



À l'heure de la mobilité connectée, quel avenir pour les données ouvertes de transport en commun ?

LAURE GUICHERD

Mastère Spécialisé Création et Production Multimédia

Promo 2012-2013

June, Twenty Two

73 rue Sainte-Anne

75002 Paris

Suiveur de Stage à l'INA

Matthieu Richy-Dureteste

Tuteur en entreprise

Joachim Breton

June,
TwentyTwo.

Sommaire

Introduction

L'open data, qu'est-ce que c'est ?

Définition 5

Les acteurs 6

Pour quels bénéfices ? 7

Les risques de l'ouverture des données 7

Histoire 9

Les courants de pensée voisins 10

Open data et Open source 11

Open data, Linked data, Big data ? 12

Le cadre juridique 13

La loi CADA et Etalab 13

Les licences 16

L'état des lieux 16

Type de données ouvertes 16

Les initiatives existantes en France 18

Les modes de mise à disposition 20

Les résultats des réutilisations 22

L'open data et les entreprises privées 24

Un contexte favorisant l'ouverture 24

La transparence pour améliorer la confiance 25

Une opportunité d'innovation 25

Rendre leurs données aux utilisateurs 26

L'entreprise réutilisatrice des données 27

Des réticences compréhensibles 27

Les données au cœur du transport public

La mutation actuelle des transports 29

De quelles données parle-t-on ? 32

Représentation des données fondamentales de transport 33

Pour quelles utilisations ? 35

Formats des données de transport 36

Les acteurs en jeu 41

La loi LOTI et les AOT 41

Le cas particulier de l'Île-de-France 42

Quelques entreprises privées qui jouent un rôle important	42
Le cas Google.....	43

Que dit la loi ?	47
-------------------------------	-----------

Les initiatives d'ouverture de données en France	48
---	-----------

Initiatives des AOT locales	48
-----------------------------------	----

Initiatives des opérateurs privés	50
---	----

La forme de cette ouverture	54
-----------------------------------	----

Les réutilisations	55
---------------------------------	-----------

Quel avenir pour les données ouvertes de transport public ?

Une volonté publique	56
-----------------------------------	-----------

Des réticences compréhensibles.....	56
--	-----------

Les pistes à explorer	57
------------------------------------	-----------

Y mettre de la bonne volonté	58
------------------------------------	----

Encourager l'innovation ouverte.....	58
--------------------------------------	----

Faire confiance aux destinataires de l'ouverture.....	59
---	----

La problématique complexe des données temps réel.....	60
---	----

Les enjeux de l'intermodalité	61
-------------------------------------	----

Faire la guerre à Google ou travailler ensemble ?	61
---	----

Vers un assistant personnel de voyage	62
---	----

Monétiser, pourquoi pas ?	64
---------------------------------	----

Conclusion

Bibliographie

Glossaire

Remerciements

Introduction

Si aujourd’hui, le numérique et le web sont devenus les fers de lance d’un marché économique mondial à la croissance rapide et aux bénéfices pharaoniques, peut-être devrions-nous, maintenant la maturité atteinte, nous pencher à nouveau sur les fondamentaux et les motivations premières derrière ces concepts. En effet, de quoi parlons-nous, lorsque nous évoquons le « numérique », de plus que d’une façon de représenter, stocker et surtout partager des **données** ? S’affranchir de l’analogique et des supports physiques a remis en cause pour toujours notre façon d’envisager l’information, mais aussi de la posséder, d’y accéder et de la partager à l’infini via le réseau internet. Nous sortons aujourd’hui d’une ère durant laquelle les puissants étaient ceux qui détenaient le savoir, pour un monde dans lequel, il est techniquement possible d’accéder par tous à l’information, sans limite de temps, de quantité, et sans restriction d'accès — voire même depuis n'importe où grâce à l'avènement du mobile — remettant en cause tous les statuts sociaux établis auparavant.

En parallèle, on constate ces dernières années un très fort engouement pour le « **faire ensemble** », imposé en partie par la crise économique mondiale, mais qui a trouvé par la suite son public via de nombreux rassemblements — de spécialistes comme de citoyens — par l’intermédiaire de sites communautaires, de *fab labs* et d’initiatives diverses.

Toutes les conditions sont réunies aujourd’hui pour que le citoyen prenne en main ces outils pour reprendre le pouvoir sur ce qui a été à l’origine du web, c'est-à-dire les données, devenues, avec la marchandisation de l’outil internet, un levier de monétisation comme un autre. Nous allons voir que des mouvements se sont soulevés pour demander la mise à disposition des données qui concernent le citoyen au premier plan, réclamant de pouvoir y accéder via les nouveaux systèmes d’information. L’open data est un mouvement complexe et protéiforme, qui trouve ses origines dans de nombreux mouvements aux motivations diverses, et qui a des enjeux et répercussions aussi bien techniques que politiques et juridiques ; c’est la raison pour laquelle j’ai choisi ce sujet pour l’établissement de cette thèse professionnelle.

J’ai par ailleurs choisi d’approfondir le sujet en me penchant sur les rapports riches et complexes de l’open data avec les transports en commun, sujet très actuel dont l’agence où j’ai effectué mon stage est une participante active, qui contribue actuellement à son évolution et en redéfinit les contours en continu.

L'open data, qu'est-ce que c'est ?

Avant de tenter toute définition de l'open data, il convient de délimiter correctement ce que l'on entend par « data », en français « donnée ». Qu'est-ce qu'une donnée ? Le Larousse nous donne la définition suivante : « *ce qui est connu ou admis comme tel* ». Une donnée serait simplement, donc, un état de fait connu. Comme nous l'avons dit plus tôt, on peut considérer que *tout* est donnée.

Il faut noter, pour qui serait tenté d'utiliser le synonyme d'« information », que les deux notions sont distinctes. En effet, une donnée est une information, mais une information à l'état brut, indiscutable et neutre. Une information, par contre, n'est pas forcément une donnée, mais peut être le résultat de l'interprétation d'une donnée.

La deuxième définition fournie par le Larousse est d'un autre ordre : « *représentation conventionnelle d'une information en vue de son traitement informatique* ». En évoquant, dans la définition même de la donnée, son utilisation informatique, on touche du doigt le sujet de ce mémoire et on entrevoit déjà les problématiques à venir : que peut-on considérer comme une donnée ? Comment représenter une donnée ? Pour quelles réutilisations ?

Définition

L'open data, traduit en français par « *données ouvertes* », est un mouvement consistant à mettre à disposition des données numériques sur le web, afin qu'elles soient accessibles et utilisables par tous, selon des modalités bien précises, que nous allons détailler ici.

En 2007, l'*Open Government Working Group*¹ a organisé une rencontre à Sebastopol, en Californie, pour formaliser ce qui deviendra plus tard une référence dans la définition de ce qu'est une donnée ouverte. Selon ces principes, regroupés sous le nom de « *8 principles of open government data* », il est considéré qu'une donnée est ouverte si elle est mise à disposition selon les principes suivants :

1. Les données doivent être **complètes** (avec deux exceptions fondamentales, celle des données sensibles et celle des données personnelles),
2. Les données doivent être **primaires**, collectées à la source et non modifiées,
3. Les données doivent être **à jour**,
4. Les données doivent être **accessibles** au plus grand nombre d'utilisateurs, pour le plus grand nombre d'utilisations,
5. Les données doivent permettre le **traitement automatisé** en étant structurées de façon raisonnable

¹ Le groupe de travail était composé entre autres de Tim O'Reilly, célèbre éditeur à qui l'on doit la définition du web 2.0, Lawrence Lessig, grand promoteur des licences Creative Commons et professeur de droit à Stanford, et d'Aaron Swartz, militant de l'Internet et inventeur du RSS.

6. Leur accès doit être **non discriminant**, accessible à tous sans condition d'abonnement ou souscription
7. Les données doivent être mises à disposition dans un **format non propriétaire**
8. Les données doivent être mises à disposition sous une **licence libre**, elles ne doivent pas être soumises à des droits de propriété intellectuelle.

Ces critères font certes, jusqu'ici, figure de référence quand à l'évaluation de la qualité des initiatives d'ouverture de données, mais il est important de souligner que ce sont des principes généraux et exigeants qu'il faut envisager dans une perspective idéale, et que dans les faits, ils sont rarement appliqués à la lettre, comme nous allons le voir lorsque nous détaillerons les initiatives existantes.

OPEN DATA

Suivant ces lignes de conduite, on pourra aisément distinguer données publiques et données ouvertes ; par exemple, les données publiées chaque année sur les établissements de soin en France par la HAS (Haute Autorité de Santé), bien que répondant, dans le principe, à une volonté de transparence des services publics, ne correspondent pas à la définition de données ouvertes, puisque ces données sont mises à disposition sous le format PDF, qui est un format propriétaire et non réutilisable, et qu'elles ne sont pas structurées pour la réutilisation informatique, puisque mises en forme et déjà interprétées pour faciliter la lecture directe. Par ailleurs, aucune licence n'est mentionnée qui donnerait un cadre à une éventuelle réutilisation : ces données n'ont simplement pas voca-

tion à être transformées et réutilisées, mais sont fournies telles quelles, pour une consultation directe.

Les acteurs

Bien sûr, l'open data n'est pas uniquement une affaire de données, mais aussi de personnes et d'acteurs. Trois rôles principaux interviennent dans le jeu d'ouverture des données :

- Le **détendeur des données** : une institution ou organisation qui souhaite mettre des données à disposition ; on verra par la suite que cela concerne en grande partie l'administration publique ou les collectivités locales, mais les entreprises sont aussi concernées, et même les particuliers.
- Les **réutilisateurs** : ceux qui vont utiliser les données mises à disposition afin de créer un résultat (application, visualisation, rapport...) : ce peut être n'importe qui, mais on rencontrera souvent des journalistes, graphistes, développeurs ou entreprises, voire des chercheurs.
- Les **utilisateurs finaux** : ceux qui consultent les données en direct ou via une réutilisation. Ce sont des particuliers, et souvent des activistes ou des associations engagés dans la vie citoyenne.

Ces rôles ne sont pas figés, et peuvent parfois se confondre ; en effet, il peut arriver que le détendeur des données propose en parallèle des données brutes des visualisations permettant la compréhension directe : il est dans ce cas aussi un réutilisateur. De même, un réutilisateur peut aussi être uti-

lisateur final, par exemple dans le cas d'applications spécialisées.

Pour quels bénéfices ?

Les bénéfices de l'ouverture des données sont en fait de deux ordres, qui se complètent l'un l'autre.

Le premier, et le plus cité, est un argument politique, celui de l'amélioration démocratique. En mettant des données publiques à disposition, on permet à chacun de participer à ce que plusieurs mouvements citoyens ont appelé une « **démocratie ouverte** »¹, un mode de gouvernance qui s'appuie sur trois piliers fondamentaux :

- **Transparence** : contrer la défiance des citoyens envers les institutions en rendant l'action publique compréhensible et claire. Cette philosophie a beaucoup à voir avec la notion d'*accountability*, chère aux anglo-saxons, pour qui la politique est envisagée comme une interaction continue — voire une stimulation indispensable — entre le pouvoir en place et l'opposition parlementaire, bien loin de la notion d'état-protecteur à la française, et qui requiert de rendre des comptes au regard public de façon continue pour justifier son action. Cette demande de comptes s'appuie sur le principe que quiconque participe au fonctionnement du pays en payant des impôts doit pouvoir accéder aux données et aux projets produits de façon libre et gratuite.

- **Participation** : en ouvrant l'accès aux données, on donne aussi un moyen de participer à l'action publique en donnant des retours, en posant des questions et en ouvrant des débats. C'est ce que l'on appelle en anglais l'*empowerment*² du citoyen. C'est un moyen pour les institutions de se mettre à l'écoute des citoyens, et de les consulter en continu sur des questions qui concernent tout un chacun.
- **Collaboration** : au-delà d'une participation individuelle du citoyen, les données peuvent stimuler et donner les moyens à divers composantes de la société civile (entreprises, associations, *think tanks*, journalistes, collectivités) de s'organiser comme acteurs à part entière de l'action publique, et ainsi dynamiter les structures verticales pour être force de proposition de façon rapide et efficace ; à noter que cet axe se rapproche de la deuxième famille des arguments en faveur de l'open data :

Le deuxième argument est un argument économique, qui envisage l'ouverture des données comme un **facteur d'innovation**, et une opportunité de créer de nouveaux services pour améliorer la vie quotidienne. Nous reviendrons plus tard longuement sur les enjeux sous-tendus par cet arguments, ses limitations et les défis à relever.

Les risques de l'ouverture des données

Bien sûr, des critiques, voire des limitations, sont à relever. En tout premier lieu, des critiques liées à la nature même de ces données ouvertes, qui, si el-

¹ Les anglophones parlent d' « Open Government ».

² la traduction française “capacitation” n'est ni élégante, ni parlante, car l'on perd l'idée de prise de pouvoir, voire de renversement des positions établies.

les sont sensées être parfaitement neutres, ne peuvent en réalité jamais l'être tout à fait ; une donnée brute n'existe pas en soi, mais est toujours le résultat d'une interprétation (par l'initiateur ou par l'éditeur de la donnée) et d'un choix éditorial (le choix des données à montrer est forcément orienté). Il ne faut pas non plus oublier les marges d'erreur dues à la mesure (bruit du capteur, approximations...).

Par ailleurs, il est évident que l'on ne peut pas tout libérer. Comme nous le verrons plus tard, les données sensibles et personnelles sont naturellement exclues, mais de façon plus générale, il est utopique de penser que tout est résumable à un jeu de données, parce que l'aspect structuré de ces données empêche toute nuance et contextualisation. Aussi, cette absence d'interprétation et de clés de lecture, intrinsèque au concept de données ouvertes, a ses revers : dès lors que l'on a besoin d'un intermédiaire pour extraire, interpréter et mettre en forme les données, l'*empowerment* se trouve limité, puisque le citoyen lambda a besoin de réelles connaissances techniques, politiques et même juridiques pour pouvoir réutiliser la donnée brute en toute connaissance de cause, ainsi que pour la diffuser : on touche au risque de **ne pouvoir s'adresser qu'aux plus capables**, en oubliant les vraies cibles du mouvement que sont tous les citoyens, quelque soient leurs capacités.

Plus généralement, le mouvement open data étant encore très jeune, de nombreux malentendus subsistent à son encontre, du fait que son adoption a été très rapide, que de nombreuses composantes (techniques, juridiques, politiques) sont à prendre

en compte, et même, peut-être, à cause de l'imprécision du terme : *open data* — terme anglais, ce qui ne facilite pas l'assimilation par le public francophone — fait référence à la fois à l'objet (la donnée ouverte) et à la démarche politique qui l'entoure.

Tout n'est donc pas rose dans le monde de l'*open data*, et l'ouverture à tout prix n'est pas ce qu'il faut chercher à atteindre. Les observateurs notent plusieurs risques à cette ouverture, si elle n'est pas bien encadrée, ou faite dans des conditions mal choisies. Une fois les données mise à disposition, il faut rester vigilant sur les réutilisations et les conséquences, économiques en particulier, qui peuvent en découler. Par exemple, le principe "par défaut" de gratuité de la réutilisation est régulièrement critiqué, pointé du doigt pour le risque de **fragilisation des modèles économiques** de production qu'il induit.

Est souvent évoqué aussi le risque de **privatisation de la donnée publique** : comment empêcher les entreprises privées de faire des bénéfices en utilisant ces données, puis en invoquant la concurrence pour en rester les seules propriétaires¹ ?

Le fait de considérer les données comme indépendantes de leur support et de leur propriétaire pose aussi des questions sur la **désintermédiation de la donnée** : en montrant des informations hors de leur contexte d'origine, ne perd-t-on pas un peu de leur intérêt, et même plus, les intermédiaires d'autrefois ayant trouvé un modèle économique ne risquent-ils pas de disparaître ?

¹ Nous verrons plus loin que le choix de la licence OdBL limite fortement ce risque, en obligeant à une réutilisation avec les mêmes limitations.

Tout au bout de la chaîne de valeur, il ne faut pas oublier les destinataires des réutilisations, confrontés au risque de se retrouver face à une pléthore d'objets de qualité très diverses : comment éviter l'**éparpillement des réutilisations** et le manque de visibilité pour le public ?

Histoire

À l'origine, l'ouverture des données trouve ses racines dans le monde de la recherche scientifique. La première apparition de l'expression *open data* se trouve d'ailleurs dans un document d'une agence de recherche américaine en 1995¹. Ses auteurs y évoquaient le besoin impérieux d'abandonner les droits de propriété intellectuelle et de mettre à disposition le résultat de leurs recherches, en justifiant ce choix par le fait que la connaissance appartient à tous, et que chacun doit pouvoir y accéder pour pouvoir enrichir et apporter sa pierre à l'édifice du savoir commun. Bien avant l'invention même d'internet, en 1942, Robert King Merton, célèbre sociologue des sciences, avait théorisé ce concept de « pot commun » du savoir scientifique.

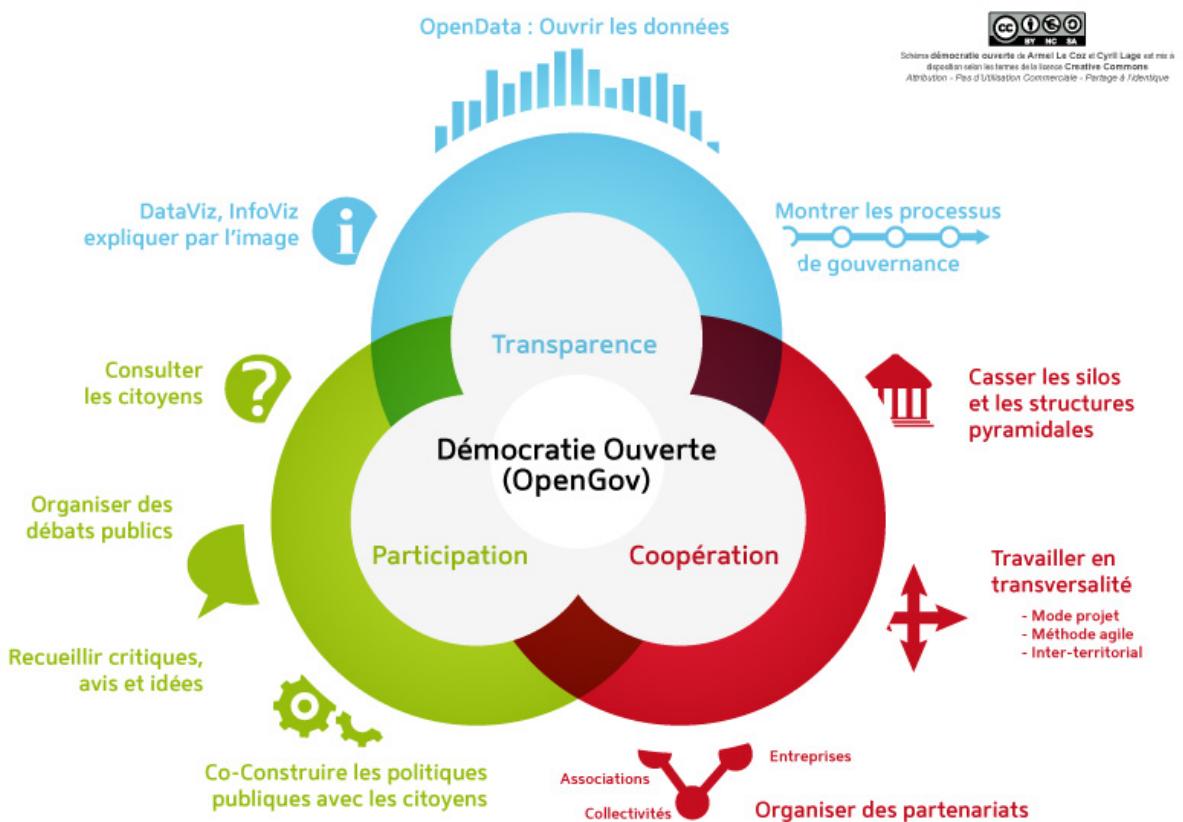
Une douzaine d'années plus tard, le travail de l'*Open Government Working Group*, cité plus haut, met un grand coup d'accélérateur dans la propagation de ces idées nouvelles. Leur formalisation de la définition de données ouvertes a pour but, un an avant les élections présidentielles américaines, de faire adopter la démarche d'ouverture des données sous l'angle politique. Ils sont appuyés par

de nombreux militants à travers le monde, mais aussi par la voix des médias, comme le *Guardian*, au Royaume-Uni, qui va oeuvrer activement dans la même perspective. Ces efforts vont porter leurs fruits : lors de sa campagne de 2008, Barack Obama va se faire un ardent défenseur de l'*open government*, à travers son slogan « *We, president* », qui souligne cette idée d'amélioration de la démocratie. Le jour de son élection, il signe un memorandum² — geste fort qui va orienter l'action de son gouvernement pour les années à venir — dans lequel il s'engage à « *exploiter les nouvelles technologies pour mettre en ligne et rendre facilement accessibles au public des informations sur leurs activités et décisions prises* » (transparence), « *offrir aux citoyens des possibilités accrues de contribuer à l'élaboration des politiques et de fournir à leur gouvernement les bénéfices de leurs savoirs et expertise collective* » (participation), « *utiliser des outils innovants pour coopérer entre eux, à tous les niveaux de gouvernement, mais aussi avec les ONG, des entreprises et des particuliers du secteur privé.* » (collaboration). Son gouvernement, depuis, multiplie les initiatives, dont une des plus notables est le lancement du portail *data.gov* en mai 2009.



¹ « *On the full and open exchange of scientific data* », National Research Council, 1995

² « *Transparency and open government* », http://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment



Source : <http://www.flickr.com/photos/46274765@N07/6938502167/>

Depuis cet événement qui a posé un jalon dans le développement des démarches open data, de nombreuses administrations dans le monde ont entrepris d'ouvrir leurs données, et à l'heure actuelle le débat est plus animé que jamais sur les enjeux, les bénéfices et les risques de cette ouverture.

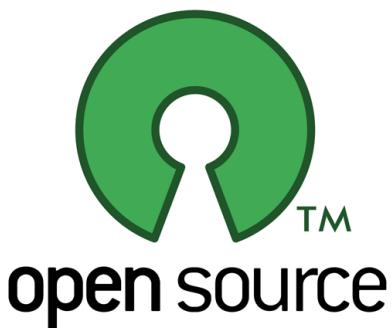
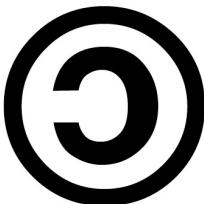
Dans ses fondements tout d'abord, on constate une parenté très forte avec la philosophie des « **biens communs** », notion très ancienne définie par le droit romain, qui considère comme « biens communs » les choses qui appartiennent à tous et donc ne peuvent appartenir à personne, comme la mer et l'air. Ce concept va évoluer petit à petit, jusqu'à être fréquemment appliqué aux idées environnementalistes (la couche d'ozone, l'oxygène), mais aussi au monde des idées, avec une différence fondamentale : si les autres biens communs ne sont pas inépuisables, et doivent être partagés par la communauté, les idées sont dites « non-soustractables », c'est-à-dire que leur usage par l'un n'empêche par leur usage par les autres.

Les courants de pensée voisins

Comme on a pu l'entrevoir, l'open data, dans ses origines, ses acteurs et ses objectifs, est au croisement de plusieurs cultures, qui s'entrecroisent, et même parfois se contredisent.

