux bases de données et leur utilisation tels que le prévoit la loi dans le cadre du code sur la propriété intellectuelle. Toutes ces précisions semblent être nécessaires afin de mieux comprendre le cadre dans lequel doivent s'inscrire les licences qui sont censées encadrer le monde de l'*open data*.

3.2 Les licences

À l'heure actuelle, il existe dans le monde de nombreuses licences sous lesquelles on peut distribuer les données ouvertes [12]. Comme on l'a déjà évoqué, dans le système régi par le principe de copyright, l'application d'une quelconque licence est irrévocable, pleine et absolue. On a déjà vu que dans le système français, cela n'est pas vraiment possible (dans le cadre des lois existantes), ce qui est surtout dû à la nature du droit d'auteur. Il fallait alors, dans

12. Pour plus de détail, voir par exemple : `http://fr.jurispedia.org/index.php/Licence_libre_\%28fr\%29`.

ces conditions, voir de quelle manière de telles licences peuvent être établies en France. Par ailleurs, ce va et vient, qui accompagne le processus de la mise en place d'une licence ou d'une transposition dans la législation française d'une licence existante ailleurs, peut être démontré, par exemple, sur le cas de Rennes. La collectivité Rennes Métropole et la Ville de Rennes au moment du lancement de leur site internet dédié aux données ouvertes a d'abord opté pour la licence CGR (Conditions générales de réutilisation) de l'Agence du Patrimoine Immatériel de l'État (APIE) qui était à la base chargée de créer une licence à cet effet. Étant donné qu'il s'agissait d'une simple retranscription des conditions déjà prévues par la loi CADA, et que celle-ci s'avérait insuffisante pour les besoins auxquels elle était destinée, la collectivité Rennes Métropole et la Ville de Rennes ont alors décidé d'élaborer leur propre licence dit « *Licence de Rennes Métropole en accès libre* ». Celle-ci, à son tour, a été remplacée par la licence « *Open Data Base License* » ou autrement « *ODbL* » de l'Open Knowledge Foundation. Cela a d'ailleurs été le cas d'autres participants à la course à l'ouverture des données.

Une licence ? Pour quoi faire ? À l'égard de la loi CADA il n'est pas vraiment nécessaire d'élaborer une licence à part, vu que la mise à disposition et la réutilisation des données publiques est régie directement par cette loi, comme cela est stipulé dans l'article 10 : « *Les informations figurant dans des documents élaborés ou détenus par les administrations mentionnées à l'article 1er, quel que soit le support, peuvent être utilisées par toute personne qui le souhaite à d'autres fins que celles de la mission de service public pour les besoins de laquelle les documents ont été élaborés ou sont détenus* » (Loi, 1978b). Cela toutefois à condition que les données « *ne soient pas altérées, que leur sens ne soit pas dénaturé et que leurs sources et la date de leur dernière mise à jour soient mentionnées* » (Loi, 1978b, article 12). La nécessité de faire appel à une licence relève des organismes qui prévoient, par exemple, une redevance ou ceux dont les données sont gérées selon un régime spécial (EPIC [13]) comme c'est le cas par exemple de la RATP ou de la SNCF qui peuvent eux-mêmes décider des conditions de la libération des données.

Comme on l'a vu, dans le cadre de la loi CADA, les trois conditions imposées sont :

— l'indication de la source des données ou autrement dit de la paternité ;
— l'indication de leur date de mise à jour ;
— le respect de l'intégrité des données (ne pas altérer et dénaturer les données).

Ces trois conditions, et en particulier la dernière, représentent toutefois quelques difficultés. Parmi elles, les questions qui concernent le fait qu'il ne faut pas altérer et dénaturer les données. Par exemple, une « traduction » d'un format à l'autre ou l'enrichissement ou l'agrégation de ces données constitue une entrave au règlement ou non ? C'est d'ailleurs pour ces raisons-là que les licences semblent être une nécessité afin que ce genre d'incertitudes soit levé [14].

13. Établissement public à caractère industriel et commercial.

14. Il faut dire que certains quiproquos sont toujours présents et que la méconnaissance du fonctionnement des licences peut aboutir parfois à des situations très étranges. Par exemple, sur le site `https://opendata.hauts-de-seine.net/la-licence` des Hauts de Seine, on précise que les données sont mises à disposition avec la licence ouverte d'Etalab et en même temps, on stipule que « *conformément à l'article 12 de la loi n° 78-753 du 17 juillet 1978, les données ne doivent pas être altérées et leur sens ne doit pas être dénaturé* », ce qui va à l'encontre de ce qui est dit dans la licence ouverte qui permet, quant à elle, de réaliser les modifications sur les données mises à disposition (page web consultée le 5 mai 2015).

Les licences en usage Le monde des licences est constitué de paysages riches et en constante évolution. Cette réalité est liée au fait que de nombreuses tendances, volontés et besoins s'entrecroisent et il faut dire que tout le monde est, en fin de compte, concerné. Aussi bien le secteur public que privé. Ainsi, afin de pouvoir mettre les données à disposition de chacun, plusieurs licences sont d'ores et déjà prêtes à être utilisées, parmi lesquelles on peut mentionner les licences de *Creative Commons* [15], de l'*Open Knowledge Foundation* [16] ou encore celle d'*Etalab* [17]. Toutefois, dans certaines conditions, certains acteurs peuvent se mettre à élaborer leur propre licence si les licences existantes ne les satisfaisaient pas ou tout simplement s'ils préféraient garder une maîtrise plus large sur la libération de leurs données comme cela était le cas à une certaine époque, par exemple, de la Ville de Rennes avec sa « *Licence de Rennes Métropole en accès libre* ». D'autres, comme par exemple la Direction de l'Information Légale et Administrative (DILA) qui à travers sa licence « *Information publique librement réutilisable* » [18] (IP) a préféré mettre en place une licence spécifique capable de refléter davantage les différentes contraintes que celle-ci voulait imposer aux utilisateurs finaux. La Licence IP est de ce fait très contraignante car elle est non seulement décernée à titre personnel mais en plus pour une durée maximale d'un an (mais renouvelable).

Il y a également ceux qui sont obligés par la loi d'élaborer des conditions spécifiques et adaptées d'accès et d'usage pour leurs données si elles sont soumises, par exemple, à une redevance ou une autre forme de payement.

D'autres organismes, comme la RATP ou encore la SNCF, ont, en vertu de leur statut d'EPIC, le droit de décider librement de quelle manière et sous quelles conditions ils mettent à disposition leurs données. Par exemple la RATP met la plupart de ses données sous la licence ODbL ou encore sous la Licence ouverte d'Etalab et sa propre licence [19] est appliquée uniquement à certains éléments, qui se trouvent soumis au droit d'auteur ou à d'autres lois limitant leur utilisation et surtout la libération, comme par exemple les divers plans (où, semble-t-il, c'est surtout la charte graphique qui est protégée de cette manière). La SNCF, pour sa part, a opté pour une licence maison faite sur mesure pour prendre également en compte certains côtés techniques qu'une telle libération peut parfois imposer. De ce fait, cette licence est limitative, par exemple, dans le sens que « *la mise à disposition des données et contenus de SNCF en vertu de la Licence est limitée à un volume d'Extraction de 20 (vingt) requêtes par minute et par terminal connecté* » (SNCF, 2015). Comme on peut le voir, les raisons de choix ou de création d'une licence à part sont nombreuses et pas toujours évidentes à percevoir au premier coup d'œil.

Il semble, toutefois, qu'à l'heure actuelle, malgré le fait que le monde de l'*open data* n'en est qu'à ses débuts, quelques licences tentent de gagner une position plutôt dominante. Cela concerne principalement la Licence ouverte d'Etalab et la licence ODbL (Open Data Base Licence) de l'Open Knowledge Foundation. Ces deux licences, malgré certaines ressemblances, représentent, par contre, deux approches radicalement différentes. La philosophie qui se cache derrière chacune d'elles reflète un certain positionnement vis-à-vis de la vision du monde du libre en général et vis-à-vis des données ouvertes en particulier.

15. http://creativecommons.fr/licences/les-6-licences/.
16. http://opendatacommons.org/licenses/.
17. https://www.etalab.gouv.fr/licence-ouverte-open-licence.
18. http://www.rip.justice.fr/information_publique_librement_reutilisable
19. http://data.ratp.fr/fileadmin/Documents/conditions_generales_dutilisation_0213.pdf.

La licence ODbL, mise à part le fait qu'elle soit principalement destinée aux bases de données, met en avant le principe de partage en essayant d'éviter une quelconque « privatisation » des données libérées. C'est une logique très honorable car cela permet de construire tout un écosystème autour d'un bien commun sans crainte que les efforts et les investissements ne soient détournés. Par ailleurs, c'est la raison pour laquelle lorsque tout le monde y participe, tout le monde en profite et tout le monde y gagne.

Quant à la Licence ouverte, elle prône davantage l'aspect libertaire dans le sens où le plus important est de réduire au maximum les contraintes imposées, à défaut de pouvoir les éliminer complètement, en garantissant en même temps un champ de manœuvre le plus large possible.

Ainsi, pour synthétiser, on peut dire qu'en ce qui concerne les droits accordés, ceux-ci se résument en droit à :

— « *reproduire, copier, publier et transmettre l'information ;*
— *diffuser et redistribuer l'information ;*
— *adapter, modifier, extraire et transformer à partir de l'information, notamment pour créer des informations dérivées ;*
— *exploiter l'Information à titre commercial, par exemple en la combinant avec d'autres informations, ou en l'incluant dans votre propre produit ou application.* » (Etalab, 2011).

Les différences qui ont été soulevées se trouvent surtout dans la partie qui gère les obligations et les limites d'application. Là où la Licence ouverte ne demande que le droit de paternité avec la mention de la mise à jour, la licence ODbL impose également le « partage dans les mêmes conditions ». Cela veut dire qu'une fois la licence acceptée, le résultat du travail doit être partagé sous la même licence ou une quelconque autre qui remplit cette condition.

Ces deux licences sont les représentants des deux groupes majeurs du monde de l'open data. L'un qui, comme la licence ODbL, la licence CC-BY-SA [20] ou encore la licence OGL [21] du gouvernement britannique, impose le partage dans les mêmes conditions. On parle également des licences virales ou contaminantes.

Dans l'autre groupe se trouvent les licences qui imposent uniquement la clause de paternité (attribution) avec la notion de la date de mise à jour, comme c'est le cas de la Licence ouverte d'Etalab, de la licence ODC-By [22] ou encore de la licence CC-BY [23].

Ceci dit, il existe en réalité un troisième groupe de licences. Ce groupe est composé de licences qui suppriment même l'obligation de l'attribution, c'est-à-dire que la notion de paternité n'est plus nécessaire. On y trouve les licences comme CC-0 [24] et PDDL [25]. Il faut toutefois ne pas oublier que ces licences peuvent, malgré (et en même temps à cause de) leur air libertaire absolu, poser de nombreux problèmes d'application car elles vont à l'encontre de la logique véhiculée par exemple par le principe du droit moral.

Nous proposons dans le tableau 1 un récapitulatif des licences majeures du monde de l'*open data*.

20. https://creativecommons.org/licenses/by-sa/2.0/fr/.
21. http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/.
22. http://opendatacommons.org/licenses/by/summary/.
23. https://creativecommons.org/licenses/by/2.0/fr/legalcode.
24. https://creativecommons.org/publicdomain/zero/1.0/legalcode.fr.
25. http://opendatacommons.org/licenses/pddl/summary/.

	Libertaires			Participatives
	Totalement	Partiellement		
Données ouvertes	CC-0	CC-BY	Licence Ouverte	CC-BY-SA
Bases de données	PDDL	ODC-BY		ODbL

TAB. 1 – *Licences majeures dans le monde de l'open data.*

3.2.1 Questions soulevées par l'usage des licences

Même si le paysage semble être à présent relativement dominé par les deux licences évoquées précédemment et qu'une relative stabilité s'installe, ce qui par ailleurs est une très bonne chose en soi, la situation peut ne pas durer. De par le fait que le concept même de l'*open data* reste assez récent et également du fait de diverses tensions déjà palpables comme, par exemple, les difficultés d'assumer toujours les investissements nécessaires pour libérer les données et le maintien des infrastructures et du personnel nécessaires. En même temps, certains continuent à chercher des solutions économiques qui peuvent permettre de participer à un maintien d'un équilibre budgétaire déjà grandement fragilisé.

Ainsi, par exemple, l'agglomération Grand Lyon qui a mis pendant un certain temps l'ensemble de ses données sous licence ODbL a décidé de ne pas maintenir cette logique à long terme. La volonté de valoriser économiquement les données ouvertes d'une manière ou d'une autre, sous divers prétextes, comme par exemple, celui de l'amélioration du service fourni, fait fréquemment son apparition. De ce fait, le Grand Lyon a choisi de ne pas garder une seule et unique licence (la licence ODbL) qui garantit un accès et un traitement uniforme à l'ensemble des données, et de s'ouvrir à d'autres possibilités comme passer une partie de ses données en licence ouverte d'Etalab et pour les autres de créer une licence avec authentification [26] (à l'image de l'intranet) ou encore de créer, dans le cadre de la valorisation du « *potentiel économique des données publiques* » (Grand Lyon Data, 2015), une licence payante qui se rapprocherait davantage d'une redevance. Tout cela pour concilier les besoins et les usages internes de la collectivité et la logique de l'ouverture des données et en même temps pour « *garantir un écosystème concurrentiel équitable, en évitant la formation de monopole* » (Grand Lyon Data, 2015). Mais le Grand Lyon n'est pas le seul à penser ainsi.

Le GART (Groupement des Autorités Responsables de Transport) tenait, à l'époque des fait, c'est-à-dire en 2012, à peu près le même discours : « *Dans le contexte des réflexions menées par le gouvernement et le Parlement sur la fiscalité du numérique, le GART préconise l'instauration d'une redevance liée à l'usage de ces données de service public. Il s'agirait, pour les AOT*[27]*, de percevoir une redevance assise sur les revenus générés par la publicité. Cette redevance d'usage serait dédiée au financement des transports publics* » (GART, 2012). Comme on peut le voir « *le mythe de l'argent de la publicité tombant du ciel, créé ex-nihilo, a la vie dure* » (Hervé, 2014). De plus, cela implique de nouvelles licences sur-mesure qui apportent davantage de confusion que de clarté et d'interopérabilité. La conséquence d'une telle logique peut de nouveau amener une part de flou dans cet écosystème déjà bien compliqué des licences disponibles, ce qui a été dénoncé à maintes reprises. Ainsi il faut faire très atten-

26. `http://data.grandlyon.com/connaitre-nos-licences/`.
27. Autorité organisatrice des transports.

tion à la multiplication des différents cadres juridiques car dans le passé, pas si lointain « *le manque de standard a donc généré l'utilisation et la création de licences diverses, pénalisant la lisibilité d'usage et l'interopérabilité des données* » (LiberTIC, 2011).

D'autre part, la situation autour des licences déjà existantes n'est pas pour autant dépourvue de difficultés de toutes sortes. Par exemple, la licence ODbL et l'ensemble des licences qui font partie du groupe dont elle est issue, posent certains obstacles pour une agrégation des données. Il faut se rendre compte qu'à l'heure actuelle, la majorité des applications se contente de travailler avec peu de données (en ce qui concerne leur diversité) et le passage à l'échelle n'a pas encore eu lieu. Tout cela ne sera certainement plus vrai demain ou dans un futur pas si lointain où l'agrégation des données de plusieurs (milliers ?) de sources va devenir monnaie courante. Et quand on dit plusieurs sources, on dit également, très souvent, plusieurs licences et ces licences peuvent ne pas toujours être compatibles entre elles. De ce fait, en voulant travailler avec des données ouvertes, on demeurera privé de tout un pan de données pour lesquelles les contraintes imposées par les licences ne seront pas compatibles avec les besoins et les buts de celui qui veut les réutiliser.