Open data et Open source

On le pressent en évoquant cette volonté de partage illimité : cette idée de bien commun a un rapport très fort avec le **monde du logiciel libre** (en anglais *open source*). Née en 1984 avec la *Free Software Foundation* de Richard Stallman, et son projet *GNU* (qui sera à l'origine des systèmes d'exploitation modernes basés sur Linux), la philosophie open source promeut la liberté de l'utilisateur des logiciels, via deux principes. Le premier est la liberté d'accès (et d'exécution) au logiciel et à son code source. Le second – et le plus important – est la possibilité (très encouragée) de copier ou réutiliser ce code source, et de redistribuer les logiciels en résultant, à la condition que ces réutilisations restent elles aussi ouvertes : c'est le principe (inventé par Stallman) du *copyleft*, qui à l'inverse du copyright qui restreint les droits d'utilisation, assure leur préservation. La première licence à voir le jour sous cette bannière est la licence *GPL* (pour *General Public Licence*) ; c'est aussi la plus utilisée à l'heure actuelle¹.



Le but est de donner à l'utilisateur un contrôle total sur les logiciels qu'il utilise, voire sur son ordinateur dans le cas de systèmes d'exploitation

libres. Les logiciels libres sont présentés comme l'alternative aux logiciels propriétaires, considérés comme privateurs de liberté. La définition des logiciels libres par la *Free Software Foundation* l'énonce clairement : « *Quand les utilisateurs ne contrôlent pas le programme, c'est le programme qui contrôle les utilisateurs. Le développeur contrôle le programme, et par ce biais, contrôle les utilisateurs. Ce programme non libre, ou "privateur", devient donc l'instrument d'un pouvoir injuste.* »²

Au-delà de la prise de pouvoir individuelle sur les outils se construit tout un fonctionnement de communautés de développement ; si le logiciel montre un dysfonctionnement ou des manquements, et si je n'ai pas les compétences pour participer à son évolution ou sa correction, ou si je suis simplement utilisateur final du logiciel sans volonté de réutilisation, je suis assuré que d'autres le feront. Ainsi, plus la communauté réutilisatrice est importante, plus le logiciel jouit d'une garantie de qualité et de réactivité.

Aujourd'hui, l'open source est une vague de fond, bien connue du grand public, et solidement installée dans les usages. Parmi les exemples représentatifs, on peut citer les systèmes d'exploitation basés sur **Unix** (comme Linux) très majoritaires sur les systèmes embarqués (téléphones, drones, robots) et les super-calculateurs, mais aussi le serveur **Apache**, présent sur presque trois quarts des serveurs web dans le monde, ou encore le navigateur **Firefox** de la fondation Mozilla (28% des internautes), ou encore le système d'exploitation **Android**, publié par Google sous licence open source en 2007 et qui a depuis dépassé iOS (le

¹ 68.5% des projets listés sur sourceforge.net sont publiés sous licence GPL.

² Définition du logiciel libre, gnu.org : <http://www.gnu.org/philosophy/free-sw.fr.html>

système d'exploitation d'Apple) en nombre de périphériques équipés.

Entraide, partage, communautés, participation de tous, prise de pouvoir par rapport aux solutions existantes... On constate une très grande analogie entre l'open source et l'open data, ce dernier étant un mouvement bien plus jeune, et surtout moins connu du grand public. On pourrait le considérer comme une continuité logique : si le logiciel libre a atteint sa maturité et a prouvé ses qualités, pourquoi ne pas étendre ses principes, très généraux et utopistes, à d'autres possessions citoyennes immatérielles ? Nous verrons que cela induit la nécessité de sortir le débat des sphères techniques et spécialisées, pour pouvoir s'adresser à chaque citoyen.

Preuve éclatante du lien entre les deux univers, le 15 octobre 2013 a vu la naissance de la plate-forme government.github.com, créée à l'initiative du site web Github¹. Le célèbre outil de développement collaboratif ouvre ainsi ses portes aux projets citoyens et gouvernementaux, leur offre une visibilité et donne à chacun des moyens clairs pour y contribuer, avec comme slogan « *Make government better, together* ».

Open data, Linked data, Big data ?

Plusieurs acteurs fondamentaux de l'informatique sont d'ardents défenseurs, voire des précurseurs, de l'open data. En effet, **Tim Berners-Lee**, qui n'est rien de moins que l'inventeur du web² et fon-

dateur du W3C, a proposé en 2010 une échelle permettant de jauger la qualité d'une donnée ouverte :

- ★ : Donnée (même dégradée) mise à disposition sur le web sous licence ouverte, quel que soit le format
- ★★ : Donnée structurée (tabulaire, par exemple)
- ★★★ : Donnée disponible dans un format non propriétaire
- ★★★★ : Donnée accessible de façon directe via une URI
- ★★★★★ : Donnée contextualisée et enrichie par d'autres données auxquelles elle est liée.

Cette échelle est dans la continuité des travaux de Berners-Lee sur le **web sémantique**, considéré par certains comme ce que sera le web 3.0. Berners-Lee le définit comme « *un web de données qui peuvent être traitées directement et indirectement par des machines, pour aider leurs utilisateurs à créer de nouvelles connaissances* ». Les informations disponibles sur le web ne seraient plus seulement des documents HTML ("pages") liées entre elles par des liens hypertexte, mais des données structurées et liées entre elles, qui seraient accessibles non seulement par les humains, mais aussi par les machines, qui en tireraient du sens grâce à des métadonnées précises, additionnées du contexte sémantique apporté par ces

¹ Service d'hébergement et de gestion de développement de logiciels, Github est très connu et utilisé par les développeurs du monde entier pour la création, le partage et la réutilisation de projets open source, des plus personnels aux plus célèbres. Son slogan, très parlant, est le suivant : « *build software better, together* ».

² Il a plus exactement théorisé le concept de lien hypertexte dans le cadre de ses travaux au MIT en 1989.

liens.

C'est ce qu'il définit comme le « **linked data** », qui pourra, selon lui, nous permettre de trouver, partager et analyser les données très simplement, en automatisant leurs recherches et même leur analyse, et in fine à étendre nos connaissances.

En effet, l'explosion de la quantité de données disponibles, en particulier sur le web, rend leur appréhension, leur stockage, leur partage et leur analyse et de plus complexes pour l'humain. Les enjeux et les ébauches de solutions à cette problématique sont définis par le phénomène **big data**, qui rencontre depuis quelques années un succès scientifique et médiatique grandissant. De nombreuses entreprises espèrent tirer profit de ces mines d'informations disponibles, mais pas encore reliées entre elles et analysées.

Cette tendance au traitement de données complexes, nombreuses et foisonnantes ouvre un champ passionnant à la recherche, et permet actuellement la création d'outils qui ouvriront la voie à l'ouverture des données et à leur utilisation éclairée. Cependant, le travail qui reste à accomplir est encore énorme : comment unifier ces données, quels formats adopter¹, quelles normes suivre, quelles métadonnées ?

Ainsi, l'open data se situe à la croisée des chemins, entre évolution de fond du web, recherches universitaires, utopies concrètes et intérêts économiques.

Le cadre juridique

La loi CADA et Etalab

En France, l'accès de tous les citoyens aux données publiques est un droit acquis de longue date ; en effet, l'article 15 de la Déclaration des Droits de l'Homme et du Citoyen énonce que « *la société a le droit de demander compte à tout agent public de son administration* ». Néanmoins, pour que ce droit fondamental soit encadré par une loi, il faudra attendre l'année 1978, qui verra naître deux lois fondamentales :

- La bien connue loi « Informatique et Libertés », qui promulguera la naissance de la **CNIL** (Commission Nationale de l'Informatique et des Libertés), chargée de veiller à la protection des données à caractère personnel, qui voit le jour le 6 janvier
- La loi du 17 juillet 1978, bien moins connue du grand public, porte à l'inverse sur l'accès des citoyens aux données publiques. Elle donne naissance à la **CADA** (Commission d'Accès aux Documents Administratifs), organisme chargé d'assurer le bon fonctionnement de cette loi.

La CADA encadre donc la mise à disposition des contenus, et spécifie les conditions de cet encadrement. Les données publiques concernées sont les suivantes :

- Les documents de toute sorte : « *dossiers, rapports, études, comptes rendus, procès-verbaux, statistiques, directives, instructions,*

¹ On parle beaucoup du format RDF à propos du linked data, mais aucun format n'a encore réellement été adopté.

circulaires, notes et réponses ministérielles, correspondances, avis, prévisions et décisions », quelle que soit leur date ou forme, du moment qu'ils sont achevés

- Produits par l'État, les collectivités territoriales et même les « *autres personnes de droit public ou privé chargées d'une mission de service public* »
- Dans le cadre d'une mission de service public.

Elle stipule, et c'est le cœur de cette loi, que ces autorités sont tenues de communiquer ces données à quiconque en fait la demande, et peut saisir la CADA en cas de refus.



Ne sont pas concernées : les données sensibles (secret de la défense nationale, sûreté de l'État) ou personnelles (nominatives ou relevant du secret médical), ainsi que les données de la recherche, de l'enseignement et de la culture, pour lesquelles c'est le droit d'auteur qui va s'appliquer.

Cette loi a été modifiée en 2005 par une directive du parlement européen, qui introduit le droit de réutilisation des données pour des applications autres que celles prévues initialement, tant que les réutilisateurs respectent la licence sous laquelle a été publiée la donnée, que les informations ne soient pas altérées, et que leur source originale et la date de dernière mise à jour soit mentionnée. Les réutilisations peuvent être à caractère commercial ou non.

Si l'accès aux données, jusque là, était conçu du point de vue de la transparence, ces nouvelles considérations sont clairement plus économiques, et perçoivent la possibilité de réutiliser les données publiques comme une véritable opportunité d'innovation. Il faut bien comprendre que la CADA n'impose pas aux administrations de mettre en ligne les données gratuitement ; en fait, elle prévoit même la possibilité d'instaurer des redevances de réutilisation, et définit même une façon de les calculer, en fonction du coût de la libération des données (collecte, saisie, stockage, mise à disposition, maintenance).

Une exception, et pas des moindres (Art. 10), concerne les données produites dans l'exercice d'une mission de service public à caractère Industriel et commercial, qui font l'objet d'un droit d'accès, mais pas de réutilisation. Nous verrons que cette exception est d'importance pour le sujet qui nous intéresse, puisque les autorités concernées sont les Établissements Publics à Caractère Industriel et Commercial (EPIC), statut particulier créé pour les missions de service public qui ne pourrait être correctement effectuées par des entreprises soumises à la concurrence. C'est le cas par exemple de l'Organisme Français du Sang ou de l'Institut National de l'Audiovisuel, mais aussi de la RATP et de la SNCF.

Suite à la mise en place de l'Open Government de Barack Obama après son élection en 2009, de nombreux débats vont avoir lieu sur la mise en place d'un procédé similaire en France, et en particulier sur la gratuité éventuelle de la réutilisation, soutenue par de nombreuses associations et acteurs publics. Ces débats vont mener, le 21 février 2011, à la création d'**Etalab**, service du Premier



ministre qui a pour mission la création et l'encaissement du portail interministériel des données publiques françaises : **data.gouv.fr**. Cette création s'accompagne d'un décret qui détaille les conditions de la mise en place d'une redevance de réutilisation : il sera désormais acquis que les réutilisations seront maintenant gratuites par défaut, et les administrations qui souhaiteraient faire payer leurs données devront préalablement déposer une demande auprès du COEPIA¹.

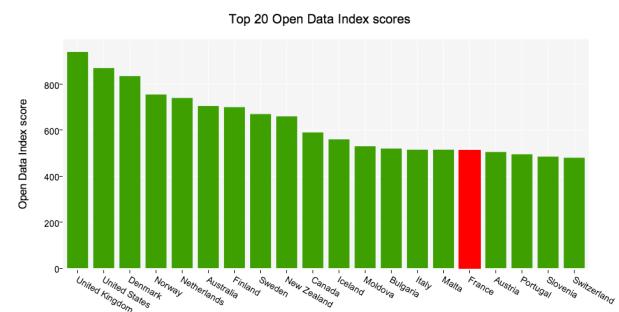
Le portail *data.gouv.fr* est inauguré en décembre 2011, et contient alors quelques 330 000 jeux de données, qui ont pour thème les questions sociales, l'emploi, l'éducation, l'énergie ou l'environnement.



Le 18 juin 2013, les pays du G8 signent la « *Charte du G8 pour l'ouverture des données publiques* », qui, dans la continuité de l'impulsion lancée par Etalab en France, réaffirme les principes de libre réutilisation et de gratuité, et engage les états à mettre disposition leurs données, qui sont « ouvertes par défaut ».

Malgré la volonté affichée par Etalab, il faut nuancer le succès de la démarche de la France par rapport aux autres pays. L'*Open Knowledge Foundation* a publié en octobre 2013 son *Open Data Index*, dont le but est de rendre compte des

initiatives d'ouverture de données gouvernementales dans le monde, et le bilan n'est pas des meilleurs pour la France, qui se situe dans le classement global en 16ème position, derrière Malte et la Moldavie.



Classement des 20 pays les mieux notés par l'*Open Data Index* de 2013 (la France est en rouge)

Source : Communiqué de presse de l'*Open Knowledge Foundation* <http://fr.okfn.org/2013/10/28/opendataindex2013/>

Il est notamment reproché à la France de ne pas publier le détail des répartitions des dépenses publiques, et à l'*IGN* de publier ses données géographiques en faible résolution uniquement, ou encore de publier les données sur les entreprises (registre SIRENE) en échange d'une redevance élevée.

Dataset	Score	Breakdown
Transport Timetables	60%	☒ ☒ ☐ \$ ☐ ☒ ☒ 🔒 ☐
Government Budget	90%	☒ ☒ ☐ \$ ☐ ☒ ☒ 🔒 ☐
Government Spending	10%	☒ ☒ ☐ \$ ☐ ☒ ☒ 🔒 ☐
Election Results	90%	☒ ☒ ☐ \$ ☐ ☒ ☒ 🔒 ☐
Company Register	35%	☒ ☒ ☐ \$ ☐ ☒ ☒ 🔒 ☐
National Map	35%	☒ ☒ ☐ \$ ☐ ☒ ☒ 🔒 ☐
National Statistics	75%	☒ ☒ ☐ \$ ☐ ☒ ☒ 🔒 ☐
Legislation	50%	☒ ☒ ☐ \$ ☐ ☒ ☒ 🔒 ☐
Postcodes / Zipcodes	0%	☒ ☒ ☐ \$ ☐ ☒ ☒ 🔒 ☐
Emissions of pollutants	65%	☒ ☒ ☐ \$ ☐ ☒ ☒ 🔒 ☐

Légende : ☒ oui ☒ non ☒ incertain

Source : résultats de l'*Open Data Index* 2013 pour la France <https://index.okfn.org/country/overview/France/>

1 Conseil d'orientation de l'édition publique et de l'information administrative.

Les licences

Si la loi fixe les conditions de mise à disposition des données publiques, reste que tout réutilisateur potentiel devra respecter les conditions de la licence ouverte qui accompagne impérativement la donnée qu'il souhaite exploiter. Il est donc fondamental, pour le détenteur des données comme pour le réutilisateur, de connaître les licences existantes, et les options qui s'ouvrent à lui.

On pourrait imaginer, au premier abord, que la démarche idéale, pour qui libérerait ses données, serait de créer une licence ouverte inédite et sur mesure, dans le but de garder une maîtrise sur ses données. Si l'intention paraît louable, ce serait une erreur, car cela rendrait l'interopérabilité des licences entre les différentes initiatives impossible. De plus cela rendrait l'apprehension de ses droits plus complexe pour le réutilisateur, qui souvent utilise des données de sources diverses pour une même application. Si chaque jeu de données nécessite un traitement différent, le déchiffrage peut devenir très rapidement un véritable casse-tête, d'où la nécessité de proposer des licences libres standard, facilement compréhensibles et adaptées à diverses politiques. Etalab l'a compris, et a créé la **Licence Ouverte** pour les données publiques en parallèle avec data.gouv.fr. Cette licence assure une large liberté, puisqu'elle n'impose aux réutilisateurs que de faire mention de la source des données, se rapprochant du modèle Creative Commons des licences "BY". La réutilisation commerciale est autorisée.



LICENCE OUVERTE
OPEN LICENCE

Une deuxième licence très utilisée en France est la licence ODbL, pour **Open Database Licence**. Elle concernait à l'origine les bases de données, est utilisée notamment par le projet OpenStreetMap, et a été traduite en français en 2010 par la ville de Paris pour l'adapter à un usage national ou territorial. Elle rencontre un franc succès depuis, et est utilisée par de nombreuses collectivités françaises (Paris, Toulouse, Nantes...). D'inspiration moins économique qu'Etalab, leur différence fondamentale réside dans le devoir qu'impose ODbL de redistribuer les données, mêmes enrichies, sous les mêmes conditions (c'est le modèle "SA" (Share Alike) de Creative Commons).

Il faut noter que les deux licences autorisent la modification des données, ce qui est fondamental pour des données temps-réel, ou des réutilisations qui tireraient parti du crowdsourcing.

L'état des lieux

Type de données ouvertes

Les données ouvertes sont très variées et peuvent être classifiées selon plusieurs types de critères.

Les **thématisques** possibles sont très nombreuses, et touchent à tous les champs de la

vie publique. La classification suivante est celle choisie par Simon Chignard dans son ouvrage « *Open data, comprendre l'ouverture des données publiques* » :

- **Vie démocratique** : résultats des votes, subventions, budgets et finances publiques...
- **Démographie** : recensements, prénoms les plus populaires...
- **Économie** : statistiques sur les demandeurs d'emploi, entreprises enregistrées, revenus fiscaux...
- **Environnement** : parcs et jardins, qualité de l'air...
- **Arts, culture et patrimoine** : équipements culturels, leurs tarifs et statistiques de fréquentation, plans cadastraux
- **Urbanisme et habitat** : occupation de l'espace urbain, projets d'aménagement
- **Transport et déplacements** : horaires, fréquentation des lignes, localisation des arrêts... (nous reviendrons plus loin dans le détail de ces types de données)
- **Équipements et services d'intérêt public** (toilettes publiques, points d'eau...), leurs horaires et localisations
- **Localisation et information géographique** : reliefs naturels, modèles de terrain...

Il n'y a pas de norme pour cette classification, et en conséquence chaque plateforme de diffusion de données ouvertes utilise son propre système de classification, ce qui ne facilite pas les comparaisons. L'*Open Knowledge Foundation* propose un

autre système de classification un peu plus complet :



Un autre critère pour classer les données concerne leur caractère statique ou dynamique. Les **données statiques** (ou « froides ») constituent la majorité des données disponibles ; elles représentent des données vraies à tout moment à partir de leur mise en ligne. À l'inverse, les **données dynamiques** (ou « chaudes ») sont mises à jour en temps réel, car elles changent en permanence. Ce type de données est très présent dans les transports en commun : prochain passage d'un bus à un arrêt donné, disponibilité des stations de vélo en libre service, état du trafic routier...

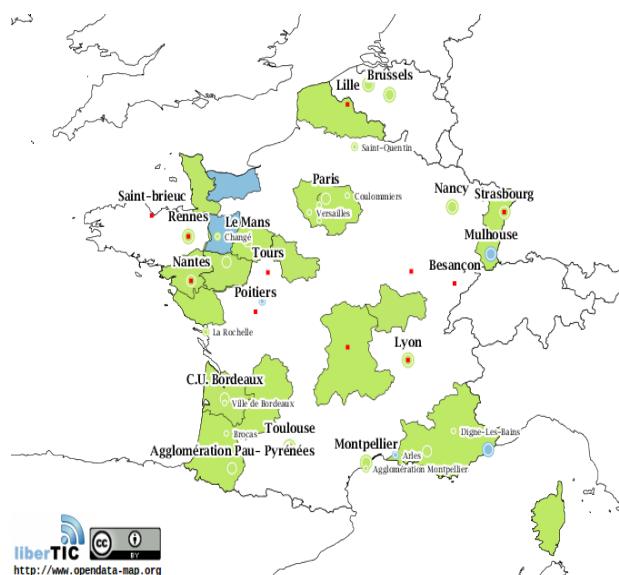
De par leur nature, les données dynamiques ne peuvent être mises à disposition comme de simples fichiers à télécharger, comme c'est le cas pour les données statiques ; elles vont plutôt faire l'objet d'une **API**, ou interface de programmation, qui va permettre leur accès, non seulement par des humains via un navigateur et une URI (adresse

web, ou URL contenant des paramètres spécifiant la demande, par exemple “*horaire du prochain bus de la ligne 62 à l'arrêt Tolbiac*”), mais aussi par des programmes (application, logiciel ou site web).

Même si leur nature dynamique oblige à mettre ces données en ligne sous cette forme, il faut signaler que la mise en place d'une telle API découle forcément de choix de design qui répondront à la question « quelles sont les requêtes auxquelles nous devons répondre ? », chacune de ces requêtes étant accessible via une URL suivie des paramètres. En conséquence, contrairement aux données statiques qui laissent au réutilisateur le soin de les filtrer, les remettre en forme ou les reclassifier avant de les diffuser sous leur nouvelle forme, les données disponibles à travers une API orientent déjà les réutilisations possibles.

Les initiatives existantes en France

Depuis l'impulsion lancée par la création de data.gouv.fr, de nombreuses collectivités françaises se sont lancées dans une démarche d'ouverture de leurs données. De nouvelles initiatives sont annoncées chaque mois, et le chantier est en perpétuelle amélioration. Finalement, certaines collectivités, comme Rennes, s'avèrent être défricheuses et à la pointe de l'innovation par rapport aux actions nationales, ce qui n'est pas surprenant en soi puisque les thématiques abordées et la volonté de se rapprocher de la vie quotidienne des citoyens sont des sujets traités de façon locale dans une France de plus en plus décentralisée.



Légende : Déjà ouvert, En cours, Mouvement citoyen

Source : <http://www.opendata-map.org> (association LiberTIC), le 3 novembre 2013

Le résultat varie grandement selon les régions, et ce sur de nombreux points : quantité des données mises en ligne, fréquence des mises à jour, animation de la communauté, formats des fichiers, licences juridiques, classification des thèmes... Le manque d'homogénéité est flagrant.

	Paris	Rennes	Nantes	Montpellier	Grand Toulouse	Bordeaux
Démographie						
Arts, culture et patrimoine						
Transports et déplacements						
Vie démocratique						
Équipements et services						
Environnement						
Urbanisme et habitat						
Localisation et info géo						
Économie						

Légende :  thématiques dominantes,  thématiques présentes,  thématiques absentes ou très peu représentées

Source : « Comprendre l'ouverture des données publiques », Simon Chignard

Il est intéressant de constater que ces démarches sont à mettre en parallèle avec l'orientation politique : la grande majorité des démarches engagées concernent des localités gérées par des majorités de gauche. Plus intéressant encore, les quelques villes de droite qui ont engagé une démarche d'ouverture (Longjumeau, Saint-Quentin ou Coulommiers), ont adopté la licence Etalab, d'inspiration plus libérale et économique, alors que les autres ont en grande majorité choisi OdBL.

Dans le foisonnement de ces initiatives diverses, nous choisirons comme exemple le portail *opendata71*, le portail de la Saône-et-Loire¹ ouvert au printemps 2011, qui a semble-t-il mûri sa démarche pour proposer une plateforme pérenne et de qualité qui se démarque des autres, en adoptant comme lignes directrices :

- **La volonté de s'adresser à tout un chacun :** le site s'adresse aussi bien aux spécialistes qu'au citoyen lambda, en proposant, en plus des données brutes, des réutilisations internes sous forme de datavisualisations pour aider à la compréhension des données par tous
- **L'importance du “faire ensemble” :** au-delà de l'effet d'annonce et de l'apparence de la transparence, le site propose un « espace partenaires » ; rarement présent sur les autres plateformes, il appelle tous les acteurs de la région à ouvrir et partager leurs données, mais

aussi tous les citoyens à les consulter et à contribuer à la démocratie participative

- La volonté manifeste de considérer cette démarche comme globale et positive pour tous, et d'en faire une opportunité de **moderniser les administrations** en ayant une meilleure visibilité des données à disposition pour aider à la décision.



Ces exigences répondent aux diverses critiques qui ont pu être faites aux initiatives locales, en particulier la difficulté à sortir du cercle des initiés et à se faire connaître des citoyens, et le besoin de sortir de l'effet d'annonce pour construire sur le long terme.

Une autre critique récurrente concerne le manque d'homogénéisation des jeux de données, que ce soit dans leur classification, leur format, leurs licences ou leurs politiques et modes de communication.



Pour pallier à ces manquements, très récemment — en octobre 2013 — s'est créée l'association *OpenData France*. Regroupant une vingtaine de grandes villes et de collectivités territoriales françaises, son but est de mettre en place un guide et une grille de libération des données, pour que

¹ <http://www.opendata71.fr/>

chacun s'y conforme ; c'est un grand pas vers l'uniformisation. L'association constitue aussi un interlocuteur unique pour les entreprises nationales et le gouvernement, ce qui facilite les échanges et prises de décisions.

Suivant le mouvement, le Sénat a aussi ouvert en octobre 2013 sa plateforme, *data.senat.fr*, contenant documents parlementaires, amendements et compte-rendus.

Dans ce panorama des initiatives d'ouverture de données françaises, il convient de ne pas oublier que les administrations ne sont pas les seules à oeuvrer : des entreprises, comme la SNCF ou la Bibliothèque Nationale de France, ont lancé des démarches similaires. Nous reviendrons plus en détails sur les entreprises par la suite.

Les modes de mise à disposition

Comme on l'a vu, les données sont dans la très grande majorité des cas mises à disposition sur le web via un **portail**. Un tel portail consiste en un site web, généralement accessible via un sous-domaine du site web de l'organisation à l'initiative de la démarche. L'usage semble s'orienter vers le choix du sous-domaine « *data* »¹, qui, s'il semblerait au premier abord calqué sur le mot anglais utilisé à l'origine par les instigateurs anglo-saxons du mouvement, s'adapte parfaitement à de nombreux pays européens en raison de son origine latine (le verbe *dare*, signifiant « donner », se déclinant comme nom commun en *datum* et *data*).

Ces portails ont de nombreuses fonctions :

- Regrouper toutes les données disponibles au même endroit, et proposer des classifications et filtres pour permettre un accès facile
- **Augmenter la visibilité** de ces données, via le référencement naturel, mais aussi grâce au positionnement éditorial du portail, qui permet de communiquer sur le positionnement politique de l'instigateur, voire de démocratiser le mouvement open data dans une optique pédagogique
- **Permettre une animation** continue autour de ces données, via des fonctions dynamiques comme des forums, la possibilité de noter ou commenter les jeux de données, voire de les corriger dans le cas des portails les plus évolués
- Mettre en avant et **promouvoir les réutilisations** réalisées par des tiers à partir des données présentes.



Ces caractéristiques étant à la fois assez générales — non inhérentes aux types de données proposées — et largement partagées par les portails existants, une solution a rapidement émergé pour répondre aux besoins techniques des détenteurs de données : le projet open source CKAN est un framework de création et de gestion de catalogues de données, conçu spécialement pour l'open data par l'Open Knowledge Foundation, et qui motorise aujourd'hui de nombreux portails gouvernementaux, comme *data.gov.uk* et *data.gov*. Etalab a signalé sa volonté de l'utiliser pour la prochaine version de *data.gouv.fr*. Bien sûr, tout est possible

¹ data.gouv.fr, data.rennes-metropole.fr, etc.

et de nombreux portails reposent sur des technologies plus répandues et généralistes comme les CMS *Drupal* ou *Typo3*.

Nous avons aussi vu que les données peuvent être mises en ligne sous plusieurs formes. La plus classique est bien sûr le fichier de données en téléchargement direct. Puisque les données sont, dans la grande majorité des cas — et comme recommandé par la communauté open data — organisées, les formats adoptés sont souvent des formats de données tabulaires :

- Le **CSV** (*Comma-Separated Values*) est un format standard, ouvert et facile à utiliser ; les fichiers .csv sont des fichiers texte dans lesquels chaque ligne de texte représente une ligne d'un tableau, contenant les données séparées par des virgules. Ils sont ainsi très faciles à lire à l'oeil nu, mais aussi à parcourir dans le cadre d'un algorithme. Tous les tableurs modernes permettent aussi d'importer des fichiers CSV pour les afficher sous forme tabulaire. Pour toutes ces raisons, les fichiers CSV représentent la majorité des données statiques disponibles sur les portails open data.
- Parfois, les données sont aussi mises à disposition au format **XLS**, le format propriétaire d'Excel (Microsoft). Excel étant disponible sur la majorité des ordinateurs, on parle dans ce cas de *standard de fait*.

Il peut arriver cependant que la forme de tableau à deux dimensions ne suffise pas à représenter la complexité des données en un seul fichier ; dans

ce cas, on va utiliser des formats tels que le **XML** (*Extensible Markup Language*), qui permet d'organiser les données à l'intérieur de balises de façon imbriquée et détaillée. Moins souvent, on utilisera du **JSON** pour représenter ce type de données complexes de façon statique.

Le cas particulier des données géographiques nécessite l'utilisation de formats spécifiques :

- Le **KML** (*Keyhole Markup Language*) est un format ouvert, qui repose sur XML et qui permet de représenter des données spatiales de façon standard. Il est possible de l'ouvrir avec des logiciels connus du grand public, comme Google Earth, ce qui le rend à la fois accessible à tous pour une simple consultation et adapté à un usage avancé¹.
- Les **Shapefiles** (**SHP**) sont un autre format moins utilisé, mais qui permet de représenter de façon très fine des objets géométriques (points, lignes, polygones) qui composent des cartes.

Bien souvent, les données spatiales sont aussi représentées par une carte interactive sur la fiche du jeu de données, à côté des fichiers en téléchargement.

Bien sûr, l'idéal est de proposer plusieurs formats par jeu de données, afin de satisfaire le plus de réutilisateurs possible.

Nous avons vu que les fichiers en téléchargement ne sont pas le seul mode de mise à disposition des données : l'accès via une **API** est indispensable pour les données temps-réel ou mises à jour

¹ Il est aussi possible de trouver des fichiers .kmz, qui contiennent, en plus des données du KML, des images et des informations graphiques à faire afficher par les logiciels.

fréquemment, mais est aussi possible pour les données statiques, même si ce cas est beaucoup plus rare. L'*Open Knowledge Foundation* recommande de toujours fournir des fichiers en téléchargement en parallèle de l'API ; en effet, en plus d'orienter les réutilisations en proposant à priori des manières d'accéder à certaines données, le principe de l'API empêche d'accéder simplement à la totalité des données pour en faire une visualisation, ou même pour une simple consultation.

Bien sûr, l'API sera augmentée de sa documentation détaillée.

Mais un jeu de données, si détaillé, précis et de qualité soit-il, n'est rien sans des **métadonnées**. Ces « données sur les données », associées au fichier, vont permettre de le catégoriser (et ainsi de le retrouver facilement) mais vont aussi fournir un contexte de lecture. Parmi ces données complémentaires, on en dénombre quelques-une qui sont indispensable dans le contexte d'ouverture de données publiques :

- Le titre et la description détaillée du jeu de données
- Les thématiques associées
- La liste des formats de fichiers disponibles
- La date de dernière mise à jour, voire la fréquence de mise à jour si cela s'applique
- L'émetteur des données (par exemple l'INSEE, l'ARCEP ou le ministère de l'Éducation)
- La licence qui s'applique
- Le territoire concerné par le jeu de données (et ses limites géographiques)

- Dans le cas des données géographiques, il est indispensable de préciser la projection et le référentiel utilisé (Lambert 93, WGS84, CC47 sont les plus utilisés), et le degré de précision

Un moyen simple de mettre à disposition ces métadonnées de façon organisée est de maintenir un « métafichier », qui liste les jeux de données disponibles sur le portail, avec leurs métadonnées associées.

Les résultats des réutilisations

Les réutilisations des données ouvertes sont très diverses, par leurs formes, leurs auteurs et leurs motivations. Elles peuvent néanmoins se classer en quatre grandes catégories :

La première est la **consultation directe**, ou le fait de lire les données en direct par le grand public, souvent pour des motivations personnelles. Ici, le réutilisateur et l'utilisateur final sont la même personne. Cela concerne des jeux de données simples, comme la liste des prénoms les plus choisis pour les naissances sur un territoire, ou les actes d'état civil.

Plus visibles dans l'espace public, les réutilisations de type **datavisualisations** construites à partir des données, et conçues pour aider le grand public à appréhender des données parfois complexes. L'utilisateur final n'accède plus à des données brutes, mais à une interface qui va lui permettre de visualiser et manipuler des données choisies. Deux types de réutilisation font partie de cette catégorie. Le premier, déjà évoqué, concerne les détenteurs de données qui proposent, conjointement aux données brutes, des visualisations sur

leur portail (le détenteur est alors confondu avec le réutilisateur).

Mais de plus en plus de datavisualisations sont produites et diffusées par des journalistes. On parle beaucoup actuellement de **datajournalisme** (ou journalisme des données en français, quoi que l'expression est rarement utilisée), mouvement qui considère que le journalisme peut (et doit) s'enrichir de l'exploitation de données brutes pour rendre compte du réel. Si le fait d'utiliser des données pour rapporter des faits a toujours été à la base du journalisme, le datajournalisme a favorisé très fortement ce mouvement dans les dernières années, grâce à la mise à disposition des données en open data et à la montée en puissance des médias en ligne, et encouragée par la défiance actuelle des citoyens envers les autorités.

Le Guardian et le New York Times proposent sur leur site de nombreuses datavisualisations¹, et font aujourd'hui figure de pionniers dans le domaine. Les applications françaises sont encore rares, et sont encore l'apanage de pure players plutôt que de médias installés, ce qui pourrait s'expliquer par le fait que ces structures ont une culture scientifique et informatique plus forte par nature, et ont déjà au sein de leur équipe des développeurs et des journalistes sensibilisés aux questions numériques. Il reste compliqué de mettre en oeuvre le datajournalisme dans les rédactions, car cela demande de travailler avec des métiers très différents (graphistes, développeurs, statisticiens...), pour un travail énorme en amont (collecte, traitement, développement), pour un résultat qui ne peut être prévisible à l'avance, ce qui reste difficile à valoriser.

ser dans un monde où l'on attend surtout des médias en ligne qu'ils soient réactifs et presque temps réel.



Mais les réutilisations qui sont de loin les plus médiatisées et les plus plébiscitées par le grand public sont les **applications**. Ici, les réutilisateurs sont assez spécialisés, puisqu'il s'agit de développeurs, et les utilisateurs finaux représentent une très large communauté. Ces réutilisations utilisent le plus souvent des données de la vie quotidienne : équipements urbains, transports... Deux cas se présentent : souvent, l'application accède aux données à distance via l'API créée par l'émetteur original des données, ce qui permet non seulement d'utiliser des données temps-réel qui ne sont accessibles que par ce biais, mais aussi d'alléger les applications en ne gardant pas les données « en dur » dans la mémoire du téléphone, et d'éviter de devoir mettre à jour l'application à chaque mise à jour des données.

Contrairement à ce que l'on imagine au premier abord, on ne parle pas seulement ici d'applications mobiles, mais aussi d'applications web (ou *web apps*). Faire la distinction serait de toute façon inutile, puisqu'on parle ici d'applications accessibles depuis un terminal, fixe ou mobile, qui ont un but précis et utilisent des données ouvertes.

Les possibilités offertes par les terminaux mobiles, comme par exemple la géolocalisation ou les notifications push, vont permettre de tirer parti au mieux de ces données, et de les présenter de fa-

¹ Le *Guardian* a même lancé en 2009 sa propre plateforme de diffusion de données, augmentée d'une API permettant de construire ses propres applications : <http://www.theguardian.com/open-platform>

çon adaptée et personnalisée aux utilisateurs.

À propos des applications mobiles développées pour iOS, Android ou autres Windows phones, il est assez ironique de constater que ces applications font souvent partie d'un microcosme très fermé et contrôlé (en particulier dans le cas d'iOS), ce qui est en très forte contradiction avec la philosophie open data.

L'open data et les entreprises privées

Comme on l'a vu jusqu'ici, l'ouverture des données concerne pour l'instant surtout le secteur public. À mesure que l'opinion publique s'habitue à l'idée de cette ouverture — voire la réclame — et que les autorités en découvrent les bienfaits et l'étendue des possibles, les entreprises, en toute logique, observent le mouvement et envisagent la possibilité d'en faire partie. Bien sûr, il faut comprendre les différences qui séparent le secteur public et les entreprises privées, afin de dégager les enjeux pour ces dernières.

Les entreprises sont productrices et utilisatrices de données depuis toujours : données clients, documents de gestion, voire même des données au centre du modèle économique pour certaines sociétés de service. Elles échangent aussi des données internes régulièrement avec leurs partenaires et sous-traitants, voire avec leurs concurrents, dans un cadre contractuel bien encadré et à des fins d'analyse de leur marché.

En revanche, les clients, et le public de façon générale, semblent oubliés dans ces démarches.

Un contexte favorisant l'ouverture

L'idée d'ouvrir ses données pour créer de la valeur n'est en fait pas nouvelle au sein des entreprises, en particulier des entreprises du web, qui souvent, mettent en place une **semi-ouverture** par le biais de leurs **API**. Par exemple, Amazon permet, via son interface de programmation « *Product Advertising API* », d'intégrer dans un site web ou une application mobile des contenus en vente sur le site *amazon.com*, sans accéder au site source. En fait, au lieu de mettre à disposition ses données en accès libre au public sans présupposer des réutilisations, Amazon fournit plutôt un moyen très contrôlé et orienté d'y accéder, dans un but précis : celui de créer des points d'entrée supplémentaires vers leur service, et ainsi d'augmenter ses activités ; on est en fait plus dans le cadre d'une ouverture de services que d'une ouverture de données.



Les entreprises du web sont aujourd'hui très nombreuses à fournir une telle API : Google Maps, Twitter, Facebook ou Flickr sont les plus connus, et ont fait évoluer l'idée que, même si elle reste contrôlée et limitée, l'ouverture est bénéfique, et peut créer de la valeur. En permettant de dissocier les données (le contenu) et l'interface qui les présente (le contenant), elles ont aussi mis en lumière

le fait que les données sont l'épine dorsale de leur chaîne de valeur. Plus nouveau, le fait d'accepter qu'autrui puisse mettre en valeur ses propres données de façon différente, voire meilleure que la sienne, fait son chemin¹, et avec lui le fait de considérer les destinataires de son activité comme des partenaires potentiels capables d'améliorer leur service.

Quelques entreprises plus traditionnelles commencent à adopter ce principe d'ouverture de leurs

services via des API ; ainsi, le Crédit Agricole a créé le « CA Store », qui permet aux développeurs de créer des applications qui elles-mêmes auront accès aux données bancaires des clients.

On est cependant encore loin de l'open data à proprement parler, puisque ces entreprises sécurisent en encadrent encore les réutilisations possibles, à l'inverse de la philosophie originelle de l'open data qui veut que l'on ouvre ses données sans présupposer des réutilisations. Pourtant, une fois la problématique posée et les enjeux bien identifiés, il y aurait des bienfaits évidents pour les entreprises à ouvrir certaines de leurs données. Nous en listeront ici les plus importants.

La transparence pour améliorer la confiance

Dans la même veine que l'open government, un des bénéfices les plus évidents, au premier abord, de l'ouverture des données d'une entreprise à ses clients est un **bénéfice d'image**. Bien sûr, le ris-

que existe de dévoiler des défauts au public, les initiatives sont donc encore rares. Il faut évoquer en particulier celle de la Lyonnaise des Eaux, qui met en ligne depuis juin 2012 l'origine et les résultats d'analyses de l'eau dans ses communes desservies, ou Nike qui diffuse la liste complète de ses usines dans le monde, pour contrer les accusations d'exploitation dans les pays en voie de développement. Mettons un bémol toutefois, en signalant que ces données ne sont ni formatées, ni accessibles via une URI, donc difficiles à réutiliser sans un lourd travail préalable de collecte et de traitement. Les entreprises sont ici encore dans une logique de service direct à l'utilisateur final, en oubliant peut-être un peu les intermédiaires qui pourraient utiliser et représenter ces données sous une autre forme.



Une opportunité d'innovation

Un autre pilier fondamental de l'open data est, comme on l'a évoqué plus haut, la **création de valeur économique** grâce à la réutilisation des données par autrui, pour créer de nouveaux services bénéfiques à l'émetteur des données, car ils faciliteront son usage. Bien sûr, c'est un argument massue pour le secteur privé, et peut-être celui qui offre le plus de promesses : en effet, quelle meilleure façon de valoriser ses données que de collaborer avec des particuliers ou avec d'autres entreprises pour inventer de nouveaux services ?

Plus loin encore, les entreprises ont à gagner à s'inscrire dans le mouvement open data, car plus

¹ De nombreuses « surcouches » sont apparues ces dernières années, utilisant en particulier les API de services de photographie, qui ont pour but de mettre en valeur le contenu de façon plus intéressante que le site original ; par exemple, ihardlyknowher.com se présente comme une alternative minimaliste à Flickr, alors que statigr.am ajoute des statistiques aux classiques galeries que l'on peut trouver sur le site ou l'application Instagram.

elles seront nombreuses à diffuser leurs données, plus les résultats des réutilisations seront complets et novateurs, et créeront de la valeur ajoutée pour les utilisateurs.

C'est aussi une possibilité pour les grandes et anciennes entreprises, traditionnellement isolées dans leurs initiatives et fonctionnant beaucoup en « circuit fermé », de s'ouvrir à un écosystème d'acteurs innovants, plus flexibles, jeunes et à l'écoute du marché, et de s'enrichir de leur expertise.

Rendre leurs données aux utilisateurs

Parmi les réticences des entreprises, on identifie la problématique des données personnelles, interdites à la diffusion, et qui constituent une grande partie du capital de données des entreprises. Cependant, s'il est impossible de les diffuser au public, un mouvement très récent promeut le partage de ces données, non pas de façon publique, mais de façon directe aux intéressés : les clients eux-mêmes concernés par leurs données, qui leurs sont personnelles et dont ils sont à l'origine. En effet, tout comme le citoyen doit avoir accès aux données créées dans le cadre des missions de service public financées par ses impôts, pourquoi ne pourrait-il pas connaître celles qui sont créées par les entreprises à son sujet ? Les débats publics très nombreux sur la confidentialité des données que nous délivrons aux réseaux sociaux — avec Facebook en ligne de mire — ont installé dans les consciences le fait que nous sommes devenus nous-mêmes le produit, et que nous créons de la valeur pour des entreprises en leur fournissant nos données privées. Seulement, si nous sommes si

nombreux sur Facebook, c'est précisément parce que ce dernier nous donne les moyens d'utiliser, mettre en forme, de valoriser et de diffuser ces données. À l'inverse, les entreprises de service plus traditionnelles font un usage intensif des données de leurs clients, pour des raisons marketing et de suivi clientèle ; la philosophie du **consumer empowerment** (prise de pouvoir par le consommateur) veut changer cet état de fait, et pousser les entreprises à rendre leurs données aux particuliers. On parle déjà du passage du *CRM* (gestion de la relation client) au *VRM* (*Vendor Relationship Management* ou gestion de la relation commerçant). Les données les plus concernées par ce débat sont, en quelques sorte, parmi les plus sensibles des données personnelles : informations bancaires, santé, voire même données de localisation dans le cas des transports en commun...

Les bénéfices sont nombreux pour les deux parties : outre le renversement salutaire de la relation de pouvoir entreprise/client, cela donne une autonomie appréciable au consommateur — qui est plus à même d'utiliser les services comme il l'entend et au mieux de son efficacité — et favorise bien entendu l'image de l'entreprise, tout en encourageant le consommateur à partager plus de données le concernant.

La plupart des fournisseurs d'eau ou d'électricité, ou des opérateurs téléphoniques fournissent déjà, via des applications dédiées, des moyens de visualiser ces données personnelles, sans pour autant permettre de les récupérer sous une forme ordonnée et réutilisable dans un autre contexte.

Ce mouvement est à mettre en parallèle avec celui du **quantified self**, ou quantification de soi, qui de façon générale, regroupe les principes et outils

destinés à l'utilisation autonome de ses données personnelles. Son croisement avec les objets connectés a donné naissance à de

nombreux produits liés à la santé et au sport, avec par exemple les produits (bracelet, balance, applications mobiles) *Fitbit*, qui mesurent les efforts sportifs et en rendent compte dans une interface de statistiques personnelle détaillée, ou le célèbre FuelBand de Nike.

Certains développeurs ou artistes ont poussé plus loin ce concept, à l'instar de Naveen Selvadurai, le fondateur de Foursquare, qui a mis à disposition sur le web son API personnelle, qui suit au jour le jour ses activités géolocalisées et datées, son poids...

L'entreprise réutilisatrice des données

Mais l'entreprise n'est pas uniquement un possible émetteur de données ; en fait, l'émergence de l'open data a aussi créé un nouveau secteur, celui des intermédiaires de données. C'est le cas, par exemple, de *Data Publica*, une société de service français qui a développé comme expertise la collecte, le formatage et la mise à disposition des données, en parallèle avec un accompagnement des émetteurs, mais qui propose aussi sur son portail de très nombreux jeux de données agrégés depuis des sources diverses.

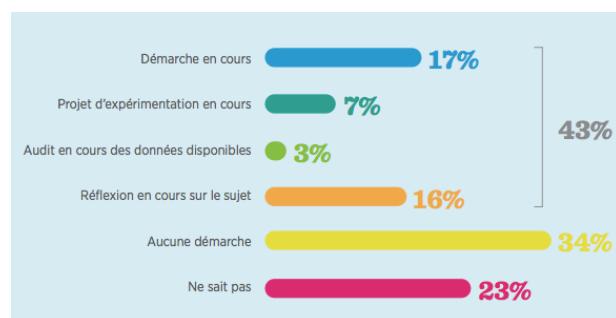
Plus en aval de la chaîne de réutilisation, des sociétés se sont développées avec comme expertise celle de la représentation visuelle des données ; c'est le cas par exemple de *Dataveyes* ou de *Mapize*.



Aussi, les startups et sociétés de service en informatique ont bien compris le potentiel des jeux de données disponibles en open data, et sont nombreuses à les utiliser pour développer des applications innovantes. Nous reviendrons en détail sur celles-ci, puisque les applications les plus demandées par le public, et par conséquent les plus développées, sont les applications de transport.