De plus, la question du passage à l'échelle soulève également une autre problématique, certes, plus technique et peut-être pas si contraignante à première vue, mais malgré tout intéressante à analyser. Il s'agit de la notion de paternité. La Licence ouverte d'Etalab n'impose qu'une seule contrainte, celle de mentionner la paternité. Toutefois, si cela ne pose pas de vrais problèmes avec peu de sources, la question peut devenir brulante dès qu'on se met à agréger des données dont les sources sont beaucoup plus importantes. Comment alors s'assurer d'un suivi efficace et fiable de l'ensemble des références qui ont le droit d'être mentionnées ? Sans oublier que pour chaque résultat issu d'une telle agrégation, il faut fournir une liste conséquente de toutes les sources qui ont été utilisées. On pourrait alors penser que la meilleure solution sera d'utiliser uniquement des licences comme CC-0 ou PDDL. Et effectivement, ces deux licences sont libres (en théorie) de toute contrainte. Par ailleurs, la logique même de l'ouverture des données semble être plutôt faite pour une telle vision des choses. Toutefois, comme on l'a déjà évoqué, le monde du libre est tiraillé par deux conceptions profondément opposées. La conception libertaire qui oscille autour de l'individu et ses libertés et la conception qui se base sur la notion du bien commun et qui vise comme unité minimale un groupe, une communauté. Ainsi, les licences CC-0 et PDDL, malgré leurs atouts forts, ne représentent qu'un versant de la réalité des choses qui régissent le monde de l'*open data*.

D'autre part, les licences, telles que l'ODbL ou CC-BY-SA, peuvent avoir aussi des côtés cachés à première vue. Par exemple, elles peuvent servir pour se protéger des différents géants du net contre lesquels ces licences semblent constituer une arme redoutable car nombreux sont ceux qui ne cachent pas leurs ambitions « économiques » vis-à-vis du monde de l'open data, comme c'est le cas du GART qui le dit ouvertement : « *Cette ouverture des données intéresse aussi tout particulièrement les opérateurs de transport, les géants du web (moteurs de recherche tels que Google, sites web ou acteurs de l'industrie informatique comme Apple) qui monétisent ces informations via l'intermédiaire de la publicité* » (GART, 2012). Ainsi, afin de se protéger, au moins un petit peu, contre ces tendances, on peut suivre le conseil trouvé sur le blog de Simon Chignard (auteur de « Open data : comprendre l'ouverture des données publiques » (Chignard, 2012c)) : « *si la cible c'est Google, alors ouvrons les données transport en privilégiant une licence ODbL ! Le moteur de recherche n'aime pas beaucoup les obligations liées à cette licence – et c'est d'ailleurs l'une des raisons de son adoption par Open*

Street Map (mémo : regardez aussi la licence utilisée par la SNCF). » (Chignard, 2012b). Cela ne date pas d'hier, que les différentes craintes, concernant les géants du net et leurs forces de frappe qui peuvent tout balayer sur leur chemin, émergent dans les esprits de tous ces acteurs qui n'occupent pas la même position qu'eux dans le paysage numérique car, comme le souligne sur son blog Simon Chignard, cette méfiance vis-à-vis des géants du net est « *partagée tant par les financeurs [du secteur des transports] que par les exploitants* » (Chignard, 2012b). Mais comme on l'a déjà dit, la chasse au ROI [28] n'épargne personne et ni Google, ni le GART, ni même le Grand Lyon, ne font exception.

3.3 Aspect technique

Un autre grand volet qui concerne les données ouvertes est celui qui traite certains des aspects techniques qui accompagnent le processus d'ouverture. Par ailleurs, la liste des 8 principes fondamentaux reflète assez bien cette séparation en deux groupes distincts : celui des aspects juridiques, déjà traités, et celui qui touche principalement le côté technique des choses. Ainsi, les principes de complétude, que des données soient brutes, mises à jour, accessibles, permettant un traitement automatisé et libérées dans un format ouvert constituent de fait, à la fois, un strict minimum pour pouvoir parler de données ouvertes et, en même temps, un réel défi car malgré une apparente simplicité de ces principes, les diverses réalités des faits du monde numérique (et pas seulement) peuvent mener la vie dure aux différents acteurs qu'il s'agisse des producteurs de données ou des distributeurs ou encore des réutilisateurs. La bonne maîtrise des aspects techniques est donc une condition sine qua non pour mener à bien des projets dans le monde de l'*open data*.

3.3.1 Nature des données

Sans vouloir rentrer dans une discussion philosophique sur la nature des données, il faut malgré tout admettre que la question n'est pas simple. Chaque domaine qui travaille avec des « données » a une représentation spécifique (Fayet, 2013), propre à lui, de ce qu'il comprend sous la notion de donnée. Fréquemment, une donnée se présente comme « *une description élémentaire, typiquement numérique pour nous, d'une réalité. C'est par exemple une observation ou une mesure* » (Abiteboul, 2012). Par ailleurs, la plupart des définitions qu'on trouve à ce sujet vont à peu près dans cette direction. Dans Wikipédia, on peut même trouver qu'« *elles sont dépourvues de tous raisonnements, suppositions, constatations, probabilités, qui, étant indiscutables ou indiscutées, servent de base à une recherche, à un examen quelconque* » (Wikipédia, 2015) ce qui, par ailleurs, va dans le sens que lui donne, par exemple, le Larousse où on peut lire qu'une donnée, c'est « *ce qui est connu ou admis comme tel, sur lequel on peut fonder un raisonnement, qui sert de point de départ pour une recherche* » (Larousse, 2015).

Globalement, une donnée, si on applique une réduction définitoire, peut être représentée comme un « fait ». Par un procédé de contextualisation, ce « fait » devient une information. Ainsi, par exemple, un tableau statistique est rempli de « faits » et le tableau en lui-même représente la contextualisation, ce qui permet d'en extraire des informations. C'est simplifié, on l'avoue, mais c'est, à peu près, de cette manière que les choses se présentent dans le cadre du sens commun par rapport à la notion de donnée. Et c'est très important d'en être conscient

28. Return On Investment, soit retour sur investissement.

car cette manière de penser et de faire peut avoir un impact même sur la manière de voir et d'appréhender les données ouvertes elles-mêmes.

Regardons maintenant, afin d'être plus concrets, dans le tableau 2, la liste qui est issue de la Charte du G8 pour l'Ouverture des Données Publiques (G8, 2013) et qui fournit des exemples de catégories d'ensembles de données (par ordre alphabétique).

Il n'est pas difficile de remarquer que la plupart des données correspondent d'une manière ou d'une autre à cette définition où un « fait », c'est-à-dire une donnée, est automatiquement (pour ne pas dire uniquement) une valeur chiffrée ou alphabétique. Leur nature et forme permettent de les manipuler aisément et de mener, par exemple, des traitements automatisés, et surtout, d'effectuer différentes analyses statistiques et visualisations. Toutefois, « *la notion de données elle-même est à géométrie variable* » (Fayet, 2013). C'est un piège de croire que toutes les données ont cette nature. En fait, cela dépend, en grande partie, du domaine de provenance de ces données. Ainsi, par exemple, « *les « données » d'un linguiste peuvent [être] des écrits ou des discours, des enregistrements de locuteurs ; les « données » d'un médiéviste sont des sources archivistiques, archéologiques, épigraphiques, iconographiques, littéraires ; les « données » d'un géologue rassemblent des coupes et observations de terrain consignées sur un carnet, des résultats de carottage, des analyses d'échantillons, des données sismographiques… Bref il n'y a pas, et loin de là, que des données quantitatives de même nature qui constitueraient des séries homogènes relativement faciles à manipuler, échanger et compiler* » (Fayet, 2013). En conséquence la situation est beaucoup plus complexe que pensent ou veulent faire admettre ceux qui comme Tim Berners-Lee ont lors d'une conférence TED en 2009 scandé après lui le fameux appel « *Raw data now* » (Berners-Lee, 2010). Cette divergence, en ce qui concerne la nature des données, représente de nombreux obstacles qui n'ont pas été tous franchis et cela même en dehors du sujet de l'*open data* en tant que tel, et pour de nombreuses questions, on ne connait pas encore de réponses entièrement satisfaisantes comme, par exemple, pour le domaine du traitement automatique des langues et l'analyse des textes, ou encore pour la reconnaissance des tableaux et autres structures similaires, et ce qui va avec le contenu des jeux de données (à titre d'exemple).

3.3.2 Données brute ou donnée enrichie ?

D'autre part, le flou qui entoure la notion de la donnée elle-même se projette bien évidemment sur l'ensemble des notions dérivées d'elle. La donnée brute en est un parfait exemple car qu'est-ce que cela veut dire une donnée brute ? À la base, on prend comme donnée brute, l'ensemble des données dites « non traitées » qui se présentent telles quelles. Les relevés, les mesures, les dates, etc., constituent de très bons exemples de ce genre de données. Toutefois, comme on l'a déjà dit, toutes les données ne se présentent pas uniquement sous ces aspects. On pense surtout aux données des diverses disciplines issues, par exemple, des sciences humaines et sociales. De plus, pour qu'une donnée reste exploitable, il faut qu'elle soit accompagnée d'un minimum d'informations, ce qui permet de connaître la nature de la réalité qu'elle est censée refléter ou du moins représenter. Il reste maintenant à savoir quel est le degré de cet enrichissement nécessaire et suffisant, car il s'agit de cela, qu'est-ce qui provoque qu'une donnée peut être encore considérée comme brute ? Et cela d'autant plus qu'un enrichissement peut s'effectuer à plusieurs niveaux. D'une part directement dans les données elles-mêmes et d'autre part par le rajout d'informations complémentaires les décrivant, appelées communément les métadonnées. Cependant, il semble, que dans le cadre de la vision des données, telles qu'elles sont

Criminalité et justice	Statistiques sur la criminalité Sécurité
Développement mondial	Aide au développement Sécurité alimentaire Industries extractives Terres
Données géospatiales	Topographie Codes postaux Cartes nationales ou locales
Éducation	Liste des écoles Valeur ajoutée Compétences numériques
Entreprises	Registre des entreprises
Environnement	Niveaux de pollution Consommation énergétique
Finances et marchés	Valeur des transactions Marchés publics attribués ou à venir Budget local ou national (prévu et exécuté)
Mobilité et protection sociales	Logement Prestations sociales Assurance-maladie et assurance-chômage
Observation de la Terre	Guichets et points de contact des administrations Agriculture, foresterie Pêche et chasse
Responsabilisation des gouvernements et démocratie	Aide au développement Résultats des élections Lois et règlements, salaires (échelles salariales) Dons
Santé	Données issues de prescriptions Données de performance
Science et recherche	Données relatives au génome humain Recherche et activités pédagogiques Résultats d'expérience
Statistiques	Statistiques nationales Recensements Infrastructure Statistiques économiques et éducatives
Transport et infrastructure	Horaires des transports publics Services à large bande

TAB. 2 – *Exemples des catégories d'ensembles de données (par ordre alphabétique) issus de la Charte du G8 pour l'Ouverture des Données Publiques.*

perçues, par exemple, par les administrations de l'État, on pense principalement aux données comme mesures, relevés, horaires, coordonnées géospatiales, etc. Dans ce cas bien précis, on peut s'attendre à des données qui se traduisent principalement par la logique dans le cadre de laquelle on attribue à un groupe de variables bien définies un ensemble de valeurs normalisées qui se présentent sous une forme non agrégée [29].

Y a-t-il vraiment des données brutes ? Cette difficulté de pouvoir définir d'une manière claire les données « brutes » provoque que certains pensent qu'en réalité une telle chose comme la donnée brute n'existe pas. On peut trouver une telle vision des choses, par ailleurs avec des exemples forts intéressants sur le blog de Simon Chignard (Chignard, 2012a). Toutes données (même celles qui prennent la forme d'une mesure ou un autre caractère chiffré ou non) sont issues d'une volonté, d'une intentionnalité, d'une interprétation de la réalité, ce qui engendre que de telles données sont par définition biaisées. De très bons exemples de cette situation se trouvent dans l'ouvrage « *Raw data is un Oxymoron* ». Plusieurs articles qui se trouvent dans cet ouvrage montrent justement à quel point il est difficile de parler des données en tant que données brutes. Par exemple, Brine et Poovey dans « *From Measuring Desire to Quantifying Expectations* » (Brine R. K. et Poovey, 2013), nous disent que les données économiques ne sont jamais brutes, au sens de non interprétées. En fait, l'économiste qui doit nécessairement passer par l'étape de *data scrubbing* (nettoyage des données) efface, en quelques sortes, l'histoire de ces données et produit, en réalité, des données enrichies afin qu'elles puissent « mieux » refléter la réalité censée décrire, prises ainsi comme objectives à l'égard de son analyse. Le choix des données (un acte délibérément réducteur) et l'ensemble des informations supplémentaires qui servent à décrire ces mêmes données (ce qui constitue leur enrichissement) vont toujours de pair. Par exemple, une « *enquête auprès d'un laboratoire de mécanique des fluides a montré que les données brutes issues des expériences étaient soumises à des modèles de validation, qui permettaient d'éliminer certains résultats non probants, les « fausses données » (sic) pour ne conserver que les « vraies données » (re-sic) : les données ici ne sont pas un matériau brut, mais le résultat d'une première opération scientifique, fruit d'une méthodologie propre à la discipline ; dans ce cas, le profane est incapable de définir les contours de la masse de données* » (Fayet, 2013). On peut trouver à peu près le même message également dans l'article « *Data Bite Man : The Work of Sustaining a Long-Term Study* » (Ribes et Jackson, 2013) où il est dit clairement que les données, sans leur entourage informationnel les contextualisant (les métadonnées, dans ce cas précis), perdent leur lien avec la réalité de laquelle elles sont issues. Ce qu'on peut traduire par une sorte de paradoxe qu'un « fait » devient une donnée brute uniquement par un enrichissement informationnel. Dans ce sens, la « *data friction* » (Edwards et al., 2011) peut constituer un pan de recherche en soi et à lui tout seul.

La question devient, par ailleurs, d'autant plus compliquée et complexe si on abandonne la vision d'une donnée comme uniquement chiffrée ou représentée d'une manière générale comme une valeur, de préférence discrète. Comme on l'a déjà évoqué, les données travaillées dans le cadre de la recherche scientifique ne représentent pas toujours une suite de nombres ou ayant un autre caractère facilement traitable. À titre d'exemple, on peut évoquer l'article « *Where Is That Moon, Anyway ?* » (Stanley, 2013), issu du même ouvrage « *Raw data is*

29. On peut trouver un bon exemple de données agrégées sur le site de balises-rhone-alpes.org qui, même si libres d'accès, ne sont pourtant pas des données ouvertes et encore moins des données brutes (`http://www.balises-rhone-alpes.org/pages/obs_loc/interrogation.php`, consultée le 15 mai 2015).

an Oxymoron », dans lequel il est montré à quel point un astronome qui veut comprendre, par exemple, l'histoire des éclipses solaires doit non seulement travailler avec les données issues des observations directes mais qu'il doit également se plonger dans tous les ouvrages historiques qui en parlent ; en sachant qu'on se trouve de ce fait dans un autre univers des données – des histoires, des chroniques, des mémoires, des tableaux (en tant que peintures) ou autres représentations textuelles, graphiques, etc. Tout cela pour dire qu'il faut être très vigilant quant à la nature des données, surtout si elles sont prises en tant que données « brutes » car les données ouvertes ne concernent pas toujours et uniquement des tableaux de l'INSEE ou des horaires de la RATP.