Des réticences compréhensibles

Nous avons listé des initiatives assez isolées, qui sont pour l'instant, pour la plupart, au stade de l'expérimentation. En réalité, les entreprises, qui ne sont soumises à aucune obligation par la loi d'ouvrir leurs données au public, sont encore défiantes, mais néanmoins curieuses ; ainsi, d'après une étude Bluenove de novembre 2011, presque la moitié des responsables d'entreprises sont convaincus de l'intérêt à ouvrir leurs données, et 43% ont entamé une démarche, même minime, en ce sens.



« Des projets d'ouverture de données en entreprise encore en phase amont », source : Bluenove

Les raisons qui freinent les entreprises sont diverses et complexes ; parmi elles la complexité juridi-



que (en particulier vis-à-vis des données personnelles) et le manque de compétences en interne, la peur de ne pas pouvoir contrôler suffisamment la qualité des réutilisations ou de dévoiler des dysfonctionnement internes.

Dans tous les cas, le mouvement open data ayant atteint, sinon une maturité, au moins un stabilité dans le secteur public, les activistes du mouvement s'accordent aujourd'hui sur le fait que la prochaine étape à franchir sera indubitablement l'adoption de ces principes par les entreprises.

David Eaves, conseiller de plusieurs administration pour l'ouverture des données, clamait fin 2011 lors de son discours d'ouverture de l'Open Data Camp de Varsovie¹, l'importance du défi pour les années à venir : « *Malgré nos succès, nous sommes loin d'avoir atteint une masse critique. Nous saurons que nous sommes en train de progresser lorsque des entreprises – grandes et petites – tout autant que des associations sans but lucratif, commenceraient à comprendre combien les données de gouvernance ouverte peuvent rendre le monde meilleur et voudront ainsi nous aider à faire progresser la cause.* »

¹ <http://eaves.ca/2011/10/21/the-state-of-open-data-2011/>

Les données au cœur du transport public

Comme nous l'avons entrevu plus haut, la mobilité occupe une place importante dans le mouvement d'ouverture des données. En effet, dans un monde de plus en plus urbanisé, quelle problématique est aujourd'hui plus quotidienne et répandue que celle des transports ? Plus que toutes les autres thématiques, celle-ci touche autant les réutilisateurs, qui sont très demandeurs de données exploitables, que le grand public, qui comme nous le verrons a besoin de solutions créatives, pratiques et utilisables au quotidien. Au-delà de ces aspects très pratiques, on touche, avec les transports, à un sujet presque politique, puisqu'il touche à l'appropriation du territoire par les citoyens.

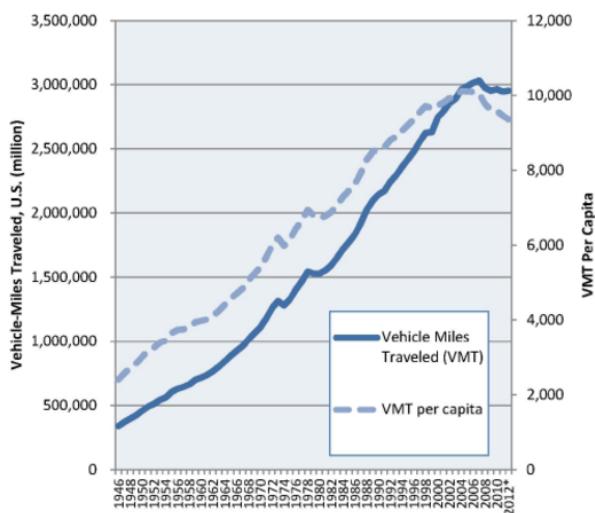
Nous allons voir que le sujet est d'une grande complexité, que ce soit au niveau des données elles-mêmes, mais aussi des acteurs en jeu, qui ont des statuts et des motivations très divers. Nous n'oublierons pas d'analyser ces subtilités en gardant à l'esprit une vision très « grand public » de cette problématique, tournée vers l'amélioration des solutions offertes aux usagers des transports que nous sommes, en y apportant un éclairage que nous espérons large et tourné vers l'avenir.

Faisons dès maintenant une mise au point sur les termes employés : si nous avons beaucoup parlé, jusqu'ici, de « *transports en commun* », il faut bien différencier ceux-ci du « *transport public* », plus

large, qui inclut aussi des solutions individuelles (transport à la demande, vélo...) mises en places par les pouvoirs publics. Même si ce n'est pas le sujet central de la présente thèse, nous nous devrons aussi d'évoquer le « *transport individuel* », solution privée (souvent la voiture), qui subit lui aussi de profondes mutations qui influent naturellement sur l'ensemble des enjeux de la mobilité.

La mutation actuelle des transports

On a très longtemps considéré la voiture comme l'aboutissement d'une société industrialisée, individualiste et complexe, et par conséquent comme le mode de transport « par défaut ». Cependant, aujourd'hui, cette hégémonie est en train de changer sous nos yeux : dans un climat de crise économique durable et de montée des prix des carburants, sensibilisés aux enjeux écologiques menant à une volonté de limiter les émissions de gaz à effet de serre, nous sommes progressivement en train de délaisser la voiture. Alors que le véhicule individuel avait traditionnellement un rôle de marqueur du statut social, il perd de son attrait dans les pays aisés : l'obtention du permis de conduire n'est plus perçu comme un rite de passage indispensable à l'accès à la vie adulte, et les jeunes connectés et urbains sont moins nombreux à le passer que par le passé. Aux États-Unis, pays de la liberté individuelle, de la mobilité tout au long de la vie et des éternels *road trips*, le nombre de kilomètres parcourus à l'année est même en décroissance depuis quelques années.



Nombre de miles parcourus par an en voiture aux États-Unis depuis 1946.

Source : « [A New Direction, Our Changing Relationship with Driving and the Implications for America's Future](#) », US Pirg

Bien sûr, le besoin de mobilité n'est pas moindre que par le passé ; il est même plus important, mais surtout différent, car il accompagne un changement rapide des modes de vie :

- La vie de famille, tout d'abord, est complexifiée par de nombreuses activités, menées toute la semaine pour tous les membres de la famille ; on assiste aussi à une explosion des familles recomposées, éclatées géographiquement, qui se déplacent régulièrement pour se retrouver.
- Le travail aussi change en profondeur : la généralisation des contrats précaires oblige à changer ses habitudes de transport quotidien régulièrement, mais aussi à accepter de travailler de plus en plus loin de son domicile. Les déplacements professionnels perpétuels sont de plus en plus fréquents, et la flexibilité des horaires et des lieux se généralise, avec le télétravail (total ou partiel) et l'utilisation d'espaces de coworking en location à la journée.

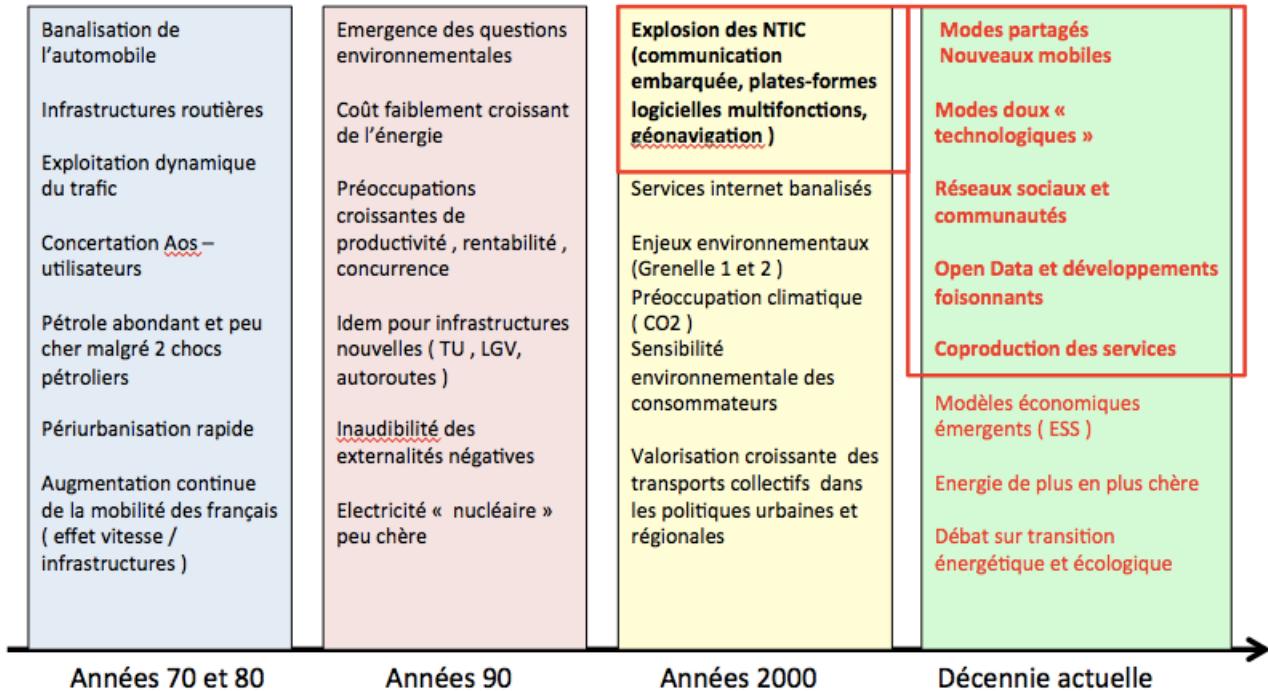
- Même les habitudes de tourisme ne sont plus les mêmes : les voyages sont moins longs, mais plus fréquents, et les destinations par conséquent plus proches. Les résidences secondaires sont souvent des résidences de week-end.

Par conséquent, la demande sur les transports en commun est forte, et les opérateurs ont, plus que jamais, le devoir de répondre de façon efficace et sûre à cette affluence.

En parallèle, on assiste à une explosion de l'utilisation de solutions collaboratives, comme les services de covoiturage, facilités par la démocratisation des smartphones et par l'utilisation d'applications permettant de planifier ces voyages en temps réel.

Cette mutation complexe de la mobilité est à penser dans le contexte de la profonde révolution numérique que nous sommes en train de vivre. Le web a déjà profondément changé le monde depuis son apparition dans les foyers dans les années 90, en apportant à tout un chacun la possibilité d'accéder au savoir global, et par extension à un possible renversement de gouvernance par le citoyen (entre autres grâce à des mouvements organisés comme celui de l'open data), mais aussi en lui offrant l'accès à des services innovants et pratiques.

Mais aujourd'hui, alors que 18 millions de français accèdent quotidiennement au web depuis leur smartphone, qu'en est-il de notre vision de cette planète parallèle qu'était le web des terminaux fixes ? À l'heure où les usages sont multiples et où l'on accède au web dans n'importe quelle situation du quotidien, le web n'est plus envisagé comme un monde externe auquel on accède ponctuellement depuis des points d'entrée bien définis :



Les nouveaux déterminants de la mobilité, tels que perçus par la SNCF

Source : [« Quelle SNCF pour demain ? »](#), LeMonde.fr, 1er oct. 2013

comme le dit Michel Serres dans son essai « *Petite Poucette* », nous vivons à l'intérieur de notre mobile, en symbiose continue avec l'univers numérique auquel il nous ouvre l'accès.

Les smartphones ne sont pas nos seules portes d'entrée au web ; on parle beaucoup du web 3.0 comme étant celui des objets connectés, et bien sûr, le web étant susceptible d'être embarqué partout, la voiture n'y échappe pas, et on assiste à la montée de ce qui est appelé par les constructeurs la *voiture connectée*.

Mais surtout, les smartphones nous permettent de combler et d'accompagner un moment de vie quotidien, qui était auparavant considéré comme un temps mort obligatoire, coupé de tout : le trajet, spécifiquement celui effectué en transport en commun, qui peut devenir un moment de vie réel et utile. L'usager des transports en commun est aujourd'hui relié au monde.

De plus, même lorsque nous sommes en dehors des moments de transport, nous avons désormais l'accès continu à des solutions nous permettant de prévoir, d'organiser et optimiser ces trajets : le smartphone nous donne la possibilité de devenir des acteurs de notre propre mobilité, de nous apprêter le territoire et nos déplacements.

Dans cette évolution profonde des transports à l'heure du numérique, plusieurs composantes sont à prendre en compte pour avoir une vision à 360 degrés, comme les infrastructures, les véhicules, l'énergie... Nous allons nous intéresser par la suite à celle qui se trouve à la croisée des chemins, entre les autorités de transports, les créateurs d'application et les usagers des transports en commun : **la donnée**. Les données du transport public peuvent être considérées, de toute évidence, comme des données publiques ; nous verrons par la suite que ce n'est pas si simple, et que la libéra-

tion des données de transport reste complexe, contradictoire et controversée. Le débat est d'autant plus intéressant qu'il concerne presque tout le monde et que les promesses de nouveaux services sont grandes. Contrairement à la plupart des autres jeux de donnée publique, les (ré)utilisateurs sont conscients de ce potentiel et la demande est très forte.

Mais l'ère de la donnée, si elle ouvre une gouvernance pour les utilisateurs, crée aussi des opportunités pour les collectivités, les entreprises et l'environnement, qui se doivent d'accompagner ces mutations de société, en proposant des solutions adaptées.

De quelles données parle-t-on ?

Il est difficile de faire un état des lieux exhaustif des données du transport en commun susceptibles d'être ouvertes, tant elles sont nombreuses et de natures diverses. Nous allons néanmoins en tenter une classification.

On l'a vu, on envisage souvent l'ouverture des données comme un geste de **transparence** envers les citoyens de la part des autorités. Dans cette optique, de nombreuses données de transport sont concernées, et les autorités de transport ont tout intérêt à faire preuve d'ouverture sur l'état des leurs réseaux et de leurs équipements, afin de (re)créer la confiance sur la sécurité des usagers. Pourtant, elles sont rares à diffuser ce genre de données, encore souvent considérées comme des

données internes. Les nombreuses catastrophes ferroviaire de 2013 ont ému les citoyens, et en France, celle de Brétigny-sur-Orge le 12 juillet a abouti à de nombreux débats sur l'état du matériel de la SNCF, et *in fine* au lancement d'une pétition réclamant l'ouverture de ces données en open data¹.

Dans une logique de transparence, on pense aussi aux statistiques de rendement et d'efficacité, en particulier sur la ponctualité des véhicules de transport en commun, ou sur le taux de fréquentation (voire de saturation, parfois), de ces véhicules. Alors que pendant des années ces statistiques étaient diffusées par le biais de communiqués de presse chaque année, ces deux jeux de données sont proposés chaque mois en open data sur le portail de la SNCF.

Un autre type de données important dans le domaine de transports publics concerne les **informations commerciales**. Visant plutôt les usagers en tant que destinataires directs, ces données peuvent être par exemple l'emplacement, les horaires, et parfois la fréquentation des points de vente, ou la liste exhaustive des tarifs et formules d'abonnement.

Mais les informations les plus cruciales, et celles qui font du transport un des domaines les plus attendus sur l'ouverture des données, sont bien les données concernant les **services** de transport en commun à proprement parler :

- Plan détaillés des réseaux et des lignes, géolocalisation des stations et arrêts
- Horaires (théoriques ou réels).

¹ <http://www.change.org/transparencessurlesrails>

Un dernier grand type de données concerne les données d'**accessibilité**, en particulier pour les personnes souffrant de handicap, mais aussi pour tous les usagers qui ont besoin d'informations précises sur la configuration des gares (poussettes, vélos...). Nous verrons qu'elles prennent une place de plus en plus importante dans les jeux de données ouvertes de transport public, et que c'est un moteur important dans les démarches d'ouverture.

Au final, on constate que les données concernées sont extrêmement diverses, que ce soit par la motivations de leur émetteur ou leurs destinataires, mais surtout par leur nature. En effet, elles comprennent des données très **statiques** (points de vente, statistiques annuelles, plans de réseaux), mais aussi extrêmement **dynamiques**, comme les horaires temps-réel et les alertes de perturbation, qui se doivent d'être d'une grande réactivité (à la minute près). Par ailleurs, l'autre caractéristique qui en fait un secteur passionnant, est le fait que les données sont globalement à la fois très **spatiales** et très **temporelles**, et ceci à un niveau de détail, et donc de complexité, très élevé.

Représentation des données fondamentales de transport

S'il est souvent assez facile de mettre en forme et ordonner des données tabulaires simples pour les préparer à la diffusion, il faut entrer un peu plus en profondeur dans la théorie des transports pour saisir que la représentation formelle de la plupart des données de transport relève de problématiques complexes.

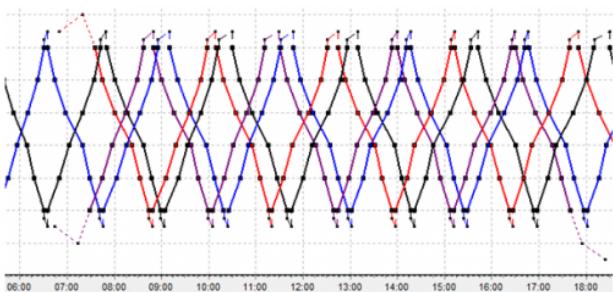
En effet, ces données doivent répondre à plusieurs exigences, en étant assez précises et exhaustives — comme le demande la philosophie open data — pour permettre des réutilisations de qualité, tout en restant compréhensibles par tous, que ce soit les réutilisateurs ou les usagers.

Les réseaux de transports sont d'une extrême complexité : à titre d'exemple, le réseau de la RATP contient 16 lignes de métro, 353 lignes de bus, plus de 12000 points d'arrêt, et pour chaque ligne, tous moyens confondus, entre 3 et 12 fiches horaires différentes, selon les périodes et les heures. Les opérateurs de transport en commun ont depuis longtemps développé des outils et des méthodes spécifiques pour manipuler cette complexité.

Paris Gare de Lyon	5:36	5:51	6:05	6:16	6:21	6:31	6:36	6:46	6:51
Châtelet Les Halles	5:40	5:55	6:09	6:20	6:24	6:34	6:39	6:49	6:54
Auber	5:42	5:57	6:11	6:22	6:27	6:37	6:42	6:52	6:57
Charles de Gaulle Étoile	5:45	6:00	6:14	6:25	6:30	6:40	6:45	6:55	7:00
La Défense Grande Arche	5:50	6:05	6:19	6:29	6:34	6:44	6:49	6:59	7:04
Nanterre Préfecture	5:52	6:07	6:21	6:32	6:37	6:47	6:52	7:02	7:07
Houilles Carrières sur Seine	5:56	6:11	6:25	6:36	6:41	6:51	6:56	7:06	7:11
Sartrouville	6:00	6:15	6:29	6:40	6:45	6:55	7:00	7:10	7:15
Maisons Laffitte	6:02	6:17	6:31	6:42	6:47	6:57	7:02	7:12	7:17

Source : extrait de la fiche horaire du RER A valable du 1er septembre au 14 décembre, du lundi au vendredi

Ainsi, derrière une simple fiche d'**horaires**, il faut comprendre que des données et des calculs complexes ont été mis en jeu : l'établissement de ces fiches provient d'un travail de *graphicage*, puis d'*habillage*, construit à partir des données spatiales des arrêts, des données temporelles (vitesse) des véhicules, et des contraintes d'exploitation des lignes (conducteurs, horaires...).



Étape de graphique d'une ligne de bus. En abscisse, le temps en minutes, en ordonnées les arrêts.

Ce travail va aboutir à l'établissement des plannings de circulation détaillés pour chaque ligne, et pour chaque période d'application (plein trafic, vacances scolaires). Les fiches horaires ne sont donc qu'une projection de cette vision métier, orientée vers le service et les usagers. Ces graphiques de circulation étant en réalité des outils très mouvants (pour certaines lignes, ils changent tous les jours !), les fiches horaires résultantes sont donc des approximations, puisqu'elles ne sont mise à jour qu'en moyenne une fois par semestre.

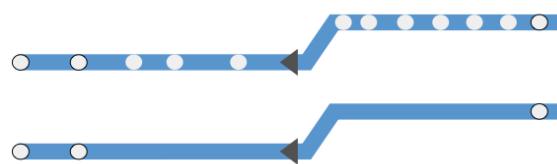
De la même façon, si l'usager des transports en commun se représente un trajet de façon assez simple (un arrêt de départ et un arrêt d'arrivée), la théorie qui se cache derrière est bien plus évoluée. En effet, on distingue dans le monde des transports plusieurs concepts pour décrire les trajectoires via le réseau :

- Une **ligne** est un sous-ensemble du réseau qui est représenté par un sigle commercial (ligne de bus 68 ou ligne de métro 4 par exemple). Géographiquement, c'est un graphe — orienté car certaines parties de lignes ne sont faites que dans un sens, aller ou retour — reliant des points dans l'espace, qui peut contenir des ramifications (dans le cas par exemple des lignes de métro contenant des fourches à leur extrémité, comme la ligne 7 ou la ligne 13), ou

même des cycles (dans le cas de la ligne de métro 7bis, et de nombreuses lignes de bus). Elle est très souvent représentée, pour le grand public, par ce que l'on appelle un *thermomètre* (ici celui de la ligne de bus 174 de la RATP) :



- Une **route** est une sous-partie de la ligne, d'une station vers une autre et dans un sens orienté. C'est ce que se rapproche le plus, pour un usager, du concept de trajet.
- Une **mission** est identique à une route, mais prend aussi en compte les arrêts desservis durant le parcours : par exemple, il est possible de faire deux missions différentes pour relier la Gare du Nord et l'aéroport Charles de Gaulle via le RER B :



- Une **course** est une mission située dans le temps, c'est-à-dire avec un horaire appliqué à chaque arrêt : c'est une colonne dans une fiche horaire.

De même, alors que l'usager envisage le concept de station — ou d'arrêt — de manière relativement

simple, nous manipulons plusieurs concepts pour les décrire côté métier :

- Une **zone d'arrêt** est un concept très orienté public : c'est une zone, dénommée par un nom unique, qui permet de rejoindre le réseau (plusieurs moyens et/ou lignes), par exemple Gare de l'Est ou Montparnasse.



- Un **arrêt** est un point précis de desserte de transports, localisé géographiquement. Ainsi, pour une même zone d'arrêt, on distingue non seulement les divers arrêts de bus, et l'arrêt de métro/RER.

Pour quelles utilisations ?

Ces données ont, bien sûr, pour vocation première d'être utilisées en interne par l'exploitant, pour son organisation et sa gestion au quotidien. Elles sont au cœur de ce que l'on appelle les **systèmes d'aide à l'exploitation et à l'information voyageur** (SAEIV), systèmes informatiques en réseau distribué permettant la supervision, le contrôle et la gestion de la flotte du réseau de transport en interne (partie SAE), mais dont une partie des résultats est utilisée pour l'information voyageur. Traditionnellement, ces informations voyageur étaient diffusées par l'intermédiaire d'équipements urbains, tels que des bornes d'information aux arrêts de bus donnant l'heure des prochains passage, ou des écrans dans les véhicules indiquant la position actuelle et l'heure estimée d'arrivée au terminus.



Une borne d'information de la RATP, source : <http://www.transbus.org/>

Aujourd'hui, il est possible de relier ces systèmes d'information aux sites internet des exploitants, voire à leurs applications mobiles, pour étendre l'information aux systèmes d'information personnels des usagers, comme c'était déjà le cas, avant l'avènement des smartphones, avec les alertes SMS. Dans le cadre du phénomène open data, les réutilisateurs réclament maintenant la libération de ces données d'information voyageur, pour s'affranchir de ces barrières, proposer leurs propres systèmes de représentation de ces données et ainsi informer les voyageurs de façon plus complète, plus efficace ou tout simplement différente.