3.3.3 L'empilement des métadonnées

Un des éléments auxquels sont confrontés tous les acteurs de l'*open data* est la question des métadonnées. Ces « données sur les données » constituent en quelque sorte la carte d'identité de la donnée qu'elles accompagnent. Pourtant, établir cette « carte » n'est pas toujours si évident qu'il n'y parait, même si les initiatives ne manquent pas [30]. Non seulement qu'il faut trouver quelles informations sont nécessaires et suffisantes mais également de quelle manière ces informations doivent être représentées afin de ne pas se retrouver dans une situation où on sera obligé de créer des métadonnées pour les métadonnées et ainsi de suite. Les informations doivent être lisibles et compréhensibles. Mises à part des informations « standards » telles que, par exemple, les identifiants de producteurs de ces données qui correspondent, par exemple, aux métadonnées de « gestion », une information un peu particulière peut toutefois porter à confusion. Il s'agit des catégories censées décrire le sujet lui-même en lui permettant ainsi de trouver une place dans un répertoire plus vaste de données – les métadonnées de description. On parle, par exemple, de l'indication de la thématique, du sujet, des mots-clés, des descripteurs, des tags, des termes d'indexation etc. Ces indications représentent un accès privilégié aux données dès qu'elles se trouvent dans un ensemble plus large et c'est pour cela qu'il est impératif de bien les choisir.

La métadonnée « catégorie », indispensable moyen d'accès à la donnée Les catégories sont, en quelque sorte, le reflet de notre façon d'appréhender le monde dans lequel on vit. Elles expriment la manière dont on s'acquitte de notre réalité et on codifie cette représentation. Et comme la compréhension de celle-ci n'est pas toujours homogène et encore moins identique, on se retrouve avec des variations plus ou moins importantes dans ces représentations. Sans oublier que dès qu'on se met à communiquer ces représentations, les choix langagiers qui sont censés les traduire aggravent davantage la confusion qui peut s'installer. Cependant, l'aspect subjectif de choix des catégories est également biaisé, sans que la liste soit complète, par la finalité de l'usage de la donnée et par le poids de l'autorité de celui qui les élabore. D'autre part, une bonne catégorisation est dépendante également du niveau de la connaissance du sujet et bien évidemment des compétences de celui qui travaille sur ces métadonnées. Tout ce qui précède a pour résultat que les catégories peuvent être, parfois, plus ou moins fantaisistes, correctes ou en adéquation avec la donnée qu'elles sont censées représenter, décrire ce qui met donc à rude épreuve l'interopérabilité et, en fin de compte, la compréhension de ces données.

30. Par exemple, voir sur le site du Conseil National de l'Information Géographique à l'adresse suivante : `http://cnig.gouv.fr/?page_id=2916`.

C'est d'ailleurs une des raisons pour lesquelles de nombreuses initiatives d'harmonisation de ce processus ont vu le jour, parmi lesquelles :

— faire réaliser cette étape de traitement des données par un spécialiste ;
— mettre en place des outils collaboratifs du type « crowdtagging » ;
— utiliser les référentiels destinés à cet effet, etc.

Le Crowdtagging, quant à lui, est une approche intéressante dans le sens où c'est la communauté toute entière qui participe à la rédaction des métadonnées « descriptives ». Celui-ci assure non seulement une large validation des choix effectués mais également un enrichissement des possibilités de description. Dans le cadre d'un usage bien encadré, cela peut s'avérer être un excellent choix.

Quant aux référentiels (d'indexation), ceux-ci permettent de mieux cibler et d'homogénéiser les pratiques en rendant les métadonnées davantage interopérables entre différents systèmes existants. Ces référentiels, libres ou contrôlés, communautaires (bottom up) ou institutionnels (top down), locaux ou généralisés sont, en principe, à l'image des domaines dont ils sont issus et des besoins qu'ils reflètent (sphère d'activité bibliographique, scientifique, d'activité politique, économique et sociale, bases de connaissances qui ne sont pas des thésaurus, manient plus les concepts que les formes).

Les schémas de catalogue de données Dans le cadre de l'interopérabilité on peut se tourner vers Data Catalog Vocabulary (DCAT) ou Catalogue Service for the Web (CSW) qui devient en France, à l'heure actuelle, un des standards de fait. Les outils ne manquent pas, mais comme on peut le voir leur quantité peut toutefois rendre perplexe.

INSPIRE, un exemple abouti de schéma de données L'importance de l'interopérabilité des données n'est plus à démontrer. C'est d'ailleurs pour cela que l'Union Européenne a mis en place en 2007, via la directive 2007/2/CE du Parlement européen et du Conseil, l'infrastructure d'information géographique dans la Communauté européenne appelée INSPIRE[31]. Le rôle de cette infrastructure est justement d'encadrer l'harmonisation et l'interopérabilité des systèmes de métadonnées à l'échelle européenne, même si dans ce cas précis, il s'agit principalement d'informations géographiques. De ce fait « *les infrastructures d'information géographique dans les États membres devraient être conçues de façon à ce que les données géographiques soient stockées, mises à disposition et maintenues au niveau le plus approprié, qu'il soit possible de combiner de manière cohérente des données géographiques tirées de différentes sources dans la Communauté et de les partager entre plusieurs utilisateurs et applications, que les données géographiques recueillies à un niveau de l'autorité publique puissent être mises en commun entre les autres autorités publiques, que les données géographiques soient mises à disposition dans des conditions qui ne fassent pas indûment obstacle à leur utilisation extensive, qu'il soit aisé de rechercher les données géographiques disponibles, d'évaluer leur adéquation au but poursuivi et de connaître les conditions applicables à leur utilisation* » (Directive, 2007). Il faut dire que c'est une très bonne initiative, certes, peu connue pour le moment, mais qui constitue un grand pas en avant pour une unification des pratiques et cela à l'échelle européenne. La seule chose qu'on puisse regretter est le fait que cela ne concerne, pour le moment, que des données bien précises, en l'occurrence les données environnementales même si d'autres possibilités d'usage apparaissent.

31. `http://inspire.ec.europa.eu/`.

Format d'échange du CKAN Etalab À une échelle plus modeste, car française, un autre « standard » qui semble s'imposer, sans qu'il soit encore dominant à l'heure actuelle, est l'initiative de la mission Etalab qui a repris l'application de gestion des données CKAN (Comprehensive Knowledge Archive Network), soutenue par Open Knowledge Foundation, en l'adaptant à la réalité d'utilisation locale (au sens national). À la base CKAN est une plateforme de stockage et de distribution des données adoptée par de nombreux organismes à travers le monde. Etalab a retenu la logique de gestion des données par cette plateforme en adoptant le format d'échange que cela impose. De par son soutien par Etalab et de par son utilisation par le site data.gouv.fr, la solution CKAN a su gagner ses lettres de noblesse et devenir une alternative incontournable dans le paysage des données ouvertes en France.

Autres solutions, autres schémas Ceci dit, il y a aussi d'autres solutions qui peuvent, le cas échéant, être prises en compte au moment de la publication des données ouvertes. Parmi elles, on trouve la solution SaaS de l'OpenDataSoft [32] et celle se basant sur Typo3 [33] auquel peut être intégré un module de dépôt de données initialement développé par la société In Cité Solution [34]. Toutefois, à l'heure actuelle, la société In Cité Solution n'existe plus et le maintien de sa solution par les acteurs encore présents peut ne pas être assuré à long terme.

Les données ouvertes en route vers les Linked data Le processus de la mise en place des données ouvertes peut être évalué également à l'aide d'un système basé sur le décernement d'étoiles, comme dans le guide Michelin, qui indique quel est le degré d'ouverture de ces données (voir figure 1 et tableau 3).

⋆	make your stuff available on the Web (whatever format) under an open license
⋆⋆	make it available as structured data (e.g., Excel instead of image scan of a table)
⋆⋆⋆	use non-proprietary formats (e.g., CSV instead of Excel)
⋆⋆⋆⋆	use URIs to denote things, so that people can point at your stuff
⋆⋆⋆⋆⋆	link your data to other data to provide context

TAB. 3 – *Système d'étoiles de l'échelle d'ouverture des données. Source : http ://5star-data.info/*

Il faut toutefois préciser que ce système, développé par Tim Berners-Lee, va légèrement au-delà des 8 principes fondamentaux de l'*open data*. En principe, quand on regarde les 8 préconisations, on s'aperçoit qu'elles s'arrêtent au point 4 de l'échelle étoilée. C'est-à-dire que la mise à disposition de la donnée, en indiquant un lien pour pouvoir y accéder, doit largement suffire pour pouvoir parler de données ouvertes. Toutefois, ce qui est visé par cette nouvelle échelle est de ne pas uniquement mettre à disposition les données telles quelles mais de les relier entre elles afin d'en constituer un large réseau (mondial) de données [35]. On ne relie plus uniquement une page d'un site internet à une autre ou des bases de données entre elles mais

32. `https://www.opendatasoft.com/open-data-solutions/`.
33. `http://typo3.org/`.
34. `http://www.incitesolution.fr`.
35. A titre d'exemple on peut mentionner le site `http://lod-cloud.net/`.

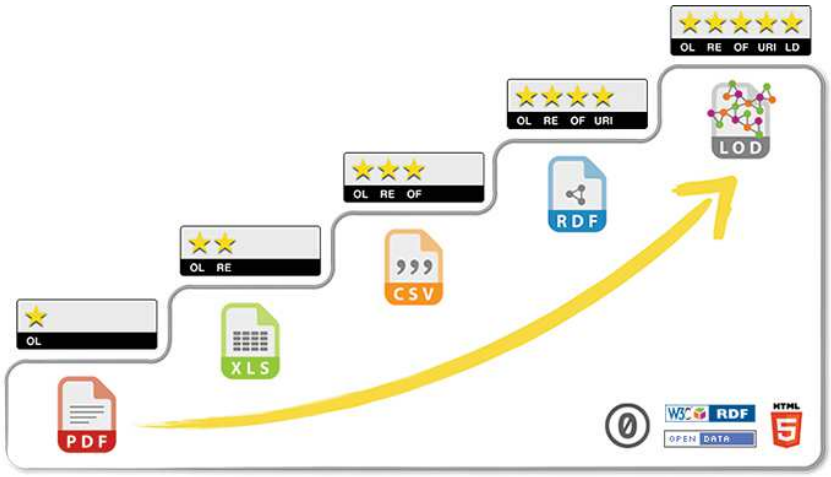


FIG. 1 – *Echelle d'ouverture des données. Source : http ://5stardata.info/, licence CCO 1.0 Universal 1.0 Transfert dans le domaine public.*

on relie les données elles-mêmes. On rentre de ce fait dans un monde des données liées ou autrement de Linked data. Cette vision des choses est très intéressante à plusieurs égards mais le plus important est le fait qu'on minimise au maximum la nécessité de duplication des données car on peut y accéder à tout moment quelle que soit leur position. Par ailleurs, de nombreux acteurs comme Eurostat ou le portail data.gouv.fr tentent d'intégrer cette logique des choses par défaut dans leur processus de mise à disposition de leurs données. Certes, à ce jour, on ne rentre que rarement dans des documents eux-mêmes afin d'en extraire les données qu'on peut relier entre elles à l'aide du RDF (Resource Description Framework). Cette approche se limite, semble-t-il, à l'heure actuelle, principalement à rendre interopérables les métadonnées qui accompagnent ces données, comme peut en témoigner le projet Datalift [36] qui a été élaboré pour pouvoir transformer les différents formats de structuration et de représentation des métadonnées en format compatible avec la logique de Linked data. On peut trouver une de ces transformations sur le portail de data.gouv.fr en cherchant, par exemple, le jeu de données concernant les ressources pédagogiques pour l'enseignement de l'histoire des arts [37].

3.3.4 Données statiques ou dynamiques ?

Les données (brutes ou non) peuvent être regardées également par le prisme de leur « finitude ». Certaines données sont dites « définitives » (on ne prévoit pas que ces données changent à l'avenir) ; d'autres doivent, ou sont censées, être modifiées par une mise à jour, plus ou moins fréquente, et d'autres encore constituent une sorte de flux plus ou moins important, plus ou moins constant. Connaitre la nature des données, en ce qui concerne cet aspect peut s'avérer

36. `http://datalift.org/?page_id=2`.
37. `https://www.data.gouv.fr/fr/datasets/jeux-de-donnees-de-ressources-pedagogiques-pour-l-enseignement-de-l-histoire-des-arts-en-rdf/`.

important dans le sens où l'on considère souvent la mise à jour comme un critère de « fiabilité », à la fois des données et de celui qui les met à disposition. Toutefois, il faut se rendre compte que ce critère peut dans de nombreux cas être même préjudiciable. Certaines données ne sont tout simplement jamais mises à jour car cela n'est pas nécessaire et même dénué de sens. Se baser ainsi sur ce critère peut ne pas refléter la réalité des choses.

On peut alors distinguer trois principaux groupes de données en ce qui concerne la fréquence de leur mise à jour :

— **données statiques** - Par exemple, le nombre de naissances ou de décès pour l'année 2000 n'est pas censé changer en principe, même si des corrections ne sont pas à exclure dans le futur (Chignard, 2012a). On peut appeler ces données – statiques ;
— **données dynamiques** - Un autre groupe de données, dites dynamiques, est représenté par les données dont des modifications, ou des mises à jour, sont à prévoir ou obligatoires et les changements, même si fréquents, sont effectués dans une limite raisonnable. Cela concerne, par exemple, des données comme les valeurs hebdomadaires des prix des carburants ou encore les horaires dans les transports publiques ;
— **données fluides** - Puis, on peut parler des données fluides, qui constituent un flux plus ou moins important et « ininterrompu » de données, dont le traitement s'effectue pour la plupart en temps réel (même si cela ne constitue pas non plus un critère absolu). Les informations sur la circulation ou sur le positionnement des Vélib', représentent de parfaits exemples de ce type de données.

3.3.5 Formats des fichiers de distribution

La question du format de distribution a une importance à la fois du point de vue de l'accès, de l'interopérabilité et de la pérennité. Il faut s'assurer que le format distribué peut être facilement accessible, c'est-à-dire que l'accès aux données n'est pas « verrouillé » par la nécessité de faire appel à des logiciels spéciaux dont l'utilisation est soumise, par exemple, à des contraintes financières ; il doit être sans difficulté transformable (dans le sens de traductible) d'un format à l'autre, et bien sûr il doit être pérenne.

Formats ouverts vs formats fermés Ces trois conditions évoquées auparavant peuvent être également traduites comme une opposition entre les formats ouverts et fermés. Un des critères pour les données ouvertes part du fait que celles-ci doivent être (dès que cela est possible) publiées dans un format ouvert, c'est-à-dire non propriétaire, un format qui ne dépend pas des « caprices » d'un quelconque « monopoliste ». C'est très important car non seulement les formats propriétaires sont souvent payants (en principe via les logiciels de gestion et de traitement de ces données, sans que cela soit toujours le cas), limités par des clauses d'utilisation définies par des licences et n'offrent pas toujours un accès à l'ensemble des spécifications que leur traitement permet ou impose. L'incompatibilité entre les formats et les logiciels est monnaie courante et constitue par ailleurs, dans de nombreux cas, la philosophie même du fond de commerce des éditeurs et gestionnaires de ces logiciels et de ces formats. De ce fait, la règle de rigueur est de les éviter dès que cela est possible. Il faut toutefois dire que cela n'est pas toujours évident, et cela d'autant plus que certains formats, même si propriétaires, ont acquis des positions de standard de fait. C'est le cas, par exemple, des formats qui sont gérés par la suite bureautique Office de Microsoft comme xls/xlsx pour Excel ou doc/docx pour Word ou encore le format pdf de la société Adobe. Cependant, malgré le fait que les spécifications pour le

format pdf sont entièrement accessibles (sur plus de 1300 pages !) (Adobe, 1993) et qu'Adobe encourage le développement d'un écosystème autour de ce format, ou encore que xlsx et docx représentent des versions « ouvertes » de leurs prédécesseurs, cela ne change rien au fait que ces format sont couverts par divers droits de propriété intellectuelle détenus par des sociétés respectives. Il faut toutefois admettre qu'à l'égard de l'article 4 de la loi n° 2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique, même ces formats peuvent, le cas échéant, être utilisés en tant que formats de base car selon ce dispositif « *on entend par standard ouvert tout protocole de communication, d'interconnexion ou d'échange et tout format de données interopérable et dont les spécifications techniques sont publiques et sans restriction d'accès ni de mise en œuvre* » (Loi, 2004). Ainsi, le critère des droits de propriété ne peut pas être utilisé d'une manière globale et irréfléchie car il ne fait pas partie des conditions éliminatoires pour une non utilisation de tel ou tel format tant que ces spécifications complètes sont disponibles (au moins selon la loi). Pourtant, le fait qu'un format soit propriétaire, « toléré » par la loi et standard d'usage dans les faits, cela ne remet pas en cause le bien-fondé des réticences vis-à-vis d'un tel format du point de vue, par exemple, de la pérennité d'usage car rien ne garantit que la « politique » de la société à laquelle ce format appartient ne change pas un jour et cela pour une quelconque raison. Par ailleurs, l'exemple du format gif peut servir de leçon [38].

Principe et contrainte du traitement automatisé Dans le point 5 de la liste des 8 principes fondamentaux de l'open data, on parle de la possibilité de pouvoir traiter les fichiers d'une manière automatisée. En d'autres termes on veut s'assurer que les formats choisis permettent d'accéder aux contenus de ces fichiers d'une manière « non-surveillée » pour en extraire des informations (ou les données) pertinentes. Regardons maintenant ce que l'on trouve à ce sujet dans la Directive 2013/37/UE du Parlement Européen et du Conseil : « *Un document devrait être considéré comme présenté sous un format lisible par une machine s'il se présente dans un format de fichier structuré de telle manière que des applications logicielles puissent facilement identifier et reconnaître des données spécifiques qu'il contient et les en extraire. Les données encodées présentes dans des fichiers qui sont structurés dans un format lisible par une machine sont des données lisibles par la machine. Les formats lisibles par la machine peuvent être ouverts ou propriétaires ; il peut s'agir de normes formelles ou non. Les documents encodés dans un format de fichier qui limite le traitement automatique, en raison du fait que les données ne peuvent pas, ou ne peuvent pas facilement, être extraites de ces documents, ne devraient pas être considérés comme des documents dans des formats lisibles par la machine. Les États membres devraient, le cas échéant, encourager l'utilisation de formats ouverts, lisibles par la machine* » (Directive, 2013).

À cet égard, il est clair que tous les formats ne se prêtent pas de la même manière à ce genre de traitement. Par exemple le format pdf qui est un format par excellence pour une visualisation, avec une forte indépendance vis-à-vis des différentes plateformes, ne permet pas toujours, d'une manière simple, d'accéder à son contenu, au sens de l'extraire. Et le format tel que xls n'est pas, malgré les apparences, mieux placé (Berro et al., 2014). Il ne suffit pas ainsi uniquement de publier les données mais également de s'assurer que les données qui y sont contenues peuvent être extraites.

Cependant, là on touche à un autre problème beaucoup plus important. À la base, les humains et les « machines » ont des approches différentes quant aux besoins de la représentation

38. Pour plus d'information, voir `http://cloanto.com/users/mcb/19950127giflzw.html`.

des données et surtout des documents qui les enferment. Là où l'ordinateur ne voit que les données, l'homme voit à la fois les données et les informations qui lui permettent d'accéder aux connaissances. La structuration de ces données fait partie intégrante du processus de transformation de ces données en informations.

Pour les humains, par exemple, un titre, sa taille, son emplacement et l'enchaînement constituent autant d'éléments nécessaires à une bonne compréhension du contenu d'un document. L'ordinateur pour sa part est, en principe, dépourvu de ce genre de « capacités ». Dans ce sens, il y a une différence de taille si les données font partie, par exemple, d'un tableau ou sont directement intégrées dans un texte. Les recherches sur l'automatisation de traitement tel que la reconnaissance, la captation et l'analyse des données présentes dans des tableaux et des textes avancent bien, même si non sans difficulté. De nombreux travaux se sont attaqués à ces sujets avec plus ou moins de succès car les contraintes sont immenses. Parmi celles-ci on peut mentionner le fait que la notion même du tableau n'est pas, en réalité, proprement définie (Embley et al., 2006) ou (Zanibbi et al., 2004), et que l'extraction des données directement du texte demeure un défi. C'est d'ailleurs une des raisons pour lesquelles le processus d'enrichissement de ces données et la création des métadonnées peuvent, voire doivent, avoir lieu. On indique à l'ordinateur, à l'aide des balises par exemple, quelle est la structuration du texte, ou tout simplement du document, qu'il est en train de traiter. Et c'est principalement grâce à ces « indications » qu'un document ou un jeu de données devient « lisible » pour un ordinateur dans le sens de la « lisibilité » telle qu'elle est perçue par les humains. On rentre de ce fait de nouveau dans le domaine de l'enrichissement des données et des métadonnées. Il faut toujours indiquer la nature des données et leur structuration car la reconnaissance automatique n'est pas encore au point et la compréhension ne fait pas partie des capacités inhérentes aux ordinateurs.

3.3.6 Principe d'accessibilité

L'élément le plus important dans le processus de la mise à disposition des données ouvertes est de s'assurer que celles-ci sont bel et bien disponibles. Cette disponibilité peut être effectuée de plusieurs manières : directe, indirecte et hybride. Par ailleurs, cela reflète également la manière d'appréhender la mise à disposition des données ouvertes en favorisant plutôt les approches tournées de préférence vers l'homme (directe et indirecte) ou celles misant avant tout sur la simplification de la communication entre les « machines » (approche hybride).

Accessibilité directe La mise à disposition s'effectue via un lien qui mène directement soit vers le jeu de données quel que soit son format soit vers le fichier qui contient l'ensemble des liens présents dans le catalogue.

Liens unitaires Ainsi, en guise d'exemple pour les liens unitaires, on peut évoquer celui de la liste des établissements d'enseignement supérieur qui se trouvent sur le portail des données ouvertes de l'Île de France [39] où les données sont présentes en plusieurs formats, prêtes à être téléchargées.

Liens vers les catalogues Certains acteurs mettent à disposition les liens vers les catalogues entiers. C'est une solution très pratique surtout si on veut permettre aux autres de

39. `http://data.iledefrance.fr/explore/`.

pouvoir accéder facilement à l'ensemble des données possédées pour gérer, par exemple, les relations entre les différentes jeux de données ou les intégrer à une structure plus large du type data.gouv.fr. Parmi ces bons élèves on trouve :

— Grand Lyon Data en CSV : http ://data.grandlyon.com/
— Montpellier numérique : http ://opendata.montpelliernumerique.fr/Les-donnees
— Open PACA : http ://opendata.regionpaca.fr/donnees.html
— Île-de-France : http ://data.iledefrance.fr/explore/
— Région Pays de la Loire : http ://data.paysdelaloire.fr/donnees/statistiques-des-donnees/

La liste n'est bien évidemment pas exhaustive.

Accessibilité indirecte Une solution très courante parmi certains acteurs est de « noyer » les liens pointant vers les jeux de données sur une page internet qui ne se prête pas facilement à un téléchargement direct et cela d'autant plus que même l'identification du fichier lui-même n'est pas des plus simples et évidents. À titre d'exemple, on peut évoquer le cas de l'INSEE.

Ainsi, si on cherche, par exemple, les informations sur le recensement de la population en 2010, on trouve :

— le jeu de données de l'INSEE via data.gouv.fr : https ://www.data.gouv.fr/fr/datasets/bases-de-donnees-et-fichiers-detail-du-recensement-de-la-population-2010/ ;
— ce même jeu de données via le site de l'INSEE : http ://www.insee.fr/fr/bases-de-donnees/default.asp ?page=recensement/resultats/2010/donnees-detaillees-recensement-2010.htm.

Formulaire – solution à éviter Une autre manière indirecte est représentée par la mise à disposition des données via un formulaire à choix multiples. À première vue, c'est une solution intéressante car on peut, via ces formulaires, choisir soi-même différents aspects sous lesquels ces données peuvent se présenter. Toutefois, cela empêche, la plupart du temps, un quelconque traitement automatisé de récupération de ces données. C'est une solution à éviter malgré les apparences d'une simplification de traitement.

Accès hybride : les API Il y a toutefois d'autres solutions pour pouvoir accéder aux données ouvertes. Parmi elles on trouve celles basées sur des API (*Application Programming Interface*). Cette « interface de programmation » représentée par « *un ensemble normalisé de classes, de méthodes ou de fonctions qui sert de façade par laquelle un logiciel offre des services à d'autres logiciels* » (Wikipédia, 2014) et qui permet une séparation entre les données et l'interface de visualisation devient de plus en plus répandue, même dans le monde de l'*open data*. Cette nouvelle approche est liée à un passage du mode « data culture » qui privilégie la mise à disposition des différents jeux de données via les liens de téléchargement à un mode régi par les API qui se basent, quant à eux, sur une logique de questions-réponses. Ainsi, on ne publie plus les données d'une manière directe mais on met à disposition de tout un chacun une interface qui permet d'y accéder via différentes requêtes. Bien évidemment, personne n'échappe à l'obligation de constituer des catalogues de données. Ce qui change c'est la manière d'y accéder.

De plus, la force des API est qu'ils permettent de répondre aux défis posés par « *la rapidité du service et la scalabilité des architectures logicielles* » (Fauré, 2012). Les API constituent donc un excellent moyen de libération des données.

Cependant, il faut également pointer quelques difficultés que cela fait surgir. Parmi les plus importantes, on peut évoquer la richesse de diversité dans le monde des API où on croise à la fois différentes solutions « maison », comme celle de Bordeaux [40], des API communautaires, comme celle de CKAN [41], ou encore les API développés par des acteurs privés du monde de l'*open data*, comme c'est le cas de l'API de la société OpenDataSoft [42]. Cela impose, par ailleurs, surtout si on se positionne au niveau du distributeur/passerelle ou de celui qui travaille sur l'agrégation des données de différentes sources, qu'on surveille constamment les cycles de vie de l'ensemble des API qu'on utilise. Par exemple, il n'est pas rare de voir les API changer leur « comportement » dès qu'ils passent d'une version à une autre. Bref, c'est une solution très commode mais à risque car une surveillance constante est de mise.

3.4 Aspect économique

Pendant de nombreuses années, les données constituaient une source de revenu direct pour les différentes structures qui les détenaient. Toutefois, depuis peu, cette logique vient de changer et les données sont devenues, au fur et à mesure, une matière première d'un nouveau paradigme économique alimentant de nouvelles approches de valorisation de ces données. Celles-ci, comme une sorte de levier, agissent indirectement sur l'ensemble de l'écosystème que cette libération des données a su initier. Ce n'est plus par la vente directe mais à travers, par exemple : différents services et applications qui renforcent la fidélisation des usagers ou des clients d'une marque ou d'une société ; à travers la volonté de transparence qui s'inscrit directement dans la logique de la communication et du marketing des acteurs concernés ; ou encore via le développement d'applications par la communauté qui s'ajoutent et enrichissent les services déjà proposés par les sociétés ou en créant de nouvelles issues des besoins et des demandes non identifiées à la base. Sans oublier que pour le secteur public, une telle libération peut être une source de revenu indirect issu des économies réalisées sur leurs dépenses par exemple. L'entrée timide de l'*open data* dans la sphère économique est pleine de promesses en sachant que l'histoire n'en est qu'à ses débuts.

3.4.1 Réflexion du la notion de retour sur investissement

Afin que l'aventure soit viable et rentable, au sens économique du terme, ce qui se traduit par des chiffres d'affaires dégageant des bénéfices attendus, il est nécessaire de concevoir un écosystème solide et raisonné qui ne se base pas uniquement sur une logique, certes admirable, du partage et du bénévolat ; en sachant que l'argument de la publicité, en tant que seul financeur des activités liées à l'ouverture des données, ne peut pas toujours servir d'unique vecteur de retour sur investissement. Sans oublier que pour rentrer dans ses frais, il faut également atteindre un certain seuil de couverture par rapport au nombre d'utilisateurs, ou autrement dit des clients, nécessaires pour qu'une entreprise soit viable comme en témoigne un de ces réutilisateurs : « *Accessoirement je suis réutilisateur de données sur Bordeaux (applis iBordeaux). J'ai participé à quelques ateliers Opendata… et j'ai essayé de démontrer qu'à l'échelle d'une agglomération ces données n'étaient pas monétisables… le nombre d'utilisateurs potentiels*

40. `http://data.bordeaux-metropole.fr/apicub`.
41. `http://docs.ckan.org/en/ckan-2.2/api.html`.
42. `https://www.opendatasoft.com/fr/2013/08/12/la-frontiere-api-interface-sefface-un-peu-plus`.

est très réduit et la donnée est librement consultable (horaires papiers, site de l'aot, fiches aux arrêts…) il faut vraiment que les AOT[43] *soient déconnectés de la réalité pour imaginer que leurs données ont de la valeur… c'est la qualité du service apporté à ses clients qui en a !* » (Chignard, 2012b). Ainsi, quoi qu'on en dise, pour faire des bénéfices, il faut un marché et que ce marché soit d'une taille critique, quelle que soit la perspective ou l'échelle.

D'ailleurs, cette situation dominée par divers quiproquos est assez bien résumée dans un autre commentaire, trouvé sur Internet, qui analyse cette situation si complexe de l'aspect économique des données ouvertes :

— *« Sans usager, sans client pas de service*
— *Sans producteur de service performant pas de service*
— *Sans données pérennes de qualité pas de service*
— *Sans producteur de données de qualité pas de service*
— *Sans chiffre d'affaire pas de producteur de données de qualité*
— *Sans chiffre d'affaire pas de producteur de services performants*
— *Sans publicité pas de chiffre d'affaire et donc…*

De nombreuses études ont démontré à plusieurs reprises :

— *qu'il existait un lien étroit entre le volume du marché de la publicité et le PIB*
— *qu'en proportion, historiquement la variation de la part de la publicité dans le PIB mondial évoluait très faiblement.*

Sans exclure quelques services de niche qui pourraient y parvenir durablement, espérer que la publicité pourra à elle seule financer les multiples services que nous utilisons tous, pourrait créer quelques désillusions ». (Chignard, 2012b)

Comme on peut le voir sur cet exemple, entre les attentes et la réalité du terrain, il y a parfois tout un fossé. C'est d'ailleurs pour cette raison qu'il faut être conscient que produire des « données » est une chose, les transformer en « données ouvertes », surtout si on garde en tête les 8 principes fondamentaux, en est une autre et en tirer des bénéfices encore une autre. Et cela concerne l'ensemble des acteurs, publics ou privés, aussi bien ceux qui décident, un jour ou l'autre, d'ouvrir leurs données aux autres, que ceux qui se mettent à construire sur cette ouverture un service viable et rentable ; si bien sûr, on met de côté, pour le moment, tout un pan d'activités basées sur le bénévolat et le partage. Tout a un coût et le retour sur investissement ne s'improvise pas et cela d'autant plus que « *faute de retours sur expérience à grande échelle ou d'études complètes et homogènes sur les sources potentielles de revenus, la question du ROI de l'Open Data reste encore ouverte* » (Econocom, 2012).