On parle d'ailleurs de plus en plus, pour décrire ces interactions étroites entre systèmes de gestion internes et systèmes d'information voyageur rendues possibles par les nouvelles technologies, de **Systèmes de Transport Intelligents** (STI, ou ITS en anglais), qui manipulent des données extrêmement diverses, pour ensuite les diffuser dans des contextes et pour des utilisations diverses.



Types de données et applications dans les systèmes d'information intelligents de transport urbain

Source : livre blanc Orange « transport collectif : l'ère du voyageur numérique »

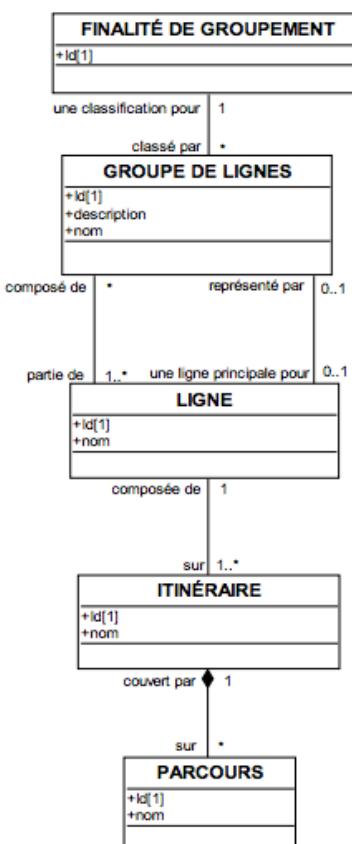
Maintenant que nous connaissons les principales données de transport public, il est temps de se pencher sur celles qui sont susceptibles d'être ouvertes et mises à disposition du public, et surtout, sous quelle forme cela va être fait.

Formats des données de transport

Même si les données de transport public sont, on l'a vu, de natures très diverses, les opérateurs de transport ont naturellement besoin de normes pour représenter, stocker, manipuler et échanger. les informations de leurs systèmes dans leur ensemble. Nous ferons ici un aperçu des normes les plus largement utilisées.

Transmodel

Transmodel est le modèle européen de données de transport public de référence. C'est une norme déposée à l'AFNOR (sous le code EN12896) qui décrit en détails un modèle abstrait et des structures de données pour tous les concepts métier du transport public. L'introduction aux concepts énoncés plus haut, est un résumé très grossier de la partie « guide réseau » de cette spécification, qui décrit aussi les données de personnel roulant, de billettique, d'information voyageur, de suivi de l'exploitation... Tous les concepts y sont détaillés sous forme de modélisation conceptuelle, selon le formalisme UML.



Source : extrait de la norme Transmodel décrivant le concept de "Ligne"

Dans la pratique, cette norme est tellement vaste, et surtout tellement abstraite, qu'elle tient surtout de guide de conception pour les systèmes d'information des opérateurs de transport, qui l'adaptent et le diminuent selon leurs besoins, et l'implémentent ensuite avec les outils techniques de leur choix. Il n'y a donc pas d'échange possible entre les systèmes de données compatibles Transmodel.

TRIDENT et Neptune

Pour pallier à cela et permettre des échanges de données entre les exploitants, une norme d'échange reposant sur Transmodel a été choisie en 2002 par un groupe de travail européen : la spécification **TRIDENT** propose une norme plus étendue permettant de représenter, en particulier, une offre de transport complète (horaires, réseau, tarifs), théorique ou en situation de perturbations, mais aussi — et surtout ! — un format de représentation et d'échange de ces données.

En 2010, la spécification TRIDENT a été augmentée par la France par de nouvelles descriptions de concepts (concernant l'accessibilité, les groupes de lignes, les horaires en fréquence, ou encore l'état des équipement) et est devenue à cette occasion la norme **NEPTUNE** (pour *Norme d'Echange Profil Transport collectif utilisant la Normalisation Européenne*). Les fichiers au format NEPTUNE jouissent d'une rétro-compatibilité avec le format TRIDENT.

Les normes d'échange NEPTUNE et TRIDENT reposent toutes deux sur le format XML.

```
<StopArea>
  <areaType>CommercialStopPoint</areaType>
</StopAreaExtension>
</StopArea>
  <objectId>SYNTHESE:StopArea:1970324837184772</objectId>
  <name>Saint Cyprien - République</name>
  <contains>SYNTHESE:StopArea:3377699720880908</contains>
  <contains>SYNTHESE:StopArea:3377704015495561</contains>
  <StopAreaExtension>
    <areaType>CommercialStopPoint</areaType>
  </StopAreaExtension>
</StopArea>
  <objectId>SYNTHESE:StopArea:1970324837184915</objectId>
  <name>Arsenal</name>
  <contains>SYNTHESE:StopArea:3377699720881231</contains>
  ...
```

Source : Extrait du fichier décrivant les horaires de la ligne 1 du métro du réseau de transport Tisseo de Toulouse, mis à disposition sur <http://data.grandtoulouse.fr> au format Trident.

Il faut toutefois noter que ces formats d'échanges n'englobent pas tous les concepts décrits dans la norme descriptive, comme les perturbations temps-réel, et couvrent uniquement la description statique d'une offre de transport.

Afin de promouvoir et d'encourager l'utilisation de ces normes, un logiciel libre a été conçu et distribué : le logiciel **CHOUETTE** (pour *Création d'Horaires avec un Outil d'Échange de données TC selon le format Trident Européen*) a pour but de consulter, mais aussi de construire et d'exporter des fichiers dans ces formats. Une version de l'outil existe aussi en ligne, pour en permettre une utilisation facile ; en effet, même si les codes sources sont disponibles, il n'est pas donné à tout le monde de les compiler pour exécuter le logiciel sur son système, et on assure ainsi un accès égal à tous.

GTFS

Mais le format de données de transport le plus connu et le plus répandu reste le format **GTFS**. Conçu en 2002 par Google en partenariat avec un ingénieur du TriMet (un opérateur de transport public de la ville de Portland), le GTFS est rapide-

ment devenu, un standard de fait aux États-Unis, et progressivement, aujourd'hui, dans le monde entier pour le partage des données de transport en commun ; en conséquence, il a été décidé de changer sa signification de « *Google Transit Feed Specification* » en « *General Transit Feed Specification* ».

Son succès est dû à plusieurs facteurs ; le premier est sa simplicité, qui le rend facile à utiliser, mais aussi à comprendre par des réutilisateurs potentiels. En effet, il ne repose pas sur le format XML, complexe et verbeux, mais sur de simples fichiers CSV. Chacun de ces fichiers décrit un concept simple :

- **agency.txt** détaille les agences émettrices des données, et leurs coordonnées (souvent, il n'y en a qu'une seule)
- **calendar.txt** liste des périodes de service, avec leur date de début et leur date de fin (c'est un moyen de regrouper les services de périodes des vacances scolaires, de fêtes de fin d'année, etc...)
- **stops.txt** est un fichier référençant tous les arrêts, leurs coordonnées géographiques et leur zone de tarification
- **routes.txt** liste toutes les *routes* possibles par le réseau : souvent, le sens aller, le sens retour, et si le plan contient des embranchements, les diverses combinaisons possibles d'un terminus à l'autre.
- **trips.txt** liste toutes les *courses* (route avec application des horaires), en les associant à un service et à une *route*.

- **stop_times.txt** liste, pour chaque arrêt référencé dans stops.txt et chaque course référencée dans trips.txt, l'horaire de passage (départ et arrivée).

Ces six fichiers sont ceux qui sont rendus obligatoires par la spécification GTFS. Comme on le voit, ils se concentrent surtout sur les horaires (aspect temporel) et les arrêts (aspect spatial). Il est cependant possible d'ajouter au jeu de données d'autres fichiers couvrant d'autres types d'information, ou complétant les premières :

- **transfers.txt** spécifie, pour tous les arrêts concernés, le temps nécessaire, en minutes, pour aller de l'un à l'autre, en précisant si le trajet se fait à pied (correspondance) ou en transport. Ainsi, si ce fichier est inclus, les calculs de temps de trajet sont facilités, et n'ont pas à être effectués par l'application.
- **fare_attributes.txt** liste les différents types de tarifs, leur prix, leur méthode de paiement et le nombre de correspondances (et durée) autorisées. Il est associé au fichier **fare_rules.txt**, qui lui précise les conditions de ces tarifs (*route* associée, *zone*...)
- **frequencies.txt** permet de spécifier les horaires non plus de façon fixe comme dans *stop_times.txt*, mais par fréquence (« le bus 68 passe toutes les 12 minutes entre 8h et 19h »)
- **shapes.txt** permet de décrire des lignes reliant des points localisés dans l'espace, qui pourront servir à décrire géographiquement des lignes de transport sur une carte

De plus, un *méta-fichier* complémentaire **feed_info.txt** contient les informations de contexte

du jeu de fichiers, comme le nom de l'émetteur des données (souvent le même que celui décrit dans *agency.txt*, mais des exceptions sont possibles dans le cas d'entités régionales), la langue, les dates de validité ou la version.

Cette façon de représenter les données de transport est un compromis intéressant : en effet, le choix a été fait de pouvoir partager des données assez complètes pour être utiles au public, mais aussi assez simples pour être facilement comprises — et donc éventuellement réutilisées — par des non-spécialistes, à l'inverse des normes Neptune ou Trident, bien plus complexes à appréhender¹.

Mais ce n'est pas la seule raison de son succès, loin de là. En effet, les données dont on a parlé jusqu'ici, et représentées par les formats standard, sont les données statiques. Mais on a vu que les données les plus critiques, dans le domaine du transport, sont les données temps-réel. Google l'a compris, et a lancé en 2011 une extension au format GTFS, appelée **GTFS-realtime**, qui permet de représenter et diffuser des données dynamiques.

Il est possible de spécifier trois types de flux de données avec la norme GTFS-realtime.

Le premier concerne les **positions des véhicules en temps réel**, permettant de connaître l'horaire estimé de prochain passage à un arrêt, ou plus globalement l'état de congestion du trafic.

Les deux autres sont des **alertes de perturbation** :

- Alertes de perturbations ponctuelles sur une course individuelle, qui permettent de définir un retard et une prédition sur les nouvelles heures d'arrivées aux arrêts, voire une annulation ou une déviation, et de contextualiser ces informations avec un degré d'incertitude donné.
- Alertes de perturbations sur un service entier (stations fermées pour travaux, portions de lignes fermées régulièrement, retards importants...), généralement utilisées pour des perturbations plus importantes ; on définit alors une date de début et une date de fin de validité de cette alerte, une raison (travaux, voyageur malade...), et surtout l'étendue des répercussions : quelles routes, courses, arrêts sont concernés ? En renseignant cette information de façon détaillée, on évite de lancer des alertes anxiogènes et inutiles pour des personnes non concernées.

Les flux doivent être mis à disposition, sous forme de fichiers (un par type de flux), sur un serveur accessible via le protocole HTTP, et mis à jour le plus régulièrement possible par l'émetteur des données, pour bien sûr être le plus pertinents possible à chaque instant.

Le format GTFS-realtime n'est pas basé sur XML, mais sur un dérivé de celui-ci, appelé **Protocol Buffers**. Ce langage de sérialisation permet de décrire des structures de données indépendamment d'un langage ou d'une plateforme de stockage, de façon simple et claire. Concrètement, ce format va être un "pont" entre une représentation de données abstraites et son implémentation dans

¹ Par exemple, GTFS ne permet pas de représenter la notion d'arrêts multimodaux, de zones d'arrêts, ou de niveaux d'accessibilité...

un langage donné : on va utiliser un fichier portant l'extension `.proto` (dans la convention GTFS-realtime, il est nommé `gtfs-realtime.proto`), et le compiler — en utilisant un compilateur fourni par Google en open source — pour obtenir ces structures de données (chaque objet étant appelé *message*) dans le langage de programmation choisi.

```
// Uncertainty applies equally to both directions of travel.
// The uncertainty roughly specifies the range of error for the vehicle's position.
// note, we don't yet define its precision.
// for the uncertainty to be 0, for example, would mean that the vehicle is exact.
message StopTimeEvent {
    // Delay (in seconds) can be positive or negative (meaning that the vehicle is early or late).
    optional int32 delay = 1;

    // Event as absolute time.
    // In Unix time (i.e., number of seconds since January 1, 1970 UTC).
    optional int64 time = 2;

    // If uncertainty is omitted, it is assumed to be 0.
    // If the prediction is unknown or invalid, uncertainty should be empty. In such case,
    // To specify a completely certain prediction, set uncertainty to 0.
    optional int32 uncertainty = 3;

    // The extensions namespace allows for extensions to the GTFS-realtime specification in the future.
    // and modifications to the spec.
    extensions 1000 to 1999;
}
```

Source : Extrait du fichier `gtfs-realtime.proto`, <https://developers.google.com/transit/gtfs-realtime/gtfs-realtime-protocol-buffer>

Le fichier `gtfs-realtime.proto` est en libre téléchargement, et va permettre de créer — par exemple, si le système interne de gestion d'information de l'émetteur de données est implémenté en C++ — les classes C++ objet décrivant les concepts utilisés par la norme GTFS-realtime, qui vont pouvoir servir d'interface pour le système propriétaire de l'émetteur de données. Il existe aujourd'hui des compilateurs de `.proto` vers les langages C++, Java et Python.

Cette étape faite, on va pouvoir créer, mettre à jour et émettre assez simplement des fichiers d'échange de données binaires (c'est-à-dire contenant des données brutes et non lisibles de façon directe par des humains), qui auront souvent pour extension `.pb`. Ce sont les fichiers que l'on va retrouver sur les portails open data des organisations de transport émettant des données dynamiques. On pourrait se demander pourquoi ne pas avoir utilisé la norme XML ; en réalité, ils sont jusqu'à 10 fois plus légers, décrivent ces données très complexes de façon beaucoup moins ambiguë, et surtout sont réutilisables rapidement, après recompilation par les logiciels, sans avoir besoin de faire parser au préalable du texte par un programme, ce qui peut s'avérer long en traitement. Au final, on répond aux besoins de flexibilité, légèreté, rapidité et automatisation intrinsèques à la nature-même des données temps-réel.

En parallèle, on peut aussi créer un fichier texte descriptif, qui va représenter les mêmes données, mais lisibles par un humain. Il ne sera pas voué à l'utilisation en production, mais plutôt à des fins d'exemple et de contrôle.

```
translation {
    text: "http://www.sometransitagency/alerts"
    language: "en"
}
header_text {
    translation {
        text: "Stop at Elm street is closed, temporary stop at Oak street"
        language: "en"
    }
}
description_text {
    translation {
        text: "Due to construction at Elm street the stop is closed. The temporary stop can be found 300 meters north at Oak street"
        language: "en"
    }
}
```

Source : exemple de représentation ASCII d'un flux d'alertes, <https://developers.google.com/transit/gtfs-realtime/examples/alerts>

Il faut aussi signaler que ces structures de données sont, de l'aveu-même de Google, pensées avant tout du point de vue de l'utilisateur final : plutôt que de réfléchir du point de vue de l'émetteur des données, par essence spécialiste, on a voulu avant tout privilégier l'information finale qui va être communiquée à l'usager des transports.

Les acteurs en jeu

En réalité, statuer sur les données de transports en France est particulièrement délicat, pour la simple raison que le milieu est extrêmement complexe d'un point de vue institutionnel et juridique.

La loi LOTI et les AOT

En France, l'organisation générale des services publics de transports est encadrée par une loi : la loi n° 82-1153 du 30 décembre 1982 d'orientation des transports intérieurs, plus communément appelée loi LOTI.

Cette loi instaure la mise en place d'**Autorités Organisatrices de Transport** (AOT), collectivités publiques dont la mission est d'organiser les transports publics à chaque niveau territorial. Sont des AOT :

- Les **communes** (parfois organisées en regroupements), qui organisent les transports urbains (on parle alors d'AOTU, Autorité Organisatrice de Transports Urbains). Elles peuvent assurer cette compétence en régie, ou

déléguer l'exploitation à un **opérateur de transport**.

- Les **départements** gèrent les transports interurbains
- Les **régions** sont en charge des transports ferroviaires sur leur territoire (TER).

Ces autorités ont donc des rôles définis et bien séparés les uns des autres. En conséquence, les données de transports, dans une délimitation géographique donnée, sont réparties entre plusieurs organisations superposées qui peuvent avoir des politiques — ou simplement des rythmes de travail — totalement différentes, ce qui rend très difficile une libération totale et cohérente des données.

La majorité des AOT françaises sont regroupées au sein d'une association appelée le **GART** (Groupement des Autorités Responsables du Transport), qui est leur porte-parole vis-à-vis du public et des autorités autres (gouvernement, Union Européenne...)

Mais la loi LOTI ne définit pas uniquement les devoirs des institutions dans la mise en place de services de transports efficaces ; elle confère aussi plus globalement aux usagers des droits fondamentaux. Ainsi, l'article 1 de la LOTI spécifie que « *tout usager a le droit de se déplacer et la liberté d'en choisir les moyens* ». Plus intéressant vu le sujet que nous couvrons, l'article 2 dit que « *le droit au transport comprend le droit pour les usagers d'être informés sur les moyens qui leur sont offerts et sur les modalités de leur utilisation*. »

Le cas particulier de l'Île-de-France

L'organisation des transports publics en Île-de-France ne relève pas de la LOTI, mais du décret n° 49-1473 du 14 novembre 1949. L'Autorité Organisatrice de Transport de la région y est le **STIF**, le



Syndicat des Transports d'Île-de-France, moins connu par le grand public que les deux opérateurs de transport ferré de la région parisienne, la **RATP** et la **SNCF**, mais qui a une grande part du pouvoir de décision en ce qui les concerne.

Quelques entreprises privées qui jouent un rôle important

Commençons par un fondamental de l'information transport nationale :



CanalTP est le leader français de l'information de transport voyageur.

Filiale de Keolis¹, la société se positionne comme fournisseur d'outils et de services sur mesure dans le domaine du déplacement et de l'information voyageur multimodale.

CanalTP propose une suite applicative clé en main, appelé **Navitia**, qui est le cœur de leur service ; complet et modulaire, Navitia propose entre autres fonctionnalités un calcul d'itinéraires multimodal, la recherche d'horaires de ligne ou à un arrêt, la visualisation sur un carte, l'information en temps-réel (alertes retards, perturbations...), tarifications, comptes et alertes personnalisées selon le profil du voya-

geur...

La société accompagne les opérateurs de transport en intégrant leur solution au SAEIV existant pour en extraire les données intéressantes (théoriques, temps-réel), et accompagne les exploitants dans l'utilisation des outils.

Actuellement, 12 régions sur les 17 françaises équipées de systèmes multimodaux utilisent la solution Navitia.

Au-delà de sa présence bien implantée en France, CanalTP joue un rôle important dans les débats sur l'open data dans les transports en commun. En effet, sa position sur le sujet est clairement orientée vers l'ouverture des données. Pour preuve, les nombreuses déclarations de ses dirigeants sur le sujet, leur participation à un certain nombre d'évènements (hackathons, conférences), et surtout la création de leur API publique **navitia.io**, présentée dès son lancement, en 2013, comme un moyen de faciliter et d'accélérer l'utilisation de données de transport ouvertes². CanalTP a mis en place cette API, accompagnée d'une explication claire et pédagogique sur les données de transport, pour chaque ville qui a mis ses données de transport à disposition, pour encourager les développeurs et hackers à utiliser ces données de façon unifiée et documentée, sans avoir à passer par les AOT concernées. Encore plus intéressant, leur système de calcul d'itinéraire y est intégré (faisant partie de Navitia), et permet ainsi à toute personne s'intéressant à ces données de les réutiliser simplement.

¹ Grand opérateur privé français de transport public.

² Voir [« navitia.io : une API de calcul d'itinéraires »](#), sur le blog de CanalTP.