3.4.2 Réflexion sur le principe de redevance

Même si la logique de la redevance semble être en déclin (pour le moment) cela ne veut pas dire que l'idée même est complètement abandonnée. C'est plutôt le contraire et cela malgré la volonté de l'État de renforcer la gratuité de l'accès aux données ouvertes. Il ne faut pas perdre de vue que pour assurer le maintien d'un système basé sur l'ouverture des données, il faut également prévoir de quelle manière cette libération des données s'inscrit dans le plan économique de l'organisme qui a procédé à la mise à disposition de ces données. D'ailleurs, certains prévoient déjà, comme c'est le cas du Grand Lyon, de reconsidérer leur approche vis-à-vis du sujet en préparant, par exemple, plusieurs niveaux d'accès aux données qui va du gratuit

43. Pour en savoir plus sur les AOT, voir `http://www.transport-intelligent.net/acteurs-politiques-sti/autorites-organisatrices-des/`.

jusqu'au payant. Ainsi, par exemple, pour le Grand Lyon, « *la tarification de certaines données à forte valeur ajoutée est dans ce cas conçue comme un instrument de soutien à l'innovation et au développement économique au profit de structures et projets dont le modèle économique est complexe à stabiliser* » (Grand Lyon Data, 2015). Cette mesure « antimonopoliste » telle que le perçoit le Grand Lyon « *se traduirait par une redevance élevée dès le franchissement de seuils relatifs aux parts de marchés* » (Grand Lyon Data, 2015), ce qui veut dire que ce qui est principalement visé est la participation aux bénéfices réalisés par les réutilisateurs. Certes, c'est une des manières de concevoir un écosystème autour des données ouvertes. Toutefois, il faut également se demander si cela ne contredit pas en quelque sorte la politique même de l'ouverture des données. Sans oublier que la notion du « monopole » reste un peu étrange surtout s'il n'est pas rare de se retrouver tout seul sur le marché car personne d'autre ne s'est encore lancé. Comme on l'a déjà dit, le côté économique de la libération des données demeure, sans que cela soit un tabou, un sujet délicat et complexe en même temps, et qui ne cherche qu'à se définir et à trouver sa place dans l'écosystème plus général des services numériques.

Ainsi, afin que l'aventure soit viable et rentable au sens économique du terme, ce qui se traduit par des chiffres d'affaires dégageant des bénéfices attendus, il est nécessaire de concevoir un écosystème solide et raisonné qui ne se base pas uniquement sur une logique, certes admirable, du partage et du bénévolat ; en sachant que l'argument de la publicité, en tant que seule financeur des activités liées à l'ouverture des données, ne peut pas toujours servir d'unique vecteur de retour sur investissement.

4 Paysage actuel de l'*open data*

4.1 Rôles, acteurs, points de développement et d'évolution à surveiller

4.1.1 Structure de la chaîne de production/utilisation

Chaque acteur peut, en principe, être classé dans un des trois groupes majeurs qui sont constitués : des producteurs ; des redistributeurs ou autrement dit des passerelles et des réutilisateurs. Les frontières entre ces trois groupes ne sont pas toujours très étanches et ce classement représente davantage la position de départ que l'étendue des champs d'actions et les prises de positions vis-à-vis des données ouvertes par chacun de ces groupes. Ceci dit, quel que soit le positionnement des acteurs, ils sont réciproquement interdépendants. Aucun ne peut exister sans égard à l'autre. De ce fait, les redistributeurs et surtout les réutilisateurs sont largement dépendants de la nature et avant tout de la qualité des données mises à disposition et des services que cela implique. De l'autre côté, les producteurs ne peuvent pas ignorer les besoins et les possibilités des acteurs se trouvant dans les étapes suivantes de la chaîne de distribution des données ouvertes. Ils doivent être toujours à l'écoute de l'autre et adapter, le cas échéant, leurs manières de faire afin que leurs efforts ne servent pas uniquement à alimenter les statistiques sans un réel apport à l'ouverture efficace de ces données. Bref, c'est un réseau d'interdépendance où chacun a son rôle et les obligations qui lui incombent afin que l'édifice que les données ouvertes représentent ne devienne pas soit la tour de Babel soit un château de cartes.

Producteurs Les producteurs représentent l'ensemble des acteurs qui mettent à disposition des autres leurs propres données dont ils disposent selon les 8 principes fondamentaux qui régissent la transformation de ces données en données ouvertes. Il faut dire que c'est un processus non trivial car de nombreuses étapes accompagnent cette ouverture. Le cycle de traitement n'est pas anodin et se compose, en principe, des étapes suivantes :

- *sélection* : étape cruciale où on définit les domaines prioritaires et la nature des données à ouvrir ;
- *extraction* : on récupère les données de leur emplacement d'origine, ce qui passe parfois par leur numérisation ;
- *nettoyage* : on contrôle la validité et l'exhaustivité des données ;
- *transformation* : on choisit le(s) format(s) de publication en s'assurant de leur exploitabilité ;
- *publication* : les données doivent être mises à jour, bien documentées et avec les métadonnées compréhensibles et les plus complètes possible ; il faut également bien choisir la licence sous laquelle ces données sont rendues disponibles ;
- *réutilisation* : les producteurs doivent surtout s'assurer que leurs données ne posent pas de difficultés d'accès ou d'autres inconvénients qui peuvent gêner leur réutilisation.

Parmi les principaux producteurs, on trouve bien évidemment les différentes collectivités, locales ou régionales, quelques initiatives citoyennes, mais avant tout les différentes instances de l'État et diverses institutions publiques, sans oublier, bien évidemment, quelques sociétés privées et publiques comme, par exemple, JCDecaux, la RATP ou Keolis.

Distributeurs/Passerelles Les distributeurs ou les plateformes passerelles sont quant à eux une sorte d'intermédiaires se situant entre les producteurs et les réutilisateurs. Un très bon exemple de distributeur est data.gouv.fr ou encore data-publica.com. Le rôle de ces « passerelles » est très important car c'est grâce à ce type de structure qu'on peut trouver la plupart des données ouvertes mises à disposition un peu partout, en un seul endroit. Toutefois, le développement d'une telle plateforme n'est pas évident car il faut maîtriser l'ensemble des solutions développées et utilisées par d'autres acteurs, ce qui n'est pas toujours une chose simple. Etre un redistributeur, c'est se confronter à une réalité des faits telle qu'elle est présentée par les producteurs avec tous les aléas, imperfections et richesses d'applications que cela représente. Par exemple, là où le producteur peut choisir le format qui lui convient le mieux, le redistributeur doit savoir gérer l'ensemble des formats utilisés par tout un chacun. Sans oublier les différentes normes de présentation de ces données et même leur encodage. Il doit également mettre en place un système de veille permanente afin de s'assurer que les liens vers les données ouvertes qu'il recense chez lui sont toujours disponibles chez le producteur et qu'il est en mesure de prendre en compte tous les changements et les modifications effectués par les producteurs sur ces données.

Utilisateurs « finaux », réutilisateurs Les réutilisateurs constituent un troisième groupe qui referme le cycle de distribution des données ouvertes. Il peut s'agir d'un simple citoyen qui à travers les données mises à disposition dans le cadre de la politique de transparence et de participation veut avoir un droit de regard sur certains aspects de la gestion de la sphère publique. Il peut s'agir également de l'ensemble des acteurs qui inscrivent la réutilisation des données ouvertes dans leur volonté de se lancer dans une aventure entrepreneuriale pour créer des ri-

chesses au sens économique du terme. Bref, la réutilisation des données ouvertes est ouverte à tout le monde quel que soit le but recherché. Ceci dit, dans sa relation vis-à-vis des données elles-mêmes, le réutilisateur se trouve quelque part entre le producteur et le redistributeur. Son positionnement et, ce qui va avec, le lot de contraintes que cela implique et impose dépendent de la nature de ses besoins, des buts recherchés mais aussi de l'échelle de l'« industrialisation » des solutions qu'il développe ou propose. Par exemple, tant qu'il n'est confronté qu'à des jeux de données unitaires pour lesquelles il met en place des applications dédiées, il se trouve plutôt dans diverses problématiques auxquelles le producteur lui-même est confronté. Si, par contre, il se lance, par exemple, dans le processus d'agrégation et de croisement des différentes sources de données, il devra faire face à toute la panoplie des obstacles à maîtriser auxquels le redistributeur lui-même est constamment confronté. Sans oublier d'y ajouter quelques nouveaux comme, par exemple, la gestion de différentes contraintes imposées par les licences, ce qui n'est pas aussi évident que cela peut paraitre à première vue.

4.1.2 Les acteurs majeurs du secteur

L'importance d'un acteur peut dépendre de différents facteurs qui reflètent à la fois leur position dans le cycle de distribution des données ouvertes, les secteurs dont ils sont issus ou qu'ils couvrent, ou encore leur masse et l'étendue de leur champ d'action. D'autre part, la position dans le paysage de l'*open data* peut également dépendre de leur statut de précurseur ou toute autre forme de « soft power » issue de la primauté dans les solutions proposées, leur inventivité et même leur courage.

Dans ce sens, d'un statut à part peut jouir, par exemple, la Ville de Rennes qui a, la première, mis en place en France un portail avec des données ouvertes et s'inscrit de ce fait dans l'histoire de l'*open data* française comme pionnière incontestable.

Data Publica [44], par exemple, est remarquable par rapport au fait que c'est la plus grande passerelle issue du secteur privé - plus de 22000 jeux de données répertoriées, dont plus de 7000 mises à disposition en libre-service [45]. Cependant, il faut mentionner que la situation autour du portail de Data Publica devient dernièrement un peu confuse. Leur moteur de recherche ne fonctionne plus (et cela depuis plusieurs mois, la dernière vérification a été réalisée le 25 mai 2015) et le blog mis en place semble s'être arrêté depuis le 16 octobre 2014, date de la dernière actualité. Depuis, il semble que Data Publica (C-Radar) a réévalué sa stratégie et s'est concentré davantage sur la vente des services dans le cadre du système de marketing prédictif.

C'est aussi cela le revers du monde de l'*open data*, la très éphémère pérennité des solutions et des acteurs confrontés à la dure réalité du monde du numérique. Dans le secteur public, le vainqueur incontestable est le portail data.gouv.fr, à la fois en raison de sa taille et de sa position.

Secteur public Dans le secteur public il faut distinguer deux niveaux de base : national et local.

Niveau national : le rôle moteur d'Etalab et data.gouv.fr Suite au décret n° 2011-194 du 21 février 2011, l'État a mis en place la mission Etalab chargée de la « *création d'un portail*

44. Devenu entre temps C-Radar, `http://www.c-radar.com/`.
45. Chiffres invérifiables à l'heure actuelle.

unique interministériel destiné à rassembler et à mettre à disposition librement l'ensemble des informations publiques de l'État, de ses établissements publics administratifs et, si elles le souhaitent, des collectivités territoriales et des personnes de droit public ou de droit privé chargées d'une mission de service public » (Décret, 2011). Le portail en question a vu le jour en décembre 2011 et il a très rapidement recensé près de 350000 jeux de données (CADA, 2011) en provenance de différentes structures de la fonction publique mais également celles de l'Eurostat par exemple. En 2013, a eu lieu une refonte du site et le nombre de jeux de données a fondu jusqu'à afficher, au 11 mars 2015, 14089 jeux de données. C'est une chute spectaculaire mais le nombre actuel semble être plus raisonnable car reflétant davantage la réalité perceptible du terrain. Malgré tout, le portail data.gouv.fr demeure l'acteur majeur, pour ne pas dire principal, du monde de l'open data en France.

Cela étant, le travail sur l'amélioration du site ne semble pas être terminé car de nombreuses incohérences persistent encore. Bien que le travail soit en cours, on va quand même essayer de montrer, sur les quelques imperfections recensées, la nature des difficultés auxquelles sont confrontés les administrateurs de la plupart des sites du type passerelle.

À titre d'exemple, on peut essayer de se focaliser sur le volume des données libérées publiées sur le portail de data.gouv.fr et évaluer le « poids » des « producteurs » locaux en regardant le nombre de jeux de données par certains des acteurs en mai 2015. On peut voir que les trois premières places étaient monopolisées par Data Publica, Eurostat et Banque Mondiale, avec (respectivement) – 7170, 5908 et 1260 jeux de données. Sauf que, Data Publica n'est pas un producteur et les deux suivants ne sont pas des acteurs locaux. Ceux-là n'apparaissent qu'à partir de la quatrième position avec le portail mutualisé de Nantes, Loire-Atlantique et Pays de la Loire [46] en tête, suivi par la région Île-de-France [47] et celui de PACA [48]. Toutefois, il semble que les données présentées ne correspondent pas tout à fait à la réalité. Par exemple, le portail mutualisé de Nantes, Loire-Atlantique et Pays de la Loire dispose, selon data.gouv.fr, de 449 jeux de données. Sauf que sur les 449 jeux de données uniquement 11 sont disponibles. C'est la même chose pour les données de la région PACA où sur 411 déclarées, uniquement 36 jeux de données sont affichés sur le site de data.gouv.fr. D'autre part, ces données sont en décalage avec les valeurs affichées sur les sites respectifs des acteurs concernés (voir le tableau 4).

Acteurs	data.gouv.fr		Nombre de jeux de données sur les portails des acteurs
	Déclarées	Accessibles	
Nantes	449	11	582
Île-de-France	420	420	527
PACA	411	36	546

TAB. 4 – *Relevé effectué sur les sites concernés le 11 mars 2015.*

De plus, certains acteurs sont mentionnés plusieurs fois via les différentes dénominations. Ainsi, la Ville de Montpellier s'y trouve deux fois et cela une fois en tant que « Ville de Mont-

46. `http://data.nantes.fr/donnees/`.
47. `http://data.iledefrance.fr/explore/`.
48. `http://opendata.regionpaca.fr/donnees.html`.

pellier »[49] avec 90 jeux de données, puis en tant que « Montpellier Territoire Numérique »[50] avec 99 jeux de données. Sans oublier que le site de « Montpellier Méditerranée Métropole »[51] avec ses 50 jeux de données n'y figure même pas. La même chose peut être dite pour la région PACA qui y figure à la fois comme OPEN PACA[52] (411 jeux de données, 36 effectives) et « Région Provence-Alpes-Côte d'Azur »[53] (130 jeux de données).

Comme on peut le voir, il semble alors que les données du portail data.gouv.fr ne sont pas toujours à jour et qu'elles sont parfois manifestement erronées. Par ailleurs, cela pose aussi la question sur la quantité réelle des données présentes sur ce portail et impose une grande méfiance vis-à-vis des valeurs fournies. Il faut se rendre alors à chaque fois directement sur le site des producteurs des données pour en savoir plus. Mais cela ne donne pas toujours non plus les informations escomptées. Par exemple, si on prend les données ouvertes de l'Éducation Nationale[54], on trouve 337 jeux de données. Mais il n'y a aucun moyen de vérifier cette valeur sur le site même du ministère[55]. La seule chose qu'on trouve, ce sont les 25 jeux de données publiées sur le site du ministère de l'Enseignement supérieur et de la recherche[56] parmi lesquelles seulement 21 sont directement issues du ministère lui-même[57]. La seule explication qu'on peut avoir semble être le fait que le ministère de l'Éducation Nationale publie directement ses données sur le portail data.gouv.fr. Ces quelques exemples donnent un aperçu des difficultés que les portails du type data.gouv.fr doivent constamment affronter. Parmi elles :

— une veille constante des sources primaires ;
— le suivi de changement des dénominations des sites sources ;
— la gestion des doublons, etc.

Pour conclure, on peut dire que maintenir un portail « passerelle » est un travail titanesque et de longue haleine. Les obstacles sont nombreux et parfois même infranchissables en l'état actuel des choses. Malgré tout, ce genre de portails est une nécessité. Cela n'est peut-être pas toujours perceptible à l'heure actuelle où les données sont « consommées » principalement localement. Toutefois, cela va certainement changer dans un futur proche, avec l'augmentation du nombre des acteurs et du volume des données. Il sera alors de plus en plus difficile d'aller chercher les données à chaque fois chez les fournisseurs primaires et cela d'autant plus que la portée de ces données va grandement s'élargir et dépasser la sphère micro afin d'opérer au niveau macro, c'est-à-dire de passer du niveau local au national et même au-delà.

Niveau local : le dynamisme des régions, villes et collectivités Au niveau local la situation autour des données ouvertes est plutôt contrastée. Seule une infime partie des différentes collectivités, villes, départements ou régions a ouvert ses données. Comme on peut le voir sur la figure 2, une large partie du territoire français n'est pas couvert.

Ce que l'on peut regretter, c'est non seulement cette énorme disparité visible à l'œil nu mais également un certain manque de coopération parmi les acteurs et les partenaires locaux qui ont quand même franchi le pas et se sont lancés dans l'aventure de l'open data. On multiplie

49. https://www.data.gouv.fr/fr/organizations/ville-de-montpellier/.
50. https://www.data.gouv.fr/fr/organizations/montpellier-territoire-numerique/.
51. http://data.montpellier-agglo.com/
52. https://www.data.gouv.fr/fr/organizations/open-paca/.
53. https://www.data.gouv.fr/fr/organizations/region-provence-alpes-cote-d-azur/.
54. https://www.data.gouv.fr/fr/organizations/education-nationale/.
55. http://www.education.gouv.fr/.
56. http://data.enseignementsup-recherche.gouv.fr/explore/.
57. Ajoutons que sur le site de data.gouv.fr, le nombre de jeu de données s'élève à 29.

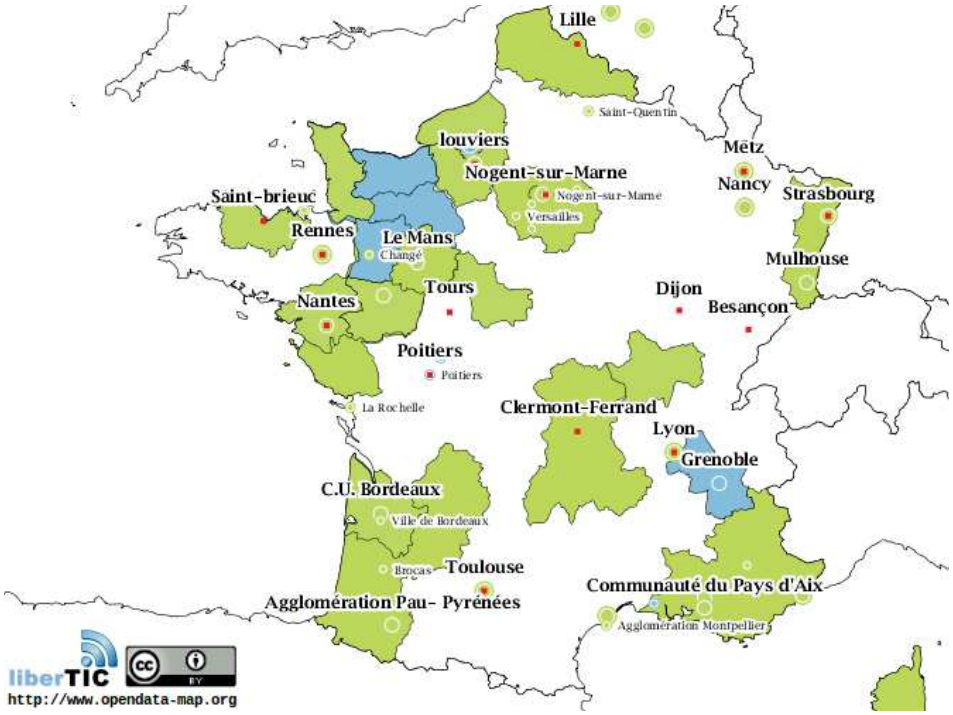


FIG. 2 – *Carte de l'open data en France. Source : LiberTIC, sous licence CC BY Attribution.*

les sites et les portails et cela dans la même région ou ville. Toutefois, on peut également se féliciter des initiatives pour mutualiser les forces comme l'ont fait par exemple Nantes, Loire-Atlantique et Pays de la Loire [58] qui ont lancé un portail commun pour leurs données ouvertes.

Secteur privé : les sociétés qui ont franchi le pas Dans le monde du privé, la situation est encore plus contrastée. À vrai dire, il n'y a pas vraiment beaucoup de sociétés qui ont franchi le pas. On peut bien évidemment citer JCDecaux, la RATP ou Keolis comme précurseurs, même si pour le moment bien isolés. Il semble que la plupart des sociétés n'ont pas encore trouvé soit les raisons soit les solutions satisfaisantes pour se lancer dans l'aventure de l'*open data*. Soit ils ne possèdent pas les données qu'ils pensent pouvoir ouvrir soit ils n'ont pas encore trouvé une manière d'intégrer une telle libération à leur politique industrielle et économique. On peut trouver une part d'explication sur le site de la RATP qui à la question « *Quels sont les jeux de données ouvertes ? Pourquoi ces données plutôt que d'autres ?* » répond ouvertement qu'« *il faut bien comprendre que nous devons préserver certaines données. Nous devons savoir trouver le bon équilibre entre de nouveaux services permis par l'open data et la nécessaire préservation de nos processus industriels* » (RATP, 2015). Détenir l'information, c'est

58. `http://data.nantes.fr/donnees/`.

détenir le pouvoir et comme la donnée est le constituant de cette information, il est tout à fait compréhensible que de nombreux acteurs ne soient pas encore prêts à la partager.

Union européenne : les projets poussent à l'ouverture L'Europe, celle de l'Union européenne ou via d'autres coopérations régionales et paneuropéennes, participe, pour sa part, à l'ouverture des données de plusieurs manières. Elle est à la fois consolidatrice et harmonisatrice des législations et des pratiques à l'échelle européenne et également instigatrice des différentes initiatives et projets liés à l'*open data*. Dans ce sens, il faut avant tout mentionner le projet qui vise l'ouverture des données à l'échelle européenne, c'est-à-dire celui de la mise en place du Portail des données ouvertes de l'Union européenne lancé en 2013.

On peut également évoquer, par exemple, le projet Homer [59] qui dans le cadre du partage à l'échelle européenne s'est donné pour but « *harmonising the Open data in the Mediterranean through better access and Reuse of public sector information* ». C'est un projet européen dont l'objectif est d'harmoniser et de favoriser l'ouverture des données dans les pays du pourtour méditerranéen de : l'Espagne, la France, l'Italie, Malte, la Grèce, la Slovénie, Chypre et le Monténégro.

4.1.3 Un secteur encore en devenir

Analyse d'une partie de l'offre de données ouvertes Parler des « faiblesses » des solutions mises en place dans le cadre de l'*open data*, ou de l'*open data* tout court n'est pas une chose aisée car soit il faut prendre chaque portail un par un, soit il faut se hisser à un certain niveau de généralité en risquant d'établir une image globale mais réductrice qui peut ne pas refléter la richesse de la réalité des pratiques et des solutions déployées. Cependant, devant l'immensité de la tâche, on prend le risque de se restreindre, malgré tout, à un aperçu plutôt général sur l'état des faits dans son ensemble en essayant de pointer quelques principales difficultés auxquelles sont confrontés les acteurs de tous bords.

Les principes de l'ouverture des données mis à mal Commençons alors par ce petit extrait de licence destiné aux données ouvertes et mises en place par la SNCF (c'est nous qui soulignons) : « *De manière générale, la Licence de la Base de données est accordée « **telle quelle** » par SNCF, **sans aucune garantie** de quelque type que ce soit, qu'elle soit expresse, tacite ou qu'elle découle de la loi, de la coutume ou de l'usage commercial. **SNCF est exonérée** en particulier **de toute responsabilité** au titre de la condition de propriété ou de toute garantie tacite, de l'absence de violation, **de l'exactitude ou de l'exhaustivité, de la présence ou de l'absence d'erreurs**, de l'adéquation à une utilisation particulière, **de la qualité marchande** ou autre ou de la **discontinuité, la suspension ou l'interruption** temporaire ou définitive de mise à disposition des Contenus* » (SNCF, 2015). On peut dire que cet extrait reflète, d'une manière ou d'une autre, tout un ensemble, non exhaustif, de différents « maux » qui accompagnent, à la fois la politique et les solutions mises en place dans le cadre de l'*open data* en ses débuts. Mais reprenons cet extrait point par point, sans s'attacher spécialement à l'ordre exact.

Avant tout, on remarque que la licence, qui est accordée « *telle quelle* », ne porte sur aucune garantie « *de quelque type que ce soit* ». C'est très important d'être conscient de ce détail qu'on retrouve un peu partout dans le monde de l'*open data*, et pas seulement d'ailleurs. La

59. `http://homerproject.eu/`.

libération des données est un acte, surtout dans le privé (mais le secteur public est également concerné), d'une volonté assumée de participer à ce mouvement de libération des données. C'est déjà en soi un signal très positif et on ne peut qu'espérer qu'il en soit tout le temps ainsi. Toutefois, certains ne se positionnent pas toujours, dans une logique de gagnant-gagnant mais uniquement, et encore, dans celle de donnant-donnant, voire donnant tout court. Cela sous-entend très souvent que l'acteur qui a procédé à la libération des données, fait un amalgame entre « éviter tout risque » et « éviter toute responsabilité » (même morale). D'ailleurs, le fait de se considérer « *exonérée [. . .] de toute responsabilité* » ne fait que renforcer davantage cette malheureuse prise de position. Il faut dire que c'est une posture très dommageable pour l'ensemble de l'écosystème que le mouvement des données ouvertes tente de construire car cela empêche de bâtir une structure plus solide en laissant libre cours à l'aléatoire et à l'incertitude.

Du point précédent découle un autre, celui de la garantie de la fiabilité des données mises à disposition. À la lecture de cet extrait, on se rend très vite compte qu'il n'y a ici aucune garantie que les données soient fiables, exactes et sans erreurs. En fait, les données sont mises à disposition « telles quelles » et on ne peut que supposer, voire espérer, qu'elles correspondent, d'une manière ou d'une autre, à une quelconque réalité.

De plus, le risque de « *la discontinuité, la suspension ou l'interruption temporaire ou définitive* » qui plane toujours sur les données n'arrange pas du tout la situation. La pérennité des données et ce qui va avec le service proposé autour de ces données est de ce fait gravement remis en question.

Ce que tout cela soulève est surtout le questionnement sur la valeur et la pérennité des données, et ce qui va avec, des services autour d'elles et en fin de compte de l'écosystème tout entier. Car comment dans ces conditions s'assurer de la viabilité d'un quelconque projet basé sur les données ouvertes si presque l'ensemble des 8 principes fondamentaux est largement mis entre parenthèses et cela sur quelques lignes d'un extrait d'une licence qui devait justement participer à solidifier les piliers de l'édifice de l'*open data* ? La « *qualité marchande* » des données ouvertes qui ne cessent de baigner dans un flou très varié reste de ce fait à établir. Il faut effectivement se demander si un quelconque écosystème est réellement viable à terme si celui-ci n'est couvert par aucune garantie « *de quelque type que ce soit* » et si tout se base uniquement sur quelques principes généraux, quelques préconisations d'usage, quelques bonnes volontés ou des gestes gracieux. Bien évidemment, les données issues de la fonction publique n'ont pas tout à fait le même statut, surtout par rapport à leur possibilité d'ouverture, que celles du secteur privé. Toutefois, même ici on peut voir très souvent à quel point cette situation, sous certains aspects proche de l'anarchie, pèse sur la qualité de cette « marchandise », que sont les données ouvertes, et de tout l'écosystème qui est en train de se construire autour. Par exemple, même le portail data.gouv.fr qui est censé être une vitrine exemplaire n'est pas épargné car celui-ci, « *comme d'autres plateformes d'ailleurs, souffre du tronquage et de la mauvaise qualité de certains jeux de données : des budgets publiés en PDF, des agrégations rendant l'interprétation impossible, des liens html présentés dans la liste de données ouvertes, des tableaux et trombinoscope en.doc, des données en vrac dans les fichiers. Naviguer sur data.gouv.fr est actuellement une expérience culinaire : telle une boîte de chocolat, on ne sait jamais à quoi s'attendre* » (LiberTIC, 2012). Ainsi, comme le souligne Claire Gallon par rapport au projet data.gouv.fr, celui-ci « *semble s'être focalisé sur la quantité au détriment de la qualité* » (LiberTIC, 2012). Toutefois, comme cela a déjà été dit à plusieurs reprises, la qualité des données et son suivi sont primordiaux. Il ne suffit pas de mettre les données à

disposition mais il faut également s'assurer que les données restent exploitables et cela d'une manière pérenne.

Sur le terrain : sondage, observations, échantillon Ainsi, pour voir à quel point cette exploitabilité « pérenne » des données est vraiment assurée, le mieux est de le vérifier directement. Pour ce faire, on a passé au crible les quelques principaux portails qui mettent à disposition les liens vers les données ouvertes issues de différentes sources aussi bien publiques que privées. Les portails analysés sont les suivants :

— Data.gouv.fr : https ://www.data.gouv.fr/fr/ ;
— Data Publica : http ://www.data-publica.com ;
— Datahub : https ://datahub.io/ ;
— European Union Open Data Portal (EUODP) : devenu http ://www.europeandataportal.eu/ ;
— Search Europe's Public Data (SEPD) : http ://publicdata.eu/ ;
— Enigma : http ://enigma.io/.

Chacun de ces portails a pour but de récupérer les données ouvertes accessibles auprès des fournisseurs primaires et de les mettre à disposition via une interface unifiée. On les a soumis à des tests à l'aide de requêtes en français et en anglais. La requête concerne les « accidents de la route » ou encore « road accidents ». On a passé au crible les 10 premiers résultats issus de chaque requête. Ce qui est visé par ce test est de voir à quel point ces portails s'acquittent de leurs rôles et fournissent les liens opérationnels vers les jeux de données qu'ils indiquent.

On peut d'abord dresser un récapitulatif dans le tableau 5 et le tableau 6.