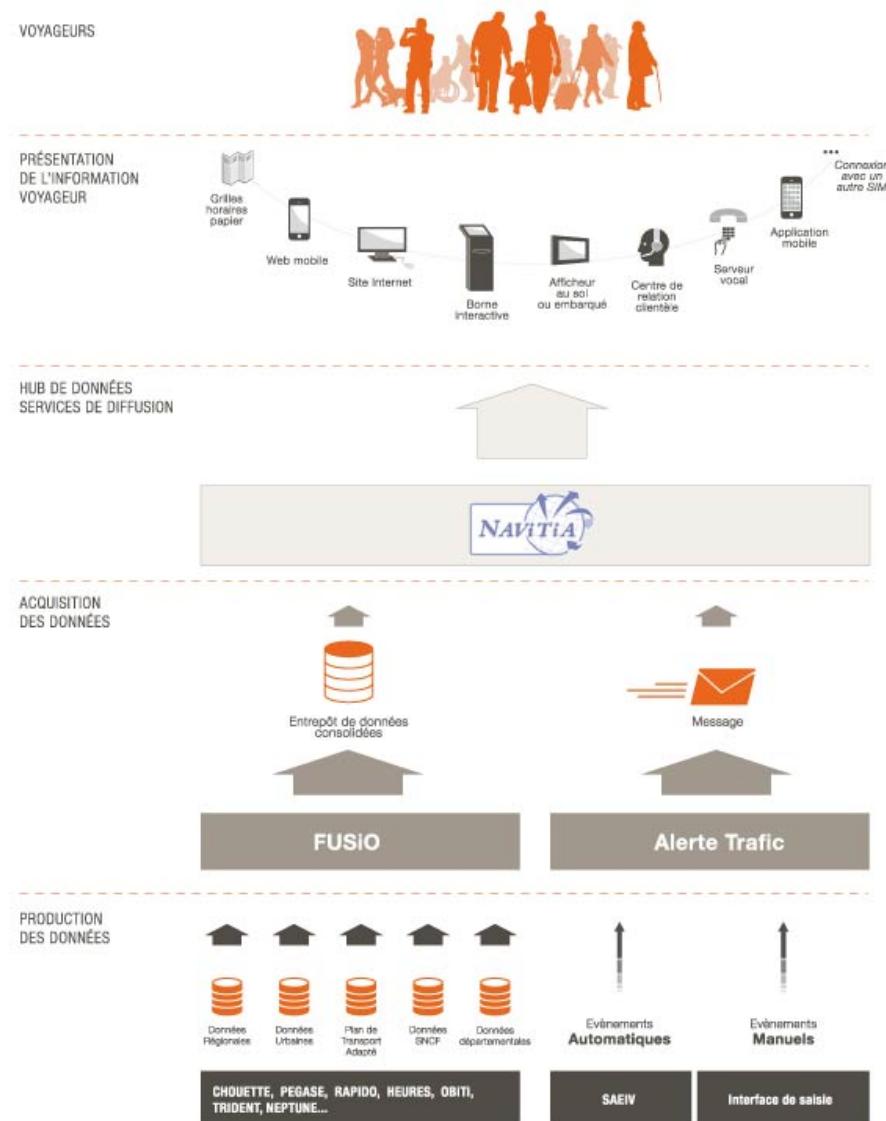


Schéma explicatif de l'offre de services de CanalTP

Source : canaltp.fr

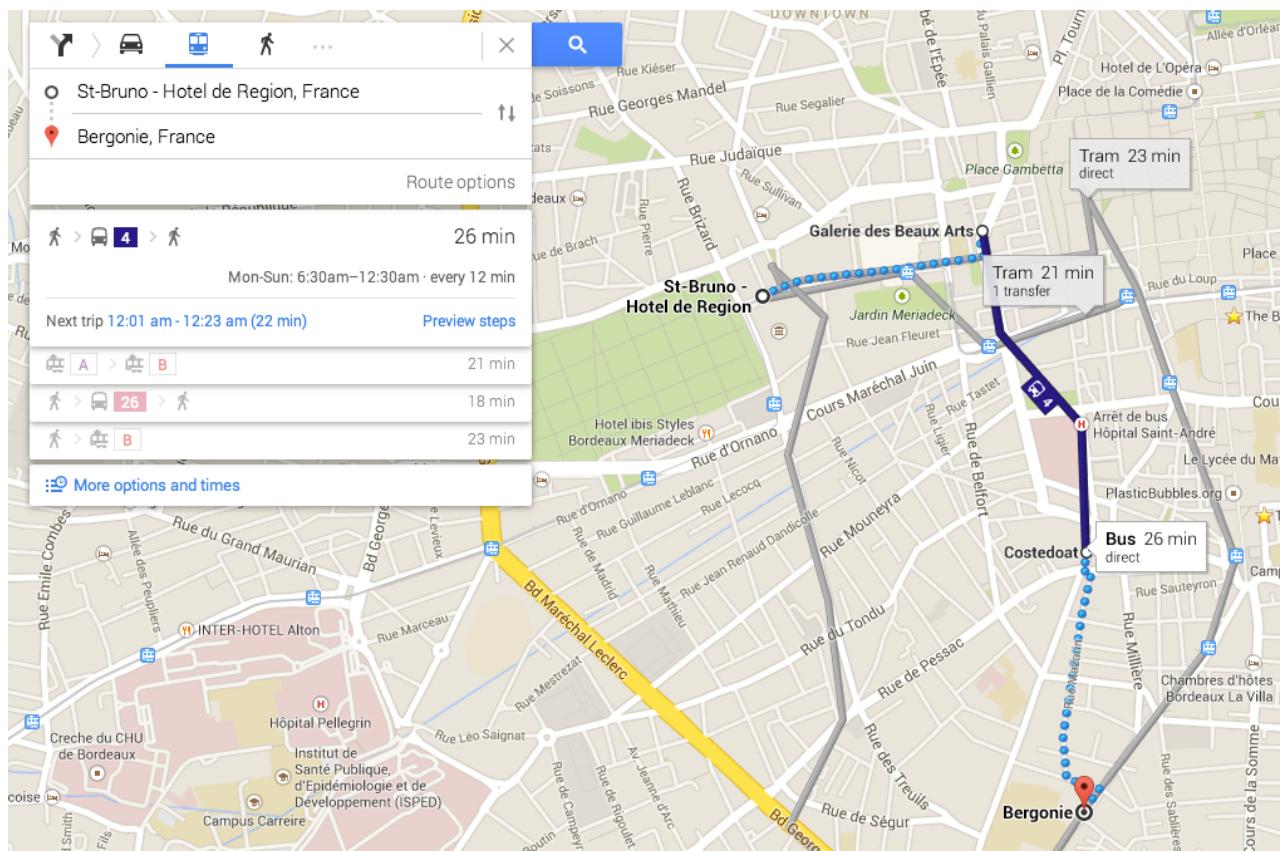
Le cas Google

Aucunement besoin de présenter Google, géant du net à l'influence incomparable dans tous les domaines du numérique ; il est ceci dit moins connu du grand public que Google s'intéresse de très près au phénomène open data, et est partenaire d'Etalab, via l'organisation du programme « **Data-connexions** », qui récompense les meilleures innovations utilisant des données publiques.

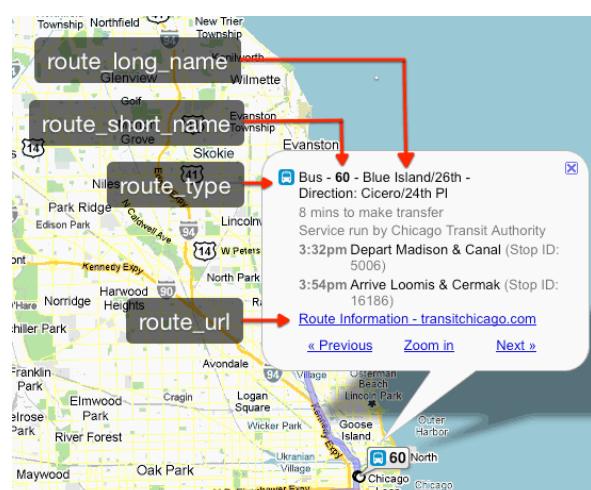
Google Transit

On a déjà longuement évoqué le cas Google lors de la présentation du standard GTFS, aujourd'hui utilisé dans la très grande majorité des cas par les administrations qui partagent leurs données de transport.

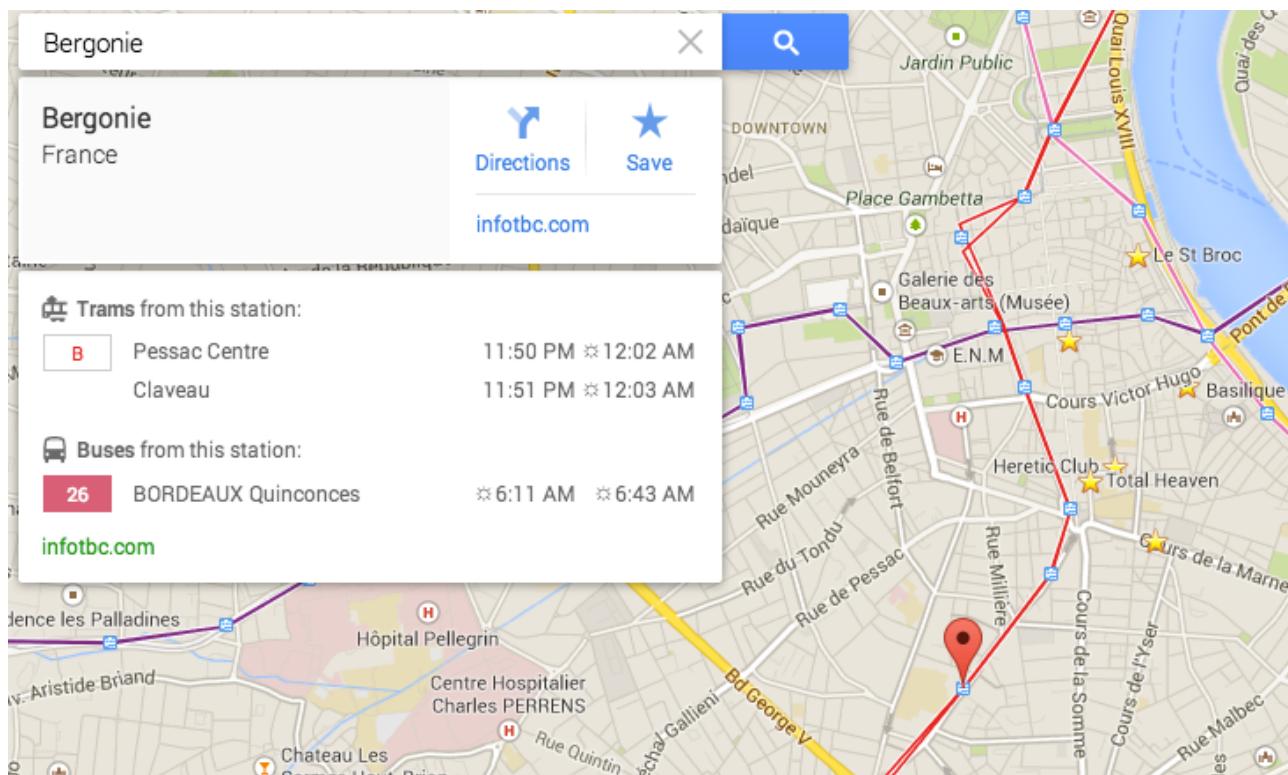
En réalité, ce format d'échange, si bien pensé soit-il, n'aurait pas rencontré un tel succès sans le moteur stratégique derrière son existence, à savoir le programme **Google Transit**. Crée en 2005, en



partenariat lui aussi avec le TriMet de Portland, c'est un portage de la norme GTFS pour Google Maps : ainsi, Google devient lui-même réutilisateur des données de transport ouvertes au format GTFS, et son célèbre Google Maps, déjà souvent utilisé pour des calculs de trajectoire en voiture ou à pied, est capable d'utiliser ces données de plusieurs façons, et ce dans le site lui-même, mais aussi dans l'application mobile, et même dans Google Earth :

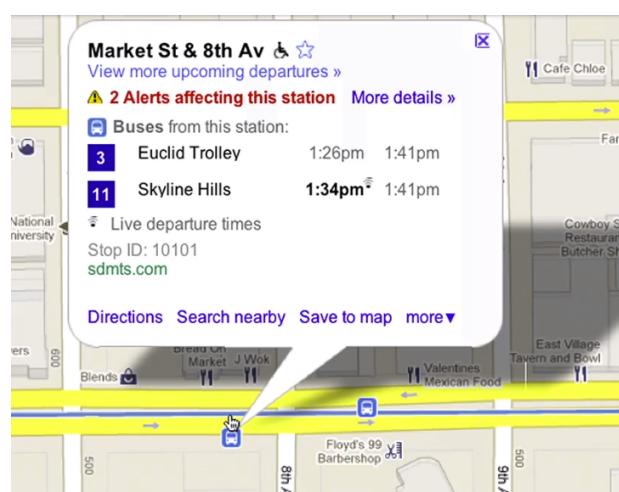


- **Affichage statique** : lignes et arrêts de transports en commun affichés comme un calque sur la carte, informations par arrêt (lignes partant de cet arrêt, ou même, si des données GTFS-realtime sont disponibles, horaire du prochain passage de telle ligne à cet arrêt)
- Mais la fonctionnalité la plus intéressante reste l'utilisation du **calcul d'itinéraire** de Google



Maps : pour les villes couvertes, en plus du choix des modes piétons et voiture, il est possible de choisir les transports en commun : Maps va alors utiliser les ressources GTFS pour proposer plusieurs trajets à l'usager. Si les données de tarification sont disponibles, Maps va calculer le prix de chaque trajet.

L'insertion de GTFS-realtime dans Google Transit permet des updates en temps-réel : dans ce cas, les horaires "réels" sont distingués par un pictogramme, et les alertes de perturbations sont affichés sur les informations des stations, et prises en compte dans les calculs d'itinéraires.



Pour apparaître dans Google Transit, les autorités organisatrices de transport sont invitées par Google¹, à travers le programme **Google Transit Partner**, à créer ces données de façon conforme à leurs spécifications, puis à les contacter pour leur soumettre. Cela est totalement gratuit pour les autorités de transport qui participent, et Google met en avant pour les convaincre des arguments percutants : amélioration du service rendu aux usagers et de la visibilité du réseau vis-à-vis du

¹ <http://maps.google.com/intl/fr/help/maps/mapcontent/transit/participate.html>

grand public, interconnexions avec les villes voisines, fluidification des trajets et donc meilleur rendement, réduction des embouteillages (pour les régions soucieuses d'une amélioration de la qualité de vie de leurs régions...).

Aujourd'hui, de très nombreuses villes sont couvertes par Google Transit dans le monde, sur les 5 continents, à des niveaux plus ou moins détaillés (de l'affichage des lignes et des stations seulement aux horaires statiques, voire les informations temps réel). En France, une cinquantaine d'autorités de transports ont franchi le pas.



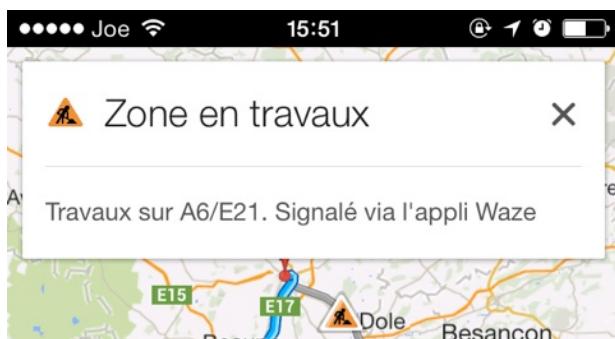
Bien sûr, même si côté Google tout cela est présenté comme un exemple parfait de partenariat public/privé, tout n'est pas si rose, et on imagine les bras de fer possibles entre les différents acteurs en jeu. La représentation des transports en commun parisiens et son évolution dans Google Maps (via Transit) est symptomatique de ces crispations. En effet, fin 2011, la « couche » Transit est apparue sur Paris, alors qu'aucune autorité de transport n'avait encore libéré ses données ; on peut alors tout simplement imaginer (et c'est d'ailleurs la seule explication) que Google a tout simplement pris les données sans autorisation, profitant de son statut d'intouchable géant du web... Le service était d'ailleurs déplorable dans les premiers mois, souffrant d'incohérence dans les calculs de trajectoire (à cause de moyens de transports non référencés, comme les bus), voire d'erreurs dans les calculs de tarification. Google étant d'habitude très soucieux de la qualité du service rendu aux utilisateurs, on peut imaginer

que ce choix a été plus politique qu'à visée réellement utile.

Signalons aussi que comme à son habitude, Google propose une **API** pour son service Google Transit, qui permet de calculer des directions et itinéraires en transport en commun en utilisant ces données.

Vers des applications crowdsourcées en temps réel ?

Cela ne concerne pas encore le transport public, mais il est certain que Google souhaite s'en rapprocher : en achetant l'application Waze, réseau social pour automobilistes permettant de signaler des incidents sur la route aux autres usagers connectés, Google a fait un grand pas en avant vers le crowdsourcing ; en effet, depuis novembre 2013, les reports de l'application apparaissent en temps-réel dans Google Maps, et sont même prises en



compte dans le calcul d'itinéraires en voiture. Nul doute qu'un même fonctionnement se profile pour Google Transit, ce qui permettrait à terme une réactivité et une exactitude inégalables dans les alertes et les calculs d'itinéraire de transport en commun, vu la masse critique d'utilisateurs.

Que dit la loi ?

On en a parlé lorsque l'on a présenté les généralités de la loi CADA plus haut : **les documents produits par des Établissements Publics Industriels et Commerciaux (EPIC) ne sont pas considérés comme des documents publics**, même s'ils opèrent dans une optique de service public, car ce sont des entreprises privées soumises à la loi du marché¹. C'est le cas de quelques grands opérateurs de transport, au premier rang desquels on compte la RATP et la SNCF, qui ne sont donc pas tenues de mettre leurs données à disposition. Cependant, l'opinion est consciente du côté paradoxal de cette limitation ; pour preuve, le Conseil National du numérique a publié un avis² en décembre 2012 qui préconisait, entre autres, d'étendre la réutilisation des données publiques aux EPIC. Conseil resté sans suite.

Mais tous les AOT ne sont pas des EPIC, et on l'a vu, les données de transport sont très demandées par les réutilisateurs potentiels, car elles concernent presque tout le monde, sont utiles au quotidien et s'inscrivent parfaitement dans le contexte d'une utilisation mobile (applications). En conséquence, La Commission d'Accès aux Documents Administratifs a été saisie de nombreuses fois par des citoyens lui demandant l'accès à ces données, considérant (ce qui relève d'une certaine logique) que ce sont des données publiques, car produites dans l'exercice d'un service public. Malheureusement, le texte de loi ne fait aucunement mention des données de transport, et ce sont les réponses

de la CADA qui font office de référence aux observateurs pour décider quelles données relèvent de ce droit.

Nous allons ici tenter de dégager les conclusions tirées jusqu'ici, et considérées comme acquises par les observateurs.

En tout premier lieu, il est aujourd'hui admis de façon générale que **les horaires théoriques de passages sont considérés comme publics et font l'objet d'un droit de réutilisation, ainsi que les coordonnées géographiques des arrêts, et celles des lignes de transport**.

Néanmoins, la frontière reste mince entre données considérées publiques par toutes les parties, et données sensibles que les opérateurs souhaitent protéger.

En témoigne l'épisode assez médiatisé qui a lancé un débat en mai 2011 : l'application *CheckMyMetro*, déjà disponible sur l'Apple Store, et dont le concept était jusque là de proposer un réseau social pour les usagers du métro, ajoute à ses fonctionnalités le plan du métro parisien, ainsi que les horaires temps-réel scrapés (récupérés) depuis le site WAP de la RATP.

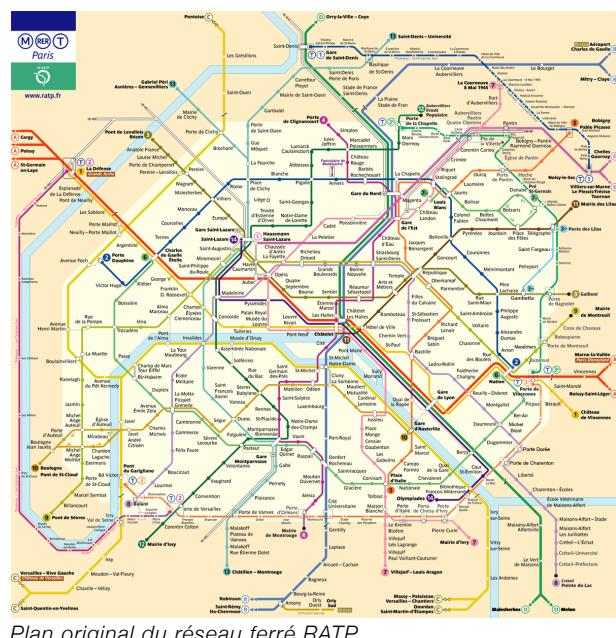


¹ On peut considérer qu'il s'agit d'une des restrictions les plus importantes concernant le développement de l'open data, d'autant plus que de nombreux établissements ayant ce statut possèdent des données de valeurs (par exemple l'ONF, l'ADEME ou l'IGN).

² Voir [« Avis n° 12 du Conseil national du numérique relatif à l'ouverture des données publiques \("Open data"\) »](#)

La RATP est alors presque immédiatement montée au créneau, en demandant à Apple le retrait de l'application de son Store, sous prétexte de violation de sa propriété intellectuelle. Les débats ont été houleux et médiatisés — l'intérêt du public pouvant aussi être expliqué par le fait qu'à ce moment-là, les horaires temps-réel ne sont disponibles que sur l'application *payante* de la RATP — et CheckMyMetro a fini par retirer les horaires et la carte de son application, et proposer un jeu concours de création originale et libre d'une nouvelle version de la carte du métro parisien¹.

En effet, le consensus dégagé par les nombreux débats sur le sujet a été de considérer la carte de la RATP comme une création originale protégée par le Code de la Propriété Intellectuelle : en effet, il s'agit plus d'un *document* que d'une *donnée*.



Plan original du réseau ferré RATP

Source : <http://www.ratp.fr/>

Mais le débat était lancé ; en effet, pour se défendre des accusations de la RATP (surtout motivée par le fait que l'application permettait de signaler des contrôleurs), le créateur de l'application Ben-

jamin Suchar décide d'accuser en retour la RATP de ne pas libérer ses données de trafic, qui pour lui relèvent de données de droit public, et sont à ce titre concernées par la CADA. La question est toujours en débat aujourd'hui, pour la raison que les données concernées par l'affaire CheckMyMetro sont précisément des données temps-réel, mais nous reviendrons sur ce point précis plus en détails.

L'affaire CheckMyMetro a en tout cas entériné le fait que les plans de réseaux ne sont pas concernés par la CADA car ils relèvent d'une création originale ; par contre les données géographiques des arrêts et des lignes, qui permettent de recréer ces plans de toutes pièces, relèvent, comme spécifié précédemment, de la loi CADA.

En revanche, la CADA n'a pas encore été amenée à se prononcer officiellement sur les données temps-réel, alors que c'est un enjeu fondamental, comme nous le détaillerons plus loin.

Les initiatives d'ouverture de données en France

Initiatives des AOT locales

Jusqu'ici, les initiatives d'ouverture de données ont surtout été le fait des collectivités locales, poussées par le volontarisme de l'État en matière d'open data, et par les demandes de militants lo-

¹ <http://www.checkmymap.fr/le-concours>

caux encouragés par les avis favorables de la Cada sur le sujet.

Nous avons déjà passé en revue certaines de ces initiatives dans la première partie de cette présentation, en mettant en avant la diversité des données mises à disposition par les collectivités ; en fait, il faut savoir que **les données de transport concernent un jeu de données sur cinq parmi les portails opendata des agglomérations.**

Le précurseur dans ce domaine a sans conteste été la métropole de Rennes, qui a lancé en 2010, dans la continuité de son portail open data (le premier en France), un concours appelé « *Rennes Métropole en accès libre* », qui avait pour vocation de récompenser les meilleures applications réutilisatrices. En partenariat avec l'opérateur de transport Kéolis STAR, elle a beaucoup misé sur les données de transport, ce qui s'est avéré payant puisque parmi les 8 applications primées, 5 utilisaient les données transport !

D'autres agglomérations, collectivités et régions leur ont rapidement emboîté le pas, et aujourd'hui, une grande partie du territoire français est "couverte", en considérant toutes les régions qui ont entrepris une démarche, même de façon séparée. On peut compter parmi ceux-là (il y en a d'autres, mais on peut considérer que ceux-ci sont les plus complets et les plus représentatifs) :

- Le Grand Toulouse et son réseau TISSEO, sur le portail data.grandtoulouse.fr
- L'agglomération de Nantes et son réseau SEMITAN, sur le portail data.nantes.fr

- La Communauté Urbaine de Bordeaux (data.lacub.fr), qui est allée loin avec une API complète et une offre de transport très variée.

Ceci étant dit, ces démarches sont très diverses dans leurs caractéristiques :

- D'abord de par le **type de données** de transport partagées : si presque toutes partagent les **coordonnées géographiques des arrêts** et les **horaires théoriques**, elles ne proposent pas toutes des **données temps-réel**. Lorsque c'est le cas, par contre, les API sont de qualité et bien documentées, et même parfois doublées par des fichiers en accès au format RSS. Parfois, les **dessins des lignes ou des zones** est mis à disposition dans un format géographique tel que le KML ou le SHP.
- Mais aussi par la fréquence de mise à jour des données, et la **contextualisation** des jeux de données : métadonnées, explications et documentations sur les formats (utiles en particulier dans le cas de Neptune ou Trident, qui sont des formats très spécialisés difficiles à appréhender)
- Ainsi, le consensus semble ne pas avoir été trouvé sur les **formats** à utiliser, et les formats XML de normes européennes Trident et Neptune sont aussi représentés que le format GTFS. Parfois, ce sont les deux qui sont mis à disposition en parallèle.

Les **licences** sont du mêmes type que celles déjà énoncées plus haut : Etalab, ODbL ou APIE. Les fichiers sont toujours mis à disposition gratuitement, et demandent de respecter cette licence en cas de réutilisation ; parfois il est évoqué la possi-

bilité de la mise en place d'une rémunération dans le cas d'une utilisation payante, mais il n'a pas encore, à notre connaissance, existé de cas dans lequel ce genre de redevance aurait été appliquée, la réutilisation se faisant uniquement sur un mode déclaratif...