Portail	Nombre de résultats FR	Nombre de liens dans les 10 premiers résultats		
		Opérationnels directement	Opérationnels indirectement	Non opérationnels
Data.gouv.fr	174	7	1	2
Data Publica	58	3	1	6
Datahub	3			3
EUODP	9	5	4	
SEPD	8			8
Enigma	0			

TAB. 5 – *Test de requête, résultats en français.*

Eléments d'analyse Le premier constat qu'on peut avancer est qu'aucun de ces portails fédérateurs ne donne entière satisfaction, et cela pour plusieurs raisons. Soit le nombre de liens non-opérationnels est beaucoup trop élevé, soit les résultats eux-mêmes sont peu nombreux. L'exception notable est Data.gouv.fr avec le plus grand nombre de résultats affichés et European Union Open Data Portal (EUODP) avec aucun lien non-opérationnel. Toutefois, là où Data.gouv.fr donne plus de 100 résultats pour chaque requête, European Union Open Data Portal n'en donne que 9 pour la requête en français et 17 pour la requête en anglais, sachant que les 9 résultats pointent en réalité vers des données en anglais. D'ailleurs, on peut faire le même reproche à Data.gouv.fr qui sort 127 résultats pour la requête en anglais mais qui concernent

Portail	Nombre de résultats EN	Nombre de liens dans les 10 premiers résultats		
		Opérationnels directement	Opérationnels indirectement	Non opérationnels
Data.gouv.fr	127	4	2	4
Data Publica	3	2		1
Datahub	25	4		6
EUODP	17	7	3	
SEPD	103	3	5	2
Enigma	28	10		

TAB. 6 – *Test de requête, résultats en anglais.*

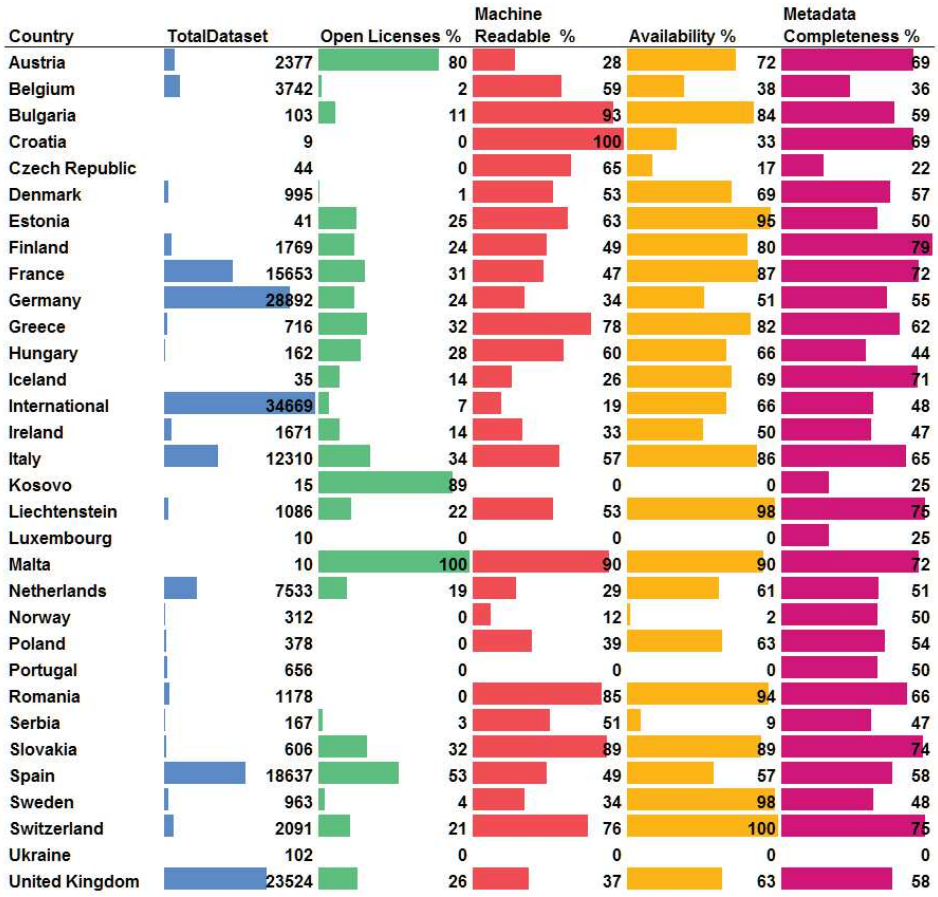
tous des données françaises. Cependant, là où European Union Open Data Portal prétend être le portail Européen (sous-entendu multilingue et avec un accès aux données « multinationales »), Data.gouv.fr est clairement affiché comme portail français et en tant que tel, il ne se préoccupe que des données françaises. Les deux autres portails qui affichent un nombre de résultats relativement élevés sont Data Publica et Search Europe's Public Data (SEPD). Toutefois, Data Publica affiche un score très élevé de liens non-opérationnels pour la requête en français et peu de résultats pour celle en anglais. Search Europe's Public Data, de son côté, donne beaucoup de résultats pour la requête en anglais mais aucun lien opérationnel pour le français. Le test nous montre alors une réalité plus que contrastée sur ce sujet.

De plus, il semble que les choses n'ont pas changé tant que ça depuis la réalisation de cette étude, c'est-à-dire depuis la fin de l'année 2014. Si on regarde à présent les chiffres qui suivent, on s'aperçoit très vite que certaines difficultés persistent. Par exemple, sur un nombre total de 173 catalogues le pourcentage des données lisibles par une machine s'élève à 45, l'accessibilité à 65 et la complétude à 58.

La figure 3 nous donne plus de détails sur ces chiffres daté du 24 mai 2016 et fournis par Open Data Monitor [60].

D'autres illustrations de déconvenues pratiques Si on reste encore un peu dans l'analyse des réalités des pratiques, en quittant les « passerelles » et en allant vers les « producteurs », on peut également mentionner, à titre d'exemple, le cas de la société SEMITAN, qui gère le réseau de transports à Nantes, dont la qualité des données a été relatée, en son temps, sur un blog (Deldicque, 2012). Celui-ci se plaint, par exemple, du fait que le fichier contenant des horaires s'étale sur 3 700 000 lignes ce qui constitue déjà en soi une difficulté technique non triviale pouvant impacter sérieusement les fonctionnalités des applications se basant sur ce genre de données. Une autre difficulté relatée sur le même blog est celle de la « fraîcheur » des données. Dans l'exemple évoqué « *les horaires d'été des transports en commun de Nantes et Bordeaux n'ont pas été communiqués aux développeurs à temps et les utilisateurs ont trouvé service-clos au lancement des applications* », et même si la coupure ne représentait qu'environ 12 heures, cela peut s'avérer fatal aussi bien pour les applications qui fournissent le service que pour la confiance envers l'écosystème que les « données ouvertes » constituent. Bien évidemment,

60. `http://opendatamonitor.eu`.



Country	TotalDataset	Open Licenses %	Machine Readable %	Availability %	Metadata Completeness %
Austria	2377	80	28	72	69
Belgium	3742	2	59	38	36
Bulgaria	103	11	93	84	59
Croatia	9	0	100	33	69
Czech Republic	44	0	65	17	22
Denmark	995	1	53	69	57
Estonia	41	25	63	95	50
Finland	1769	24	49	80	79
France	15653	31	47	87	72
Germany	28892	24	34	51	55
Greece	716	32	78	82	62
Hungary	162	28	60	66	44
Iceland	35	14	26	69	71
International	34669	7	19	66	48
Ireland	1671	14	33	50	47
Italy	12310	34	57	86	65
Kosovo	15	89	0	0	25
Liechtenstein	1086	22	53	98	75
Luxembourg	10	0	0	0	25
Malta	10	100	90	90	72
Netherlands	7533	19	29	61	51
Norway	312	0	12	2	50
Poland	378	0	39	63	54
Portugal	656	0	0	0	50
Romania	1178	0	85	94	66
Serbia	167	3	51	9	47
Slovakia	606	32	89	89	74
Spain	18637	53	49	57	58
Sweden	963	4	34	98	48
Switzerland	2091	21	76	100	75
Ukraine	102	0	0	0	0
United Kingdom	23524	26	37	63	58

FIG. 3 – *Etat des ressources en open data par pays en mai 2016. Source Open Data Monitor, sous licence CC Attribution 4.0.*

personne n'est à l'abri des dysfonctionnements mais il faut s'assurer que ceux-ci soient limités au maximum et la bonne gestion des mises à jour des données en fait partie.

Sans oublier qu'il faut également prévoir et faire face aux différentes malveillances comme par exemple le piratage des données comme cela était le cas avec le site de Keolis dont le gestionnaire a été contraint d'avouer que le site avait été compromis : « *Certaines parties du contenu du site www.star.fr ont été récupérées par un utilisateur malveillant* » (Ouest-France, 2013). Les obstacles sont donc nombreux et les défis immenses.

Mise en évidence de quelques points à surveiller Pour résumer on peut tenter de présenter quelques points qui posent ou qui peuvent poser certaines difficultés.

Mosaïque de formats La (trop) grande variété des formats de distribution des données ouvertes : la nature des données ouvertes, malgré les 8 principes fondamentaux, reflète l'ensemble de la réalité des pratiques et des besoins. Cela provoque que ces données ne sont pas uniquement au format tabulaire, comme on pouvait s'y attendre, mais également textuel, sonore, graphique, etc. Ainsi, il est compréhensible que le seul format CSV ne puisse pas s'appliquer automatiquement à tout le spectre des données. Ceci-dit, un effort peut être fait pour restreindre au maximum l'utilisation des formats jugés peu adaptés pour le monde de l'*open data*, comme c'est le cas, par exemple, de xls ou pdf, afin de simplifier l'usage et l'interopérabilité.

Qualité relative des données La qualité des données est un point difficile à aborder car que veut dire la « qualité » des données ? En principe, on s'attend à des données complètes, sans erreurs, bien documentées et avec les métadonnées les plus exhaustives possibles d'un côté et s'appuyant sur certains des référentiels existants de l'autre. Malheureusement, cela n'est pas toujours le cas. Un travail en amont semble être alors nécessaire.

Eclatement des moyens de distribution des données La distribution des données est loin d'être uniforme : malgré la présence de certaines solutions spécialisées pour la mise à disposition des données libres comme celle de l'OpenDataSoft (payante) ou CKAN (gratuite, utilisée par l'Etalab pour le data.gouv.fr) le paysage demeure très varié. De plus, même si un réel effort est fait pour mettre en place, par exemple, des API, l'accès à des données reste assez fastidieux. La distribution des données est encore très dépendante des différentes interfaces et ne se présente pas toujours comme une solution à part, indépendante et permettant l'accès direct aux catalogues sans l'intermédiaire d'une interface tournée vers l'homme. Ce qu'il faut résoudre, pour pouvoir automatiser au maximum le traitement des données, c'est de trouver des solutions du type machine-machine et non machine-homme qui sont déjà en partie existantes mais pas complètement adaptées aux besoins du monde de l'*open data*. L'alpha et oméga du monde de l'*open data* est d'avoir un accès direct et inaltéré aux sources.

Pérennité fragile des données L'éphémère pérennité des données : la qualité des données se mesure également par leur maintien à jour (pour celles où c'est nécessaire ; et il faut dire que très souvent la seule mise à jour des données est celle de leur mise à disposition, ce qui est largement insuffisant) et leur maintien tout court. Il n'est pas rare, pour ne pas dire très courant, que les données disparaissent tout simplement après un certain temps ou deviennent inaccessibles pour une raison quelconque. Cette insécurité quant à la pérennité des données est donc très préjudiciable pour l'ensemble de l'écosystème que les données ouvertes constituent. Cela implique, de ce fait, qu'une réflexion soit menée pour trouver une solution afin que l'accessibilité des données mises à disposition soit à la fois maintenue dans le temps et les données tenues à jour. Une des solutions envisageables (même si complexe, on l'avoue) peut être, par exemple, le développement d'un grand dépositoire (obligatoire le cas échéant) dans lequel toutes les données ouvertes seraient déposées. Bien évidemment, cela concerne principalement les données statiques et dynamiques car les données fluides demandent un traitement spécial. De plus, la question de financement d'une telle solution impose une réflexion sur l'étendue et la nature de la « gratuité » du service.

Manque de standards Le manque de standards propres au monde des données ouvertes : il faut dire que pour le moment il n'y a pas beaucoup d'éléments qui encadrent d'une manière sans équivoque le monde de l'*open dat*a. Certes, il y a les fameux 8 principes fondamentaux et ses 7 principes auxiliaires, sans oublier les 72 bonnes pratiques, mais il s'agit avant tout de principes et de préconisations d'ordre général qui n'imposent rien. Bien évidemment, il est tout à fait compréhensible que les débuts de l'*open data* soient aussi peu encadrés et encore moins restrictifs. L'*open data* est un phénomène nouveau qui doit d'abord trouver sa place dans le paysage numérique. Une manière de faire est justement de ne pas en faire trop, au moins au début, et d'attendre de voir comment l'environnement de l'*open data* évolue et s'auto-définit lui-même. Alors, il est peut-être temps de passer d'une idéalisation de la nécessité à un encadrement plus fouillé des pratiques du terrain en élaborant des règles plus strictes.

Couverture territoriale inégale La granularité territoriale de la mise à disposition des données ouvertes : cet aspect est particulièrement visible sur la carte de couverture territoriale des solutions open data mises en place. De grandes zones blanches montrent la frilosité des différents acteurs vis-à-vis du mouvement de l'ouverture. Ces carences sur l'axe horizontal de distribution des données posent de grands problèmes surtout si on envisage de sortir de la logique locale de l'utilisation des données ouvertes et de développer des services à une échelle plus grande. Mais si on regarde la situation sur l'axe vertical, c'est-à-dire selon la profondeur structurale de propagation de la logique et des solutions d'*open data*, on s'aperçoit que l'image qui en ressort n'est pas uniquement extrêmement contrastée mais également très complexe. Tout d'abord, toutes les collectivités ne sont pas obligées de publier leurs données. De plus, comment décider de la qualité et de l'utilité de telle ou telle donnée que quelqu'un se décide à publier ? Sans oublier que pour certaines données, il n'est même pas établi qui doit les libérer. Par exemple, une communauté met à disposition un jeu de données sur les arrêts de bus dans son village. Toutefois, est-ce à la communauté de le faire ou à la société qui gère les transports communs dans la communauté ? En plus, toutes ces informations, dispersées un peu partout, remettent en cause la viabilité économique même, car trop localisée la plupart du temps, d'une quelconque initiative entrepreneuriale. La propagation de l'ouverture des données doit alors être systématisée et d'une couverture suffisamment large et profonde pour qu'on puisse parler réellement d'une quelconque valeur et même utilité des données libérées.

5 Conclusion

Même si l'histoire de l'*open data* n'en est qu'à ses débuts, on peut d'ores et déjà affirmer que les données ouvertes ont su établir une place très importante dans la réalité de notre quotidien et que leur poids ne va que s'accentuer. On ne peut que difficilement prévoir à quoi ressemblera le monde de demain régi par les données, mais il est incontestable que le monde d'aujourd'hui est et continuera d'être bouleversé par son entrée dans l'ère du numérique et des données. Bien sûr, certains peuvent dire que l'ouverture des données publiques doit être une évidence en soi et non une sorte de « cadeau » au bon peuple. Il faut toutefois admettre que cette ouverture à laquelle on assiste n'est pas un geste politique de plus mais le signe d'un changement majeur à la fois politique, sociétal et économique. C'est une (r)évolution qui, attendue par les uns, crainte par les autres, nous propulse dans une autre dimension, dans laquelle rien ne sera comme avant et cela aussi bien pour le meilleur que pour le pire.

Références

Abiteboul, S. (2012). Sciences des données : de la logique du premier ordre à la Toile. In *Sciences des données : de la logique du premier ordre à la Toile : Leçon inaugurale prononcée le jeudi 8 mars 2012* (Collège de France, Fayard ed.)., Leçons inaugurales.

Adobe (1993). PDF Reference and Adobe Extensions to the PDF Specification. http ://www.adobe.com/devnet/pdf/pdf_reference.html. Adobe Developer Connection. Consulté le 5 mai 2015.

Algan, Y., T. Cazenave, et E. Macron (2016). *L'État en mode start-up* (Eyrolles ed.).

Assemblée Nationale (2014). Agenda et comptes rendus des réunions du Bureau. Réunion du mercredi 12 novembre 2014. http ://www.assemblee-nationale.fr/14/agendas/cr-bureau.asp#20141112. Consulté le 5 mai 2015.

Berners-Lee, T. (2010). TED 2010. The year open data went worldwide. http ://www.ted.com/talks/tim_berners_lee_the_year_open_data_went_worldwide. Consulté le 5 mai 2015.

Berro, A., I. Megdiche-Bousarsar, et O. Teste (2014). Transformer les Open Data brutes en graphes enrichis en vue d'une intégration dans les systèmes OLAP. https ://hal.archives-ouvertes.fr/hal-01110098/document. Consulté le 5 mai 2015.

Berthelot, C. (2013). Étude d'opportunité sur l'ouverture des données publiques de la région bretagne. Etude d'opportunité, Conseil régional de Bretagne. Consulté le 5 mai 2015.