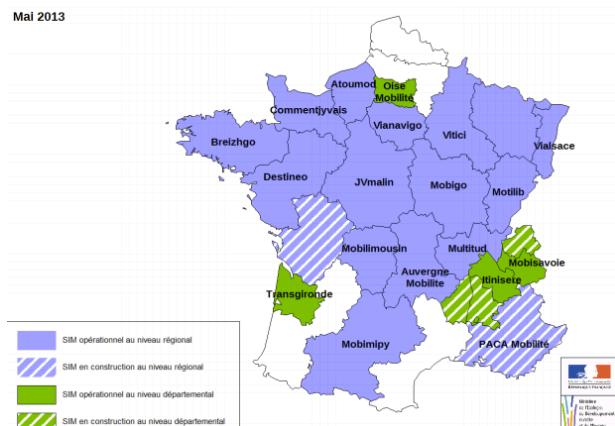


Mais les données ne sont pas le seul enjeu pour le public ; les territoires ont donc aussi mis en place des « **systèmes d'information multimodaux** » (ou SIM), pour aider les visiteurs et résidents à organiser leurs déplacements dans leur région. Parmi ceux-là, le premier,

qui a ouvert la voie à tous les autres, s'appelle [Destineo](#), existe depuis 2006 et propose un moteur de calcul d'itinéraires motorisé par la technologie Navitia de CanalTP pour la région des Pays de la Loire. De nombreux sites à vocation similaire ont suivi, par exemple [commentjyvais.fr](#) pour la Basse-Normandie, [jymalin.fr](#) pour la région Centre ou encore [Breizgo](#) pour la région Bretagne. Le SIM de la région Île-de-France est le bien connu [ViaNavigo](#).

Il faut cependant signaler que ces calculs d'itinéraire, si pratiques et orientés utilisateurs soient-ils, sont rendues disponibles uniquement via ces sites, ou parfois via une application liée ; il n'est pas encore question de les mettre à disposition via une API pour une utilisation externalisée. Au vu de cette observation, on comprend encore mieux pourquoi l'API *navitia.io* de CanalTP est une opportunité importante pour la communauté des développeurs...

Mai 2013



Les SIM existants au niveau régional, état des lieux en 2013

Source : http://www.setra.equipement.gouv.fr/IMG/pdf/JourneeITS_12_septembre-ATELIER_1_Informations_multimodale.pdf

Initiatives des opérateurs privés

Parfois, l'AOT n'est pas l'initiatrice de la libération des données de transports, et c'est l'exploitant qui s'en charge. On a déjà parlé du rôle de pionnier et du volontarisme de l'opérateur privé Kéolis, en partenariat avec l'agglomération de Rennes, qui a joué un rôle de modèle pour la libération des données de transport en commun. Nous allons maintenant opérer un focus sur les opérateurs de transport parisiens, c'est-à-dire la RATP et la SNCF, qui subissent naturellement depuis le début de ces débats une forte pression pour libérer leurs données, alors qu'ils n'y sont pas tenus de par leur statut d'EPIC.

Comment gèrent-elles cette pression et quelle est leur politique concernant ce sujet ?

La RATP : un virage rapide et récent

On a pu l'entrevoir avec l'affaire CheckMyMetro : la RATP s'est montrée très réticente, jusqu'à il y a peu, à partager ses données, considérant qu'elles faisaient partie intégrante de leur modèle écono-

mique, et que les partager serait synonyme de manque à gagner ; d'ailleurs, jusqu'à l'année 2011, le seul moyen d'accéder aux horaires était via leur application mobile Premium, à 1,29€ sur les stores.

Mais l'affaire CheckMyMetro, qui a éclaté en parallèle de la médiatisation de l'open data en France, a clairement mis en lumière ces réticences — on pourrait dire des *manquements* —, et sous pression de l'opinion, la RATP a commencé alors à envisager l'ouverture de ses données, et les avantages que cela pourrait lui amener. Après l'organisation d'un *barcamp* portant sur la cartographie des transports en commun fin 2011, ayant permis de créer un premier contact avec la communauté de développeurs et de militants, elle a libéré quelques premiers jeux de données sur le portail gouvernemental data.gouv.fr, puis est passée à la vitesse supérieure en septembre 2012 en lançant en grandes pompes son propre portail open data data.ratp.fr.



Les données mises à disposition sur ce portail concernent l'ensemble du réseau d'Île-de-France, soit les 16 lignes de métro (+ Orlyval), les 2 lignes de RER (A et B), les 4 lignes de tramway et 353 lignes de bus.

Elles sont au format GTFS, et constituent en fait le jeu de fichiers rendu obligatoire par cette norme. Il est donc possible de télécharger, soit pour toutes les lignes, soit pour une seule ligne choisie, l'ensemble des données suivantes :

- les horaires théoriques, selon chaque période du calendrier ; elles ne sont jamais fournies en fréquence (pas de fichier *frequencies.txt*), mais le portail demande aux réutilisateurs de présenter par eux-mêmes l'information sous forme de fréquence si le contexte le demande
- L'ensemble des arrêts nommés et géolocalisés (dont 12000 arrêts de bus !)
- Les temps de correspondance à pied ou entre deux stations (*transfers.txt*) ; on imagine que la mise à disposition de ce fichier, le seul qui n'était pas obligatoire, répond à une volonté de la part de la RATP de contrôler les temps de voyages annoncés par les calculateurs, plutôt que de les laisser annoncer un temps de trajet plus long qui les desservirait...

Ces jeux de données, assurément les plus intéressants pour les développeurs potentiels, sont complétés par des jeux de données plus classiques (au format CSV) concernant par exemple la billettique (liste géolocalisée des revendeurs agréés) ou des statistiques (trafic entrant par station et par an), destinés plutôt à une consultation directe ou à une réutilisation par des journalistes.

Un jeu de données assez complet donc, accompagné d'une politique assez volontaire, et d'un discours très encourageant sur les réutilisations. De plus la licence choisie pour les données GTFS est la licence ODbL, d'inspiration moins "économique" qu'Etablab et encourageant la réutilisation et le "repartage" (et ce même si la RATP entreprend cette démarche en étroite relation avec Etablab), et surtout, la RATP accompagne cette démarche de nombreuses rencontres et rendez-vous avec la communauté open data, ponctuant l'avancée de

son ouverture de données en tissant des liens forts et en restant à l'écoute de cette communauté.

Ainsi, en plus de divers *barcamps*, la RATP a lancé un grand concours d'applications pour encourager et médiatiser les réutilisations de ses données, nommé **OpenDataLab**, dont la cérémonie de lancement a coïncidé avec le lancement du portail, et dont la remise des prix a été donnée le 23 octobre.

Globalement, la démarche de la RATP a été très bien reçue par les acteurs de l'open data. Un bémol est à mettre, toutefois, sur la fréquence de mise à jour des données. En effet, les données horaires sont rarement mises à jour (sur plus de deux mois glissants), alors que l'offre réelle change bien plus fréquemment, ce qui a pu créer des problèmes et réclamations lors de hackathons.

Le STIF, Autorité Organisatrice de Transport à part

Le STIF, par contre, n'a pas changé de politique en matière d'open data depuis des années. En effet, il ne met aucune donnée à disposition, mais privilégie une approche « *open service* », sur deux modèles :

- Pour les demandes ponctuelles à visée limitée (recherche scientifique, études internes...), le STIF libère gratuitement un jeu de données, pas actualisable, sans autorisation de partage ou de réutilisation
- Une API d'accès dynamique à leurs données, sous licence payante, est prévue pour la mise à disposition d'un site ou d'un service d'informations voyageurs. C'est ce qui a été fait avec Mappy.fr, qui propose donc un calcul

d'itinéraire via les transports en commun en Île-de-France, en utilisant cette API.



Le STIF a aussi son propre système de calcul d'itinéraires (motorisé par Navitia), appelé ViaNigo, qui consiste en un site et une application. Là encore, pas de données ouvertes, et surtout, une offre assez incompréhensible pour le grand public, qui a donc le choix entre les services de la RATP, du STIF et de Transilien, par exemple, pour calculer ses itinéraires en Île-de-France, alors que ces entités sont sensées être unies par une même politique, par le biais du STIF.

La SNCF, une politique réfléchie et volontaire

La SNCF a entrepris des démarches après la RATP, mais a bien vite rattrapé son « retard » en la matière.

En juin 2012, elle ouvre son portail data.sncf.com, mettant à disposition les premières données (statiques et théoriques, au format GTFS) des lignes Transilien, TER et Intercités.

L'initiative est bien sûr très bien reçue, avec une retenue de la part des observateurs, qui concerne le choix de la licence accompagnant les données. En effet, la SNCF a choisi de créer sa propre licence, nommée « *Licence Open Data* ». Solution tentante pour qui serait tenté de contrôler au maximum les réutilisations, le choix est politiquement très mal vu, car comme expliqué plus haut,

une licence originale est plus compliquée à comprendre par les réutilisateurs, et surtout légalement complexe à partager pour qui serait tenté de les remettre à disposition.

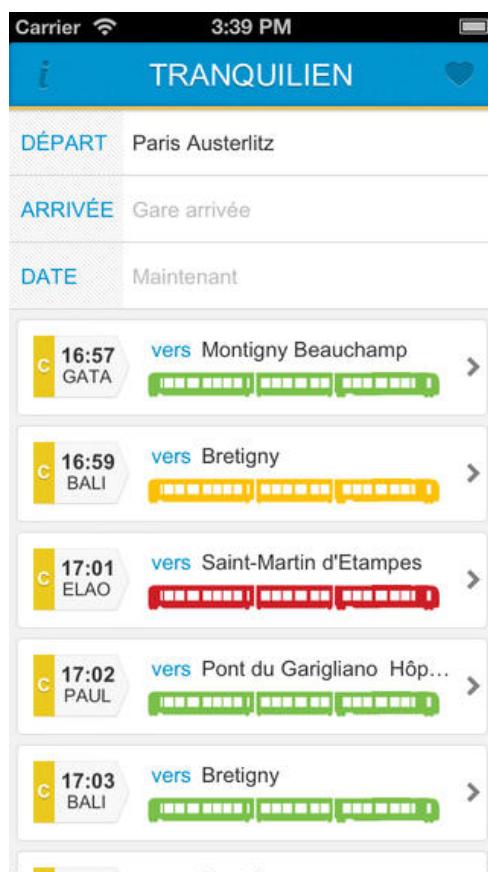
En parallèle, pour stimuler l'opération, elle lance le concours d'idées « **Open App** ». On pourrait croire que, sur le modèle de la RATP, ce concours serait destiné aux développeurs uniquement, mais le projet est ici un peu différent : il s'agit, avant de se lancer dans la création d'applications, de demander aux usagers des trains des idées d'applications qui pourraient être créées à partir des données. Ainsi, on s'adresse à tout un chacun en gardant à l'esprit l'intérêt des utilisateurs finaux tout au long du processus. Un site¹ est lancé en support à l'opération, pour recueillir les idées et les votes.

L'idée gagnante s'appelle *Transifoule* : le principe est une application crowdsourcée, qui permettrait aux usagers de signaler la saturation des rames ou des lignes en temps-réel, afin de réguler par eux-même l'abondance.

Le concours d'idée est suivi, très vite, des « **Hack Days** », hackathon de 48 heures réunissant des développeurs, qui, forts des idées des voyageurs, sont chargés de réaliser des prototypes d'applications. La SNCF en profite pour présenter pour la première fois des API temps-réel en test, donnant les prochains horaires des lignes de transiliens C et L.

Une émulation se fait, des prototypes sont créés, des équipes se forment, mais avec le recul, la seule réelle application qui sortira de cette opération est l'application **Tranquilien** : prototype créé durant les Hackdays à partir du concept Transi-

foule, il devient le moteur de l'équipe, qui monte une entreprise, **SNIPS**, dans la foulée pour porter le projet. C'est réellement le premier exemple d'application qui trouve un public à long terme, et qui va au-delà du concept et du prototype pour trouver une maturité.



Capture d'écran de l'appli Tranquilien

Source : <https://itunes.apple.com/fr/app/tranquilien/id660721122>

Puis en 2013, la SNCF (et plus spécifiquement, Transilien) s'est lancée dans un nouveau projet de grande envergure, qui diffère des précédents par sa thématique : le projet **Hackcess** est lui centré sur l'accessibilité, dans les trains comme en gare, et a pour but d'« accélérer la création de services connectés centrés sur les besoins des personnes à mobilité réduite ».

¹ <http://opendata.transilien.com/>, produit par June 21st.

HACKCESS TRANSILIEN

Dans la même veine qu'OpenApp, le programme comprend quatre grandes étapes, dont la toute première est nommée « *Imaginer* » : on n'est plus dans le crowdsourcing en ligne, mais un atelier de 48 heures est organisé le 14 septembre pour poser les problématiques et commencer à dessiner des solutions, avec des équipes composées de profils divers, dont des designers, des développeurs et des personnes à mobilité réduite.

L'étape 2 consiste à « *Cartographier* » les gares, et est faite main dans la main avec OpenStreetMap. On assiste alors à la naissance de nouvelles problématiques : comment représenter des problèmes d'accessibilité en gare ?

Plus globalement, ce projet est intéressant car il tend à sortir des éternels questions « horaires + arrêts », pour se pencher sur d'autres problématiques, comme celle de la vie en gare, et oblige à se pencher sur l'usager en tout premier lieu... Par ailleurs, la problématique du handicap créé des limitations dans les possibilités de solutions, et stimule la créativité en poussant les développeurs à sortir des éternelles applications mobiles, pour se pencher sur des solutions plus tactiles et plus humaines, comme des objets connectés. Aussi, le fait de situer l'utilisateur en gare permet d'utiliser les équipements, comme les projections sur les écrans en gare.

La forme de cette ouverture

On vient de le voir, si les opérateurs de transport ouvrent leurs données, c'est souvent, conjointement avec leurs portails, par l'intermédiaire de rencontres avec les acteurs de l'open data, et toute la communauté de réutilisateurs. Bien sûr, c'est une très bonne chose (voire même une action indispensable). Mais lorsque l'on fait le tour des réutilisations des mois plus tard, l'on réalise que le bilan est faible. En effet, la très grande majorité des applications ne vont pas au-delà du statut de projet, parfois de prototype. On pourrait attribuer cet état de fait au format des **hackathons** et autres **barcamps**. En effet, il est difficile de faire vivre un projet au-delà des 24 ou 48 heures durant lesquelles il voit le jour; et de faire durer la volonté de créer un réel objet à destination des usagers, une fois l'agitation terminée.

La couverture médiatique autour de ces évènements peut être vue, pour les producteurs de la donnée, comme une opportunité de communiquer sur leur bonne volonté, et par les participants de se faire bien voir de la communauté en étant au bon endroit au bon moment.

De même, si les **concours** sont un bon moyen de stimuler la concurrence en posant un besoin, se pose le souci de la redondance des solutions proposées, ainsi que de leur pérennité.

Il y a encore des choses à faire de ce côté-là, pour trouver de meilleures façons de travailler, à long terme, dans la sérénité, mais aussi de façon appuyée, en posant des jalons. Cela fait certainement partie des défis à relever à l'avenir ; en effet, en terme d'open data, tout est encore à inventer.

Les réutilisations

Les réutilisations des données de transport sont nombreuses, de par leurs formes, leurs auteurs et leurs thèmes. Toutefois, la forme la plus répandue est l'application mobile, qui comme on l'a vu, en plus d'être dans l'air du temps, est totalement en adéquation avec l'usage des transports en commun et la mobilité. Nous utiliserons néanmoins le terme « application » de façon indifférenciée pour dénommer les applications mobiles et les applications web.

Nous pouvons faire une catégorisation assez fidèle et générale de ces applications :

- Les applications propriétaires des AOT ou des opérateurs, qui sont ici à la fois diffuseurs et réutilisateurs de leurs données : presque tous les SIM des régions françaises ont développé, en plus de leur site portail, une application mobile associée.
- Les applications que l'on peut qualifier de « juniors », réalisées souvent dans le cadre de concours ou de hackathons, par des développeurs individuels ou des équipes issues d'écoles d'ingénieurs
- Les applications « locales », réalisées par des entreprises spécialisées dans le transport (par exemple Transilien et SNIPS)
- Les applications « transverses », qui se concentrent sur un seul mode de transport alternatif, comme le vélo ou la voiture électrique
- Et enfin les applications, largement utilisées, des grands acteurs privés impliqués dans les

problématiques de mobilité, comme Google, Apple, Nokia ou TomTom

Finalement, on constate à l'usage qu'il est assez difficile pour des créateurs d'applications de mobilité autres que les AOT ou les grands acteurs de se faire une place au soleil, probablement victimes de la concurrence d'image, de qualité et de communication avec les grands acteurs, mais aussi de la difficulté de pérenniser, non seulement leur service face à des offres de données sans cesse redéfinies, mais aussi leur modèle économique dans cet écosystème complexe.

Quel avenir pour les données ouvertes de transport public ?

Une volonté publique

Dans la continuité de la création d'Etabl, l'État français affiche une volonté d'innovation et d'ouverture en terme d'open data. Dans le cadre de la Charte du G8 pour l'ouverture des données publiques, le plan d'action pour la France¹ prévoit d'ailleurs la tenue de six débats thématiques, dont un concerne les transports, dans le but de légiférer de façon plus précise. Il serait question de remettre en cause le principe d'exception pour les Etablissements Publics d'Intérêt Public...

L'Europe aussi pousse en direction de l'incitation à l'ouverture. Déjà en 2007, bien avant le lancement de ces débats, la directive européenne INSPIRE (*Infrastructure for Spatial Information in the European Community*) donnait des directives aux pays membres pour partager leurs données géographiques, selon des règles de mise en œuvre communes, ce qui montrait déjà une sensibilité pour le partage et pour la culture de la donnée.

Le décret français de 2012 qui établit le principe de réutilisation gratuite par défaut était par ailleurs la transposition d'une directive européenne, connue comme la « directive PSI » pour « Directive on the re-use of public sector information ».

Face à ces rapides évolutions, et à la prise de conscience du public, une forte pression est mise sur les opérateurs de transport, au premier rang desquels la RATP et la SNCF.

Des réticences compréhensibles

Libérer ses données relève d'une philosophie assez compliquée à aborder pour des entreprises privées telles que la RATP ou la SNCF, chez lesquelles subsiste une **culture très forte de protection de leurs informations**. L'ouverture des données demande de changer de point de vue, voire de philosophie, sur sa propre organisation, car cela nécessite d'accepter que l'innovation et l'apport de valeur puisse venir de l'extérieur : pas facile pour des entreprises anciennes qui se sont construites sur un modèle de monopole et de contrôle total de leurs process.

Et puis, il n'est pas facile de libérer des données sans avoir d'à priori sur leur réutilisation future. La cible, c'est-à-dire les réutilisateurs, est difficile à identifier en amont, et ce problème est d'autant plus difficile à résoudre que les réutilisateurs sont rarement en mesure de définir eux-même leurs besoins. Pour les exploitants, cela revient plus ou

¹ voir http://www.gouvernement.fr/sites/default/files/fichiers_joints/plan-action-france_version_francaise.pdf

moins à **naviguer à vue** sans promesse de retour de valeur.

Les opérateurs ont surtout une appréhension, à terme, de devoir **rendre compte de produits de réutilisations de qualité inégale**, et de voir leur image assimilée à des applications de mauvaise qualité ou non maintenues.

Le sujet étant relativement nouveau, les opérateurs de transport se heurtent aussi à un **manque de moyens** important : mettre ses données à disposition demande beaucoup de temps, de l'argent et des procédés complexes, pas encore documentés et donc relativement méconnus.

Il ne faut pas oublier, dans le cas de la RATP et de la SNCF principalement, la **peur de voir leurs services phagocytés par ceux de Google**. En effet, Google propose déjà avec Google Transit une application de calcul d'itinéraire efficace, mais surtout auquel les utilisateurs font confiance, vu la renommée de l'entreprise. La peur de voir Google mettre la main sur des données complètes de transport est amplifiée par la pression mise indirectement par Google, via le signalement, pour tout calcul d'itinéraire à Paris dans Google Maps, de la phrase suivante : « *Ces résultats peuvent être incomplets : toutes les agences de transport en commun ne nous ont pas fourni leurs informations* ».

12:37 - 12:59 (22 min)

Ces résultats peuvent être incomplets : toutes les agences de transports en commun de cette zone ne nous ont pas fourni leurs informations.



En ce qui concerne la SNCF, on peut imaginer que leur agence de voyage, Voyages-SNCF, redoute l'entrée possible de Google (via Google Wallet) dans le marché des agences de voyages proposant des trajets dans leurs trains, comme le fait par exemple Capitaine Train depuis leur condamnation de l'Autorité de la Concurrence en 2009.

Les pistes à explorer

Dans tous les cas, les opérateurs ont déjà pris conscience de l'importance du mouvement : **l'ouverture des données est inéluctable**, car demandée par trop d'acteurs extérieurs, autant parmi les citoyens qu'au sein de l'État.

De toute façon, ils ont aussi réalisé que **s'ils n'ouvrent pas leurs données, ceux qui les veulent le feront par eux-mêmes** :

- En utilisant le *scraping* des données depuis les sites de l'exploitant, comme ont été nombreux à le faire les développeurs¹ dans le cas de la RATP depuis son site mobile (ce qui a été le cas de CheckMyMetro)
- Ou carrément, pour les horaires théoriques, copiés depuis les tables papier ou PDF éditées par les exploitants, ce qui ne garantit pas que les réutilisations utilisent des données "fraîches" et pourrait être préjudiciable au service final
- Soit par des moyens mixtes, voire indéterminés (on pense aux données de transport d'Île-de-France qui se sont retrouvées sur Google Transit sans aucun accord préalable, y compris les tarifs, qui ne sont pourtant pas partagés en open data)

Nous proposons ici des pistes à envisager, en regard de ces observations, pour les détenteurs de données de transport qui souhaitent libérer leurs données.

Y mettre de la bonne volonté

Bien sûr, à long terme, rien ne peut fonctionner si le propriétaire des données entreprend une démarche d'ouverture de données sans réelle volonté de la faire fonctionner. Cela peut paraître évident, mais il faut **aller au-delà de l'effet d'annonce**, voire du « coup de communication » que l'on a pu percevoir dans les évènements organisés par la RATP par exemple.

Il faut savoir animer une communauté et la faire se sentir partie intégrante du projet, mettre les choses à plat, et admettre aussi que créer et entretenir un écosystème prend du temps. En d'autres termes, il faut **faire confiance à la pérennité** et aux bénéfices du projet, même à long terme.

Parfois, cela nécessite de **convaincre aussi en interne**, surtout dans les grandes entreprises où les organigrammes sont larges et morcelés. Dans cette perspective, un bon exemple est celui de la SCNF, qui a organisé un hackaton en interne, le « Cheminots Hack », pour faire comprendre à ses employés qu'ils étaient les bienvenus pour faire partie intégrante de la démarche d'ouverture.

Encourager l'innovation ouverte

En fait, la démarche d'entreprendre sans présupposer des résultats à l'avance, mais en attendant des bénéfices, porte un nom : c'est ce que l'on appelle **l'open innovation**, ou innovation ouverte en français.

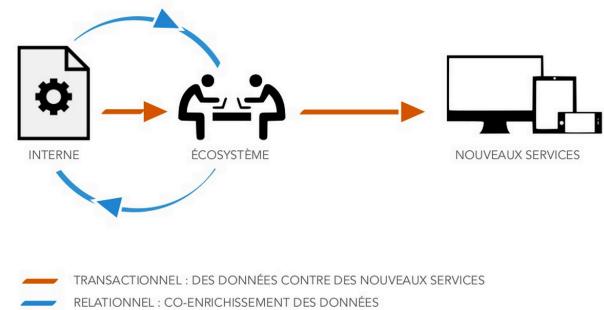
L'innovation ouverte incite à l'ouverture, au partage et à la collaboration entre acteurs, ce qui est presque antinomique avec les valeurs d'entreprise, mais dans la continuité totale de l'esprit open data.