BOAI (2010). Budapest Open Access Initiative. http ://www.budapestopenaccessinitiative.org/. Consulté le 5 mai 2015.

Bouchoux, C. (2014). Refonder le droit à l'information publique à l'heure du numérique : un enjeu citoyen, une opportunité stratégique (rapport). rapport d'information n° 589 (2013-2014) de Mme Corinne Bouchoux, fait au nom de la MCI sur l'accès aux documents administratifs, déposé le 5 juin 2014. Technical Report 589, Sénat.

Brine R. K. et M. Poovey (2013). From Measuring Desire to Quantifying Expectations : A Late Nineteenth Century Effort to Marry Economic Theory and Data. In L. Gitelman (Ed.), *"Raw data" is an oxymoron*, Infrastructures series, pp. 61. The MIT Press.

CADA (2011). Lancement du portail data.gouv.fr. http ://www.cada.fr/lancement-du-portail-data-gouv-fr,20114477.html. Consulté le 5 mai 2015.

CE (1989). *Lignes directrices pour améliorer la synergie entre secteur public et secteur privé sur le marché de l'information*. Commission Européenne.

Chignard, S. (2012a). En finir avec le mythe de la donnée brute. http ://donneesouvertes.info/2012/06/01/en-finir-avec-le-mythe-de-la-donnee-brute/. données ouvertes. Consulté le 5 mai 2015.

Chignard, S. (2012b). Monétiser les données du transport public... chiche ? http ://donneesouvertes.info/2012/11/15/monetiser-les-donnees-du-transport-public-chiche. données ouvertes. Consulté le 5 mai 2015.

Chignard, S. (2012c). *Open data : comprendre l'ouverture des données publiques* (Éditions FYP ed.). Collection Entreprendre.

CNN (2012). Avis n° 12 du conseil national du numérique relatif à l'ouverture des données

publiques (« Open data »). http ://www.cnnumerique.fr/wp-content/uploads/2012/06/2012-06-05_AvisCNNum_12_OpenData.pdf. Consulté le 5 mai 2015.

CPI (2015). Code de la propriété intellectuelle. http ://www.legifrance.gouv.fr/affichCode.do ?cidTexte=LEGITEXT000006069414. Consulté le 5 mai 2015.

Creative Commons (2015). CC0 1.0 Universel (CC0 1.0). Transfert dans le domaine public. https ://creativecommons.org/publicdomain/zero/1.0/legalcode.fr. Consulté le 5 mai 2015.

Déclaration (1789). Déclaration du 26 août 1789 des droits de l'homme et du citoyen. Version consolidée au 05 mai 2015. http ://www.legifrance.gouv.fr/affichTexte.do ?cidTexte=LEGITEXT000006071192&dateTexte=29990101. Consulté le 5 mai 2015.

Décret (2005). Décret n° 2005-1755 du 30 décembre 2005 relatif à la liberté d'accès aux documents administratifs et à la réutilisation des informations publiques, pris pour l'application de la loi n° 78-753 du 17 juillet 1978. http ://www.legifrance.gouv.fr/affichTexte.do ?cidTexte=JORFTEXT000000265304. Consulté le 5 mai 2015.

Décret (2011). Décret n° 2011-194 du 21 février 2011 portant création d'une mission « Etalab »chargée de la création d'un portail unique interministériel des données publiques. http ://legifrance.gouv.fr/eli/decret/2011/2/21/PRMX1105072D/jo/texte. Consulté le 5 mai 2015.

Décret (2014). Décret n° 2014-1050 du 16 septembre 2014 instituant un administrateur général des données. http ://legifrance.gouv.fr/eli/decret/2014/9/16/PRMX1421510D/jo/texte. Consulté le 5 mai 2015.

Deldicque, B. (2012). Bilan OpenData, collectivités, entreprises et développeurs. http ://www.benoit-deldicque.com/blog/bilan-opendata-collectivites-entreprises-et-developpeurs/. Consulté le 5 mai 2015.

Directive (2003). Directive 2003/98/CE du Parlement européen et du Conseil du 17 novembre 2003 concernant la réutilisation des informations du secteur public. http ://eur-lex.europa.eu/LexUriServ/LexUriServ.do ?uri=OJ :L :2003 :345 :0090 :0096 :FR :PDF. Consulté le 5 mai 2015.

Directive (2007). Directive 2007/2/CE du Parlement européen et du Conseil du 14 mars 2007 établissant une infrastructure d'information géographique dans la Communauté européenne (INSPIRE). http ://eur-lex.europa.eu/legal-content/FR/TXT/HTML/ ?uri=CELEX :32007L0002&from=EN. Consulté le 5 mai 2015.

Directive (2013). Directive 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. http ://eur-lex.europa.eu/legal-content/FR/TXT/HTML/ ?uri=CELEX :32013L0037. Consulté le 5 mai 2015.

Directive INSPIRE (2015). Conseil national de l'information géographique (CNIG). http ://cnig.gouv.fr/ ?page ?_id=1177. Consulté le 5 mai 2015.

Econocom (2012). Peut-on parler de ROI pour des données gratuites ? http ://blog.econocom.com/blog/peut-on-parler-de-roi-pour-des-donnees-gratuites/. E-media, the Econocom blog. Consulté le 5 mai 2015.

Edwards, P. N., M. S. Mayernik, A. L. Batcheller, G. C. Bowker, et C. L. Borgman (2011).

Science friction : Data, metadata, and collaboration. *Social Studies of Science 41*(5), 667–690.

Embley, D. W., M. Hurst, D. Lopresti, et G. Nagy (2006). Table-processing paradigms : a research survey. *International Journal of Document Analysis and Recognition (IJDAR) 8*(2-3), 66–86.

Etalab (2011). Licence Ouverte / Open Licence. https ://www.etalab.gouv.fr/licence-ouverte-open-licence. Le blog de la mission Etalab. Consulté le 5 mai 2015.

Etalab (2013). Publication de la Directive du 26 juin 2013 révisant la Directive PSI. https ://www.etalab.gouv.fr/publicationdeladirectivedu26juin2013revisantladirectivepsi. Consulté le 5 mai 2015.

Fauré, C. (2012). APICulture et DataCulture à la lumière du facteur temps. http ://www.christian-faure.net/2012/12/19/openculture-et-dataculture-a-la-lumiere-du-facteur-temps/. Hypomnemata : supports de mémoire. Consulté le 5 mai 2015.

Fayet, S. (2013). "Données" de la recherche, les mal-nommées. http ://urfistinfo.hypotheses.org/2581. UrfistInfo. Consulté le 5 mai 2015.

G8 (2013). Charte du G8 pour l'Ouverture des Données Publiques. http ://www.modernisation.gouv.fr/sites/default/files/fichiers-attaches/charte-g8-ouverture-donnees-publiques-fr.pdf. Consulté le 5 mai 2015.

GART (2012). Pour une redevance liée à l'usage des données des transports publics. Groupement des Autorités Responsables de Transport. http ://www.gart.org/S-informer/Salle-de-presse/Pour-une-redevance-liee-a-l-usage-des-donnees-des-transports-publics. Site du GART. Consulté le 5 mai 2015.

Gaudrat, P. (2006). Intérêts de l'investisseur contre droit d'auteur. http ://www.liberation.fr/tribune/2006/05/04/interets-de-l-investisseur-contre-droit-d-auteur_38159. Libération Tribunes. Consulté le 5 mai 2015.

Gaudrat, P. et G. Massé (2000). La titularité des droits sur les œuvres réalisées dans les liens d'un engagement de création. http ://miage.univ-nantes.fr/miage/DVD-MIAGEv2/Droit_files/Rapport%20Gaudrat%20Droit%20d%27auteur.pdf. Consulté le 5 mai 2015.

Grand Lyon Data (2015). Comprendre la démarche. http ://data.grandlyon.com/comprendre-la-demarche/. Consulté le 5 mai 2015.

Hervé, N. (2014). JCDecaux expulsé de Grenoble : une décision salutaire et subversive. http ://www.liberation.fr/debats/2014/11/24/jcdecaux-expulse-de-grenoble-une-decision-salutaire-et-subversive_1149817. Libération Idées. Consulté le 5 mai 2015.

Lacombe, R., P.-H. Bertin, F. Vauglin, et A. Vieillefosse (2011). Pour une politique ambitieuse des données publiques. http ://www.ladocumentationfrancaise.fr/rapports-publics/114000407/. Rapport public. Ministère de l'industrie, de l'énergie et de l'économie numérique. Consulté le 5 mai 2015.

Larousse (2015). Donnée. http ://www.larousse.fr/dictionnaires/francais/donnée/26436. Dictionnaire de français Larousse. Consulté le 5 mai 2015.

LiberTIC (2011). Pourquoi n'y a-t-il pas de consensus sur une licence Open Data en France ? https ://libertic.wordpress.com/2011/07/05/pourquoi-ny-a-t-il-pas-de-consensus-

sur-une-licence-open-data-en-france/. Consulté le 5 mai 2015.

LiberTIC (2012). « L'open data est très mal estimé ». Interview de Claire Gallon de l'association Libertic. http ://www.data-publica.com/content/2012/11/lopen-data-est-tres-mal-estime-interview-de-claire-gallon-de-lassociation-libertic/. Consulté le 5 mai 2015.

Loi (1794). Loi du 7 messidor an II (Loi du 25 juin 1794) concernant l'organisation des archives établies auprès de la représentation nationale. http ://www.legilux.public.lu/rgl/1794/A/0002/Z.pdf. Consulté le 5 mai 2015.

Loi (1978a). Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. http ://www.legifrance.gouv.fr/affichTexte.do ?cidTexte=JORFTEXT000000886460. Consulté le 5 mai 2015.

Loi (1978b). Loi n° 78-753 du 17 juillet 1978 portant diverses mesures d'amélioration des relations entre l'administration et le public et diverses dispositions d'ordre administratif, social et fiscal. http ://www.legifrance.gouv.fr/affichTexte.do ?cidTexte=JORFTEXT000000339241. Consulté le 5 mai 2015.

Loi (2004). Loi n° 2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique. http ://www.legifrance.gouv.fr/affichTexte.do ?cidTexte=JORFTEXT000000801164. Consulté le 5 mai 2015.

McKinsey Global Institute, J. Manyika, M. Chui, P. Groves, D. Farrell, S. Van Kuiken, et E. A. Doshi (2013). Open data : Unlocking innovation and performance with liquid information. http ://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information. Consulté le 5 mai 2015.

Merton, R. K. (1973). The normative structure of science. In N. W. Storer (Ed.), *The sociology of science : theoretical and empirical investigations*, pp. 273. The University of Chicago Press.

NAP (1995). On the Full and Open Exchange of Scientific Data. The National Academies Press. http ://www.nap.edu/readingroom.php ?book=exch&page=summary.html. Consulté le 5 mai 2015.

NAUK (2010). Open Government Licence for public sector information. http ://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/. The National Archives UK. Consulté le 5 mai 2015.

Nicolas, L. (2012). Interview de la Commissaire européenne Neelie Kroes : « data is the new oil ». http ://www.taurillon.org/4730. Le Taurillon, magazine eurocitoyen. Consulté le 5 mai 2015.

OGD (2010). The Annotated 8 Principles of Open Government Data. http ://opengovdata.org/. Consulté le 5 mai 2015.

OKF (2009). Where Does My Money Go ? http ://wheredoesmymoneygo.org/about.html. Open Knowledge Foundation. Consulté le 5 mai 2015.

OKF (2015). The Open Definition. http ://opendefinition.org/. Open Knowledge Foundation. Consulté le 5 mai 2015.

Ordonnance (2005). Ordonnance n° 2005-650 du 6 juin 2005 relative à la liberté d'accès aux documents administratifs et à la réutilisation des informations publiques. http ://www.legifrance.gouv.fr/affichTexte.do ?cidTexte=JORFTEXT000000629684.

Consulté le 5 mai 2015.

Ouest-France (2013). Rennes. le site internet du réseau de transport star/keolis piraté. http ://www.ouest-france.fr/rennes-le-site-internet-du-reseau-de-transport-star/keolis-pirate-155323/. Ouest-France.fr. Consulté le 5 mai 2015.

PMAP (2014). Administrateur général des données : rencontre avec Henri Verdier. http ://www.modernisation.gouv.fr/laction-publique-se-transforme/en-ouvrant-les-donnees-publiques/administrateur-general-des-donnees-chief-data-officer-interview-henri-verdier. Portail de la modernisation de l'action publique. Consulté le 5 mai 2015.

RATP (2015). Open Data FAQ. http ://www.ratp.fr/fr/ratp/v_133048/demarche-faq/. Consulté le 5 mai 2015.

Rennes DGIC (2015). L'open data : c'est quoi ? http ://www.data.rennes-metropole.fr/notre-demarche/l-open-data-c-est-quoi/. Rennes métropole en acces libre. Consulté le 5 mai 2015.

Ribes, D. et S. J. Jackson (2013). Data Bite Man : The Work of Sustaining a Long-Term Study. In L. Gitelman (Ed.), *"Raw data" is an oxymoron*, Infrastructures series, pp. 147–166. The MIT Press.

Salaün, J.-M. (2014). La (ré)utilisation et l'exploitation des données ouvertes favorisent-elles leur économie ? http ://archinfo24.hypotheses.org/2426. Economie du document. Consulté le 5 mai 2015.

SNCF (2015). Licence SNCF Open Data. https ://data.sncf.com/licence. Consulté le 5 mai 2015.

Stanley, M. (2013). Where Is That Moon, Anyway ? The Problem of Interpreting Historical Solar Eclipse Observations. In L. Gitelman (Ed.), *"Raw data" is an oxymoron*, Infrastructures series, pp. 77–88. The MIT Press.

Temesis (2015). La liste des 72 bonnes pratiques OpenData. https ://checklists.opquast.com/fr/opendata/. Opquast Checklists. Consulté le 5 mai 2015.

Tricoire, A. (2006). Le droit d'auteur au service de l'industrie ou la mort de l'autonomie de l'art. http ://www.observatoire-omic.org/colloque-icic/pdf/TricoirereTR2.pdf. Consulté le 5 mai 2015.

US Gov (2007a). An act to provide greater transparency in the legislative process. http ://www.gpo.gov/fdsys/pkg/PLAW-110publ81/content-detail.html. Consulté le 5 mai 2015.

US Gov (2007b). An act to promote accessibility, accountability, and openness in Government by strenghthening section 552 of title 5, United States Code (commonly referred to as the Freedom of Information Act), and for other purposes. http ://www.justice.gov/sites/default/files/oip/legacy/2014/07/23/amendment-s2488.pdf. Consulté le 5 mai 2015.

Wikipédia (2014). Interface de programmation. http ://fr.wikipedia.org/w/index.php ?title=Interface_de_programmation&oldid=107311411. Page Version ID : 107311411. Consulté le 5 mai 2015.

Wikipédia (2015). Donnée publique. http ://fr.wikipedia.org/w/index.php ?title=Donn%C3%A9e_publique&oldid=113844480. Page Version ID : 113844480. Consulté le 5 mai 2015.

Zanibbi, R., D. Blostein, et J. R. Cordy (2004). A survey of table recognition. *Document Analysis and Recognition 7*(1), 1–16.

Summary

Open data, this new subject of overall attention, that some don't hesitate to picture as the "black gold" of modern times, quickly became one of the major concerns of humanity, which no one can take lightly nor to ignore. Nowadays, one tries to pay more attention to this new societal, cultural and economical paradigm, in order to better apprehend its complexity as well as measuring the impact it may have or has already, on our societies. This study offers an update of the Open data movement and on the open data themselves. It especially introduces the key topics (at legal, technical, and economic levels) to anyone who wants to get into the Open data venture in order to better assess its nature and its various facets.