Un compromis intéressant consisterait, pour les entreprises privées, à s'inspirer de ces valeurs pour développer leurs programmes d'ouverture de données : puisque l'open data est né du monde de la recherche et de l'open source des développeurs, pourquoi ne pas s'inspirer de leurs mode de

¹ Un exemple est visible sur Github : <https://github.com/pgrimaud/horaires-ratp-api>, ou sur API Tam, une API créée par un développeur Montpelliérain pour mettre à disposition les données de l'opérateur Transdev aux autres développeurs : <http://modulaweb.fr/apitam/>

travail : rapide, adaptatif, en constante remise en question ?

Il s'agirait aussi de travailler en collaboration ouverte avec les réutilisateurs, ne pas attendre que tout soit parfait et maîtrisé avant d'ouvrir, tenir compte de leurs retours et travailler ensemble vers l'amélioration, en fait, ne pas hésiter à proposer des données en **version bêta**.



« **De l'open data transactionnel à l'open data relationnel** »

Source : présentation de Romain Lalanne (responsable open data à la SNCF) pour l'OpenWorldForum 2013

Faire confiance aux destinataires de l'ouverture

Mais une démarche open data n'est rien sans ses destinataires cibles, et il faut savoir leur faire confiance pour aborder une démarche d'ouverture de données de façon sereine.

En allant encore plus loin dans le raisonnement, si l'on considère qu'il n'y a pas opposition entre soi et l'extérieur, il serait idéal, à terme, d'**utiliser les mêmes données en externe qu'en interne**, pour assurer une fluidité et une confiance totale entre les parties.

Les réutilisateurs

En tout premier lieu, le bénéfice le plus évident est celui du dialogue avec des réutilisateurs, et la capacité d'innovation qu'ils peuvent apporter en proposant des **solutions originales** : libérer ses données, c'est aussi une opportunité d'améliorer son organisation en interne et d'acquérir de nouveaux points de vue.

La multiplication des réutilisateurs, et donc des réutilisations, n'est pas à craindre : les applications mauvaises mourront d'elles-mêmes, il ne faut pas avoir peur de **faire jouer la compétition** entre réutilisateurs. En parallèle, **être exigeant sur la qualité des données fournies** permettrait de l'être en retour sur la qualité des réutilisations : n'oublions pas que l'open data est un échange et que toutes les parties ont des bénéfices à tirer du processus d'ouverture.

Les utilisateurs

Les utilisateurs finaux des services issus de réutilisation ne sont pas à négliger : s'il est naturel d'appréhender de perdre des utilisateurs de sa propre application pour des applications réutilisées, il ne faut pas oublier le bon sens — voire même la méfiance ? — du public, qui naturellement préfèrera aller à priori vers les solutions des « grands noms » et des émetteurs premiers de la donnée.

Surtout, il est possible de faire des utilisateurs des acteurs des applications finales, en intégrant le **crowdsourcing**, en particulier en ce qui concerne les données temps-réel, pour les perturbations de trafic par exemple. Les usagers des transport sont toujours les premiers à vivre ces perturbations, et donc les plus aptes à les diffuser rapidement, et à

en rendre compte de façon fidèle. Plutôt que de présupposer qu'ils seront réticents à faire remonter ce genre d'informations, il serait peut-être plus avisé de considérer que les usagers apprécieraient de se sentir acteurs de leur voyage. Il existe des moyens simples pour s'assurer que les informations remontées seraient de confiance, et d'éviter le "bashing" : par exemple en proposant un système d'alerte très simple (limitation du champ de liberté, pas de texte mais des choix), en demandant une inscription à l'entrée, ou en sur-pondérant les participations des membres les plus anciens ou les plus actifs. Comme des alertes crowdsourcées n'ont un intérêt qu'à partir d'une certaine masse critique de contributeurs, plusieurs leviers sont possibles pour encourager la participation : *gamification* de l'application, ou, plus envisageables pour les services des émetteurs de données, avantages clients pour les plus actifs... Ce que l'application *Tranquiliien* semble avoir oublié, et rencontre des difficultés pour fédérer une communauté de participants, alors que l'application repose entièrement sur cette condition.

Aussi, il est possible de réutiliser ces contributions pour en faire un jeu de données en open data : pourquoi pas une API temps-réel d'occupation des rames de Transilien, ou des fichiers CSV de statistiques, nourries par les contributions à *Tranquiliien* ? Cela permettrait d'améliorer le service dans un cercle vertueux, sur le modèle d'open data relationnel.

La problématique complexe des données temps réel

On l'a perçu au long de cette présentation : côté transports, le nerf de la guerre réside dans les données temps-réel, plus intéressantes au quotidien pour des usagers réguliers, et adaptées à une utilisation mobile dans des situations inattendues. Ce sont donc les plus convoitées par les réutilisateurs, mais aussi les plus jalousement protégées par les opérateurs, qui ont saisi là une opportunité de garder des utilisateurs pour leurs services maison.

Mais ce n'est pas la seule raison : nous avons vu que techniquement, l'alimentation d'un flux de données temps-réel nécessite des moyens assez lourds d'intégration avec le SAEIV, et des investissements humains importants.

Par ailleurs, le "format" API pose des questions : l'obligation d'héberger la source du flux sur le serveur de l'émetteur oblige à une limitation dans les accès, pour éviter la saturation du serveur : comment envisager, limiter et financer cette infrastructure ?

Surtout, quel modèle économique pour ces API ? Faut-il envisager un modèle d'open service, à la façon du STIF ?

Les enjeux de l'intermodalité

La difficile interopérabilité

Aujourd'hui, il est très difficile en France d'envisager un système de données unifié, que cela soit :

- Au niveau géographique et **politique** (les conseils régionaux AOT sont indépendants, et ont des politiques de mise en oeuvre très différentes, si jamais elles existent)
- Au niveau **technique** : les formats et type de données sont aussi très différents d'une source à l'autre
- Au niveau **juridique** : avec la multiplication des licences, voire la création de licences maison, faire cohabiter tous les jeux de données disponibles relève du casse-tête.

Etalab est en train de légiférer sur le sujet, tout du moins de réfléchir à des normes, mais tout est à faire dans ce domaine.

L'intégration avec les moyens de transport alternatifs

Puisque les infrastructures sont matériellement limitées, et surtout pour coller à la réalité des usagers, les exploitants de transport public et les réutilisateurs ne doivent pas oublier les autres moyens de transport, et à l'avenir les intégrer dans les réutilisations ; en effet, de nombreux trajets quotidiens ne se font pas uniquement via les transports en commun : la marche, le vélo, la voiture sont souvent utilisées, en particulier en banlieue, pour rallier les gares.

La SNCF a déjà commencé à s'orienter stratégiquement dans cette direction, en lançant un programme au long cours qu'elle a dénommé « **porte à porte** » (par opposition à « gare à gare »). Dans le cadre de ce programme, elle a acheté en 2013 la société de covoiturage *123envoiture*, et se voudrait « intégrateur de transport ».

De même, la RATP se distingue sur la même ligne, puisque l'application gagnante de son grand concours OpenDataLab, en octobre 2013, a été l'application mobile *Sharette*, qui permet de planifier des trajets quotidiens en entreprise en mêlant transports en commun et covoiturage.

Faire la guerre à Google ou travailler ensemble ?

Une crainte répandue parmi les exploitants concerne Google, et le risque de le voir mettre la main sur les données pour ensuite les intégrer dans ses services, ultra-concurrentiels en comparaison de toute réutilisation possible, voire même des applications des exploitants. Des craintes compréhensibles : si Google proposait les horaires temps-réel et les alertes trafic de la RATP dans toute sa suite logicielle (Maps, Navigation, Google Now) intégrée nativement sous Android (plus de la moitié des smartphones en France), il y a fort à parier que les utilisateurs déserteraient l'application de la RATP. Ceci étant dit, une fois que les données sont libérées, elles le sont pour tout le monde (la non-discrimination des réutilisateurs fait d'ailleurs partie des principes fondateurs de l'open data). Alors comment gérer cette peur de voir un acteur phagocyter l'ensemble de l'écosystème ?

Il serait tentant de bouder le format GTFS, pour compliquer l'import et la lecture des données par Google ; mais considérant que cela ne l'a visiblement pas empêché d'intégrer des données de la RATP qui n'avaient pas été libérées, et vu l'en-gouement des développeurs et la qualité des réutilisations issues de ce format en Île-de-France, il serait fort dommage de se priver de ce standard...

À terme, la meilleure solution sera probablement de collaborer, comme l'ont déjà fait Kéolis Rennes et la Communauté Urbaine de Bordeaux, en four-nissant leurs données de transport au programme Google Transit. En effet, un scénario probable — pour Paris, notamment — est que si Google n'a pas d'accord pour utiliser les données, il les récu-pèrera de lui-même sans permission au nom du "bien public" ; il est même possible de les voir aller jusqu'au procès, et considérant que l'opinion pu-blique leur est acquise, et que les répercussions d'images seraient catastrophiques pour les exploi-tants.

Si l'exploitation des données par Google est iné-luctable, alors comment rivaliser ? Des leviers sont possibles.

Pour les réutilisateurs indépendants, il faut con-sidérer que si pour l'instant, la plupart des applica-tions proposent uniquement des fonctionnalités très classiques d'affichage d'horaires et de calculs d'itinéraire, cela est dû à la nouveauté, et peut-être à quelques manquements de la part des ex-ploitants pour présenter les informations de façon réellement centrée sur l'expérience utilisateur. Mais à terme, les applications indépendantes qui fonctionneront seront celles qui, comme pour toutes les applications disponibles sur les stores, pro-po-poseront un angle et une vision originale à l'utilisa-

teur, ou des cibles très identifiées ; ce qui n'est pas vraiment possible pour Google, qui se posi-tionne depuis toujours sur une offre de services neutre et pour tous. Pas de réel danger de concur-rence donc, mais plutôt une saine complémentari-té.

Les exploitants de transport, par contre, ont plus de souci à se faire, étant sur le même créneau. La meilleure solution, à notre avis, consisterait à s'inspirer des bons produits de Google, et de compléter ces fonctionnalités « de base » avec des services que l'exploitant peut être le seul à pro-po-poser :

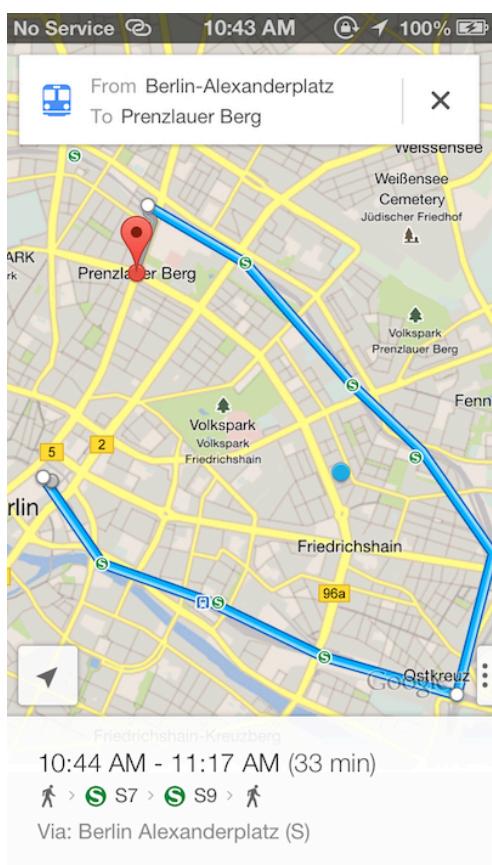
- Programmes de fidélité intégrés à l'application
- Possibilité d'intégrer des futurs services en gare (courses en lignes à récupérer, services divers), en cours de conception dans plusieurs gares parisiennes, dans l'application, pour proposer un voyage "sans couture", dans la lignée du programme « porte à porte » de la SNCF
- Pourquoi pas, à terme, intégrer des outils de billettique pour réserver ses billets au sein de l'application ?

Vers un assistant personnel de voyage

Si l'on commence à s'inspirer de Google pour voir le monde comme lui, il y aurait peut-être beaucoup à apprendre, pour les exploitants, et à imiter.

Depuis peu, Google intègre dans les smartphones équipés d'Android l'application **Google Now**. Son principe, simple mais très efficace, consiste à amener l'information à l'utilisateur sans qu'il ne

fasse aucune recherche : par exemple « Il est 19h30 et vous êtes encore au bureau, voici le trajet en transports en commun pour rentrer chez vous ». Cela est rendu possible par la connaissance profonde qu'a acquis Google sur nos vies : à force d'utiliser quotidiennement leurs services, nous leur avons fourni des informations intimes sur nos habitudes et nos besoins.



Capture d'écran de l'application Google Now

Pourquoi ne serait-il pas possible de s'inspirer de ce modèle pour proposer un tel « assistant de voyage » via les applications des exploitants ?

Cela ne se fait pas du tout encore, peut-être pour des raisons de frilosité de la part des grands groupes, qui craignent de se voir reprocher une at-

teinte à la vie privée, mais il y aurait de vrais bénéfices, pour toutes les parties, à voir se développer un tel type d'applications :

- Pour les utilisateurs, aujourd'hui habitués à fournir leurs données personnelles en l'échange de services efficaces et personnalisés
- Mais surtout pour les exploitants, qui ratent jusqu'ici une opportunité incroyable de mieux connaître leurs usagers et de s'adapter à leurs besoins. Agréger des informations de millions d'usagers des transports (tendance « big data ») serait bien plus facile que le comptage en gare (encore effectif aujourd'hui). Idéalement, il serait possible de s'inspirer, encore une fois, de Google, pour restituer ces données récupérées à l'utilisateur, sous forme de tableaux de bord, qui lui permettraient de se connaître mieux lui-même et d'améliorer son confort et ses habitudes¹.

On n'est plus vraiment, ici, dans une logique d'open data, puisque l'on parle de données personnelles, et d'un échange de ces données strictement entre l'usager lui-même et l'exploitant. Il s'agirait donc de **faire levier avec les données personnelles des usagers**, en complément des données publiques libérées.

¹ C'était le cas de Google Latitude, application de Google qui permettait de se géolocaliser en continu (et de visualiser la position de ses amis), qui proposait sur son site une page résumant tous ses déplacements et ses statistiques chaque semaine (temps passé au travail, kilomètres parcourus en un an...)

Monétiser, pourquoi pas ?

N'oublions pas que la loi prévoit la possibilité de l'instauration d'une redevance pour l'accès ou la réutilisation des données publiques. Si l'idée est loin d'être populaire dans le public aujourd'hui, peut-être faudra-t-il, à terme, envisager cette solution pour les données sensibles du transport comme le temps-réel, ce qui permettrait de contribuer au financement de leur mise à disposition — et donc de proposer des données de qualité.

Une piste intéressante serait de proposer un modèle *freemium*, c'est-à-dire un accès gratuit jusqu'à un certain nombre d'accès par jour, puis un modèle payant pour les réutilisateurs les plus gourmands. Cela permettrait de ne pas décourager les nouveaux entrants en les laissant expérimenter, puis de leur laisser le temps de tester leur modèle économique, mais aussi de faire collaborer les géants du web, comme Google, à l'économie de la donnée.

Conclusion

En tentant, au détour de ces pages, de définir les contours des relations entre open data et transports en commun, nous avons pu constater qu'ils sont de plus en plus complexes, mouvants, contradictoires ; réellement, en évolution extrêmement rapide.

Il nous faudra encore des années pour déterminer si des modèles économiques sont possibles, pérennes et surtout si chacun des acteurs concernés, de la plus grande autorité organisatrice de transport à l'usager occasionnel du métro, peut y trouver son compte. Tout est encore à faire, et on sent avec le peu de recul que l'on a à disposition, que le mouvement open data agit plus comme une lame de fond, remettant en cause les principes les plus établis et nous poussant à réorganiser profondément les institutions, que comme un séisme modifiant nos habitudes du jour au lendemain.

Les chantiers sont immenses, mais plus étrange, nous n'avons pas encore la capacité d'entrevoir leurs limites et leurs enjeux dans l'ensemble. Les yeux plongés dans le code, les applications, les hackathons, les formats de données, nous apportons chaque jour notre pierre à l'édifice de l'innovation ouverte, ce qui rend la tâche d'autant plus passionnante et nous montre tous les jours que les bénéfices ne sont pas dans la finalité, mais sur le chemin que nous parcourons.

Nous comprenons de plus en plus, au détour de ces chemins, que la donnée crée une valeur croissante grâce au partage, et qu'il serait temps de redéfinir nos usages de ces données à toutes les échelles, pour permettre à tout un chacun d'en tirer des bénéfices, économiques ou de services.

J'ai pour ma part apprécié me plonger dans ce monde passionnant qu'est l'open data, et espère bien continuer à en faire partie, voire y participer de façon plus active, dans les années qui viennent, sur la lancée de l'établissement de cette thèse professionnelle qui je l'espère aura permis d'intéresser ses lecteurs à ces problématiques.

Bibliographie

Open data, comprendre l'ouverture des données publiques, Simon Chignard, 2012 édition fyp, et le [blog](#) accompagnant l'ouvrage

[*8 Principles of Open Government Data*](#)

[*Open Definition*](#), par l'*Open Knowledge Foundation*

[*Democratieouverte.org*](#), plateforme francophone pour la promotion de l'open government

The Guardian [*Where does my money go ?*](#)

[*Transparency and Open Government*](#), memorandum publié par Barack Obama le jour de son élection :

[*Loi n° 78-753 du 17 juillet 1978* \(Loi CADA\)](#)

[*Définition du Logiciel Libre*](#) par la *Free Software Foundation*

[*Echelle des 5 étoiles de l'open data par Tim Berners-Lee*](#)

[*Linked Data, the Story So Far*](#), Tim Berners-Lee, Tom Heath, Christian Bizer

[*Déclaration des Droits de l'Homme et du Citoyen de 1789*](#)

Le site de la Commission d'Accès aux Documents Administratifs cada.fr

[*Charte du G8 pour l'Ouverture des Données Publiques*](#)

[*Carte des initiatives open data en France*](#), créée et maintenue par l'association LiberTIC

Site de l'[*Open Data Index*](#)

Site de l'association *OpenData France*

<http://opendatafrance.net/>

[*Open data : quels enjeux et opportunités pour l'entreprise ?*](#), livre blanc publié par Bluenove

[*Spécifications de la norme Transmodel*](#)

[*Spécifications du format GTFS*](#)

[*Spécifications du format GTFS-realtime*](#)

[Portail du programme de partenariat Google Transit](#)

[Loi n° 82-1153 du 30 décembre 1982](#) d'orientation des transports intérieurs (LOTI)

[Décret n°49-1473 du 14 novembre 1949](#) relatif à la coordination et à l'harmonisation des transports ferroviaires et routiers

[Blog R&D de CanalTP](#)

Libre blanc Orange [transport collectif : l'ère du voyageur numérique](#)

[Site de la PREDIM](#) (Plateforme de recherche et d'Expérimentation pour le Développement de l'Innovation dans la Mobilité)

[L'open data dans le domaine du transport : analyse des premières initiatives et recommandations](#), rapport de l'Agence française pour l'Information multimodale et la Billettique

[Transports du futur](#), blog de Gabriel Piassat sur l'évolution de la mobilité

[TransID](#), blog de Yann Le Tilly sur la mobilité connectée et l'information voyageur

Glossaire

Datajournalisme (ou journalisme de données) : désigne une nouvelle façon de faire du journalisme, en se basant sur des données brutes pour établir des faits, ou au contraire en livrant au public ces données, souvent présentées sous forme de visualisations interactives.

API (Application Programming Interface) : interface de programmation qui permet de se « brancher » sur une application pour échanger des données.

Crowdsourcing : désigne le fait d'utiliser les internautes ou les utilisateurs d'une application mobile pour améliorer le service, grâce à leur participation, sur le principe d'intelligence collective.

Scraper (de l'anglais « scrape », racler, rassembler) : action d'aspirer des données depuis un site web qui n'a pas mis ces données à disposition de façon ordonnée, grâce à un programme qui parcourt ses pages et récupère ces données, de façon légale ou non.

CADA (Commission d'Accès aux Documents Administratifs) : Autorité administrative indépendante dont la mission est de faciliter et de contrôler l'accès aux documents administratifs par les particuliers, conformément à la loi du même nom. Depuis 2005 et la modification de cette loi en faveur de la réutilisation des données publiques, la CADA possède aussi des compétences et arbitre des décisions dans ce domaine.

EPIC (Établissement Public à caractère Industriel et Commercial) : statut du droit français désignant une personne morale de droit privé ayant pour but la gestion d'une activité de service public. Ils sont privés et soumis à la concurrence, mais assurent une mission de service public. Ex : l'INA, la SNCF, la RATP, l'ONF, Méteo France...

LOTI (Loi d'Orientation des Transports Intérieurs) : loi fondamentale d'organisation des services publics de transport en France, qui définit les droits des utilisateurs des transports, mais aussi les relations entre les diverses autorités organisatrices de transport.

AOT (Autorité Organisatrice de Transports) : collectivité publique à laquelle la LOTI a confié la mission de définir la politique de transport en son sein. Les régions, les départements et les communes sont des AOT, ainsi que le STIF en Île-de-France — dans le cas des communes, on parle d'*Autorités Organisatrices de Transport Urbain*.

STI (Systèmes de Transports Intelligents) : désigne, de façon générale, l'application des nouvelles technologies de l'information au domaine des transports (information voyageurs, billettique sécurité, gestion du parc...)

SIM (Systèmes d'information multimodaux) : Services mis en place par les régions, fortes de leur statut d'AOT, proposant une information complète sur les offres de transport locales (horaires, plans), mais aussi, le plus souvent, un système de calcul d'itinéraires.

Barcamp : Rencontre ouverte sur un sujet précis, dont le principe est de faire participer les personnes présentes en leur faisant produire quelque chose pour faire avancer le sujet, en réactions aux conférences considérées comme unilatérales, académiques et moins dynamiques.

Hackathon : mot valise composé des mots *hack* et *marathon*, c'est un évènement, généralement de plusieurs jours, durant lequel des développeurs se réunissent pour faire de la programmation collaborative, et inventer des solutions très rapidement. On pourrait considérer que c'est un barcamp de développeurs.

Remerciements

Je souhaite remercier, en tout premier lieu, mes collègues de travail, qui m'ont accueillie de la meilleure façon possible, et ont contribué à faire de ce stage une période agréable et enrichissante. Particulièrement, je remercie Joachim Breton, mon tuteur de stage, pour sa disponibilité et ses conseils sur le sujet vaste et complexe de la mobilité connectée.

Je remercie aussi mon suiveur à l'INA, Matthieu Richy-Dureteste, pour l'intérêt qu'il a porté au sujet de ce mémoire, ses encouragements et sa disponibilité, et plus généralement l'ensemble de l'équipe enseignante de la formation CPM, qui a su m'orienter vers une voie professionnelle qui aujourd'hui m'apporte beaucoup et continue de le faire chaque jour.

Je remercie aussi vivement ceux de mes proches qui m'ont supportée, dans tous les sens du terme, lors de l'écriture de ce mémoire.