Pénombre

La lettre grise

Supplément à la Lettre d'information de Pénombre
association régie par
la loi du 1er juillet 1901

Troisième série. Volume XII.
Mars 2018, numéro 14
Fondée par Lucio Nero (*)

BIGOPEN DATA

14ème Nocturne de l'association Pénombre
Petite tentative pour dégonfler le big et ouvrir la discussion sur l'open

Pas un jour ne passe sans que l'on annonce qui un colloque, qui un think tank sur le big data, une docte réflexion sur l'open data…

Pénombre se devait de tenter à sa manière de lever un coin du voile sur ce sujet envahissant tous les domaines d'activités (santé, mobilité, marketing, publicité, finances, etc.)

Il faut bien reconnaître qu'il s'agit presque d'une friandise, tant les nombres sont présents dans les débats autour de ces concepts… mais il n'y a pas que des nombres et loin de là dans les chaudrons du bigopen data…

Nostalgie des nombres élaborés scrupuleusement à partir d'échantillons rigoureusement calculés face à un tapis de données prétendant plus ou moins à l'exhaustivité ?

Promesse d'applications porteuses d'utilités nouvelles, déferlement numérique de sollicitations publicitaires fondées sur des corrélations basiques et parfois surréalistes, enfermement de la vie quotidienne dans un monceau de données produites par les individus, consciemment ou non…

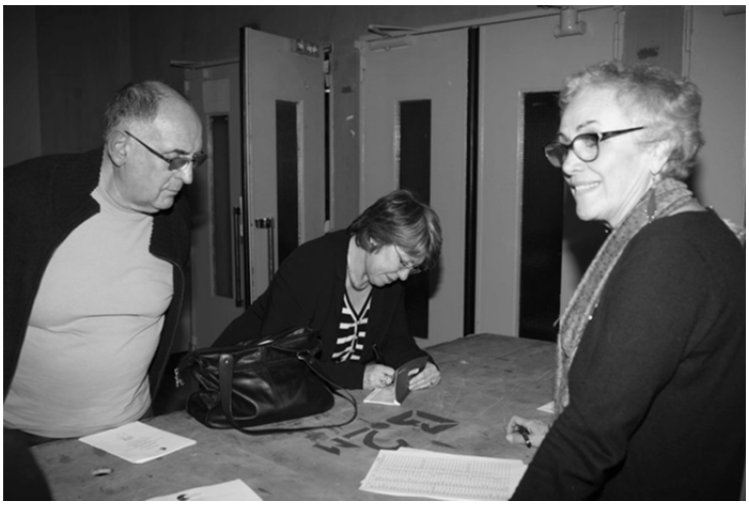
Avec une interrogation de bon sens en toile de fond : collecter, capter et engranger des données, c'est évidemment très tendance ! Mais se pose-t-on toujours, souvent, rarement ou jamais la question centrale à nos yeux : **pour quoi en faire !**

Et les enjeux sociétaux et démocratiques dans ce maelström ?

Nous rendons compte dans cette lettre grise de la nocturne consacrée à ce sujet en présentant les interventions de nos invités ainsi que des questions et apports qu'elles ont suscités dans le public.

ACCUEIL DU PUBLIC ….

Avant



Pour démarrer

Les débats sont ouverts par Fabrice Leturcq, président de Pénombre, mais ses paroles se sont envolées sans être enregistrées…

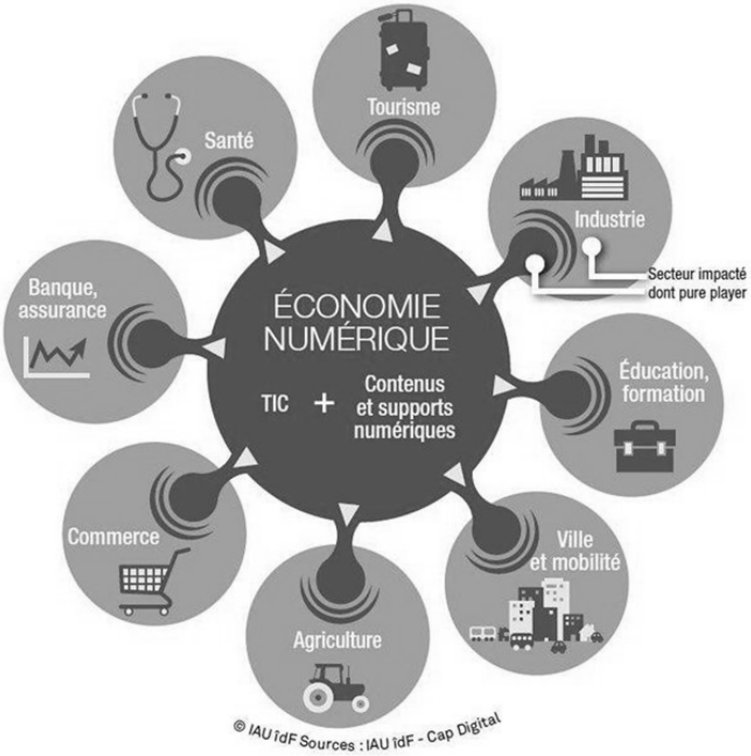
1. BIG DATA ET OPEN DATA DANS LE MÊME BATEAU ?

Alain Tripier

Jean-René Brunetière anime la soirée

Les notions de big data et d'open data sont abondamment utilisées et mises à toutes les sauces par les médias. Elles correspondent essentiellement à la profusion de données numériques existantes et exploitables.

Le big data évoque souvent l'exploitation de données à des fins commerciales, dans des applications marketing et publicitaires. Le champ d'application est beaucoup plus vaste comme l'indique le schéma suivant proposé par l'IAUIdF (Institut d'aménagement et d'urbanisme de la région d'Ile de France) pour décrire l'économie numérique.



Celui d'open data correspond assez clairement à la **mise à disposition** de tous de données issues des services des États, des services publics et assimilés. Ce qui n'exclut pas qu'une exploitation commerciale ait lieu à partir de ces données.

Bien entendu l'open data suppose de disposer de data, plus ou moins big, structurées plus ou moins également !

Pour les puristes de la langue française, on peut proposer une traduction à partir du thème retenu pour les journées de la statistique SFdS 2014 « Données en grande dimension ou Big data ».

L'affaire ne date pas d'hier puisque, selon le journal suisse *Le Temps* :

> Le premier scientifique de données est John Tukey, un statisticien (1915–2000) qui aurait fait de la science des données une éthique, celle du «Data First». Plutôt que de partir d'une hypothèse et d'en vérifier la valeur, il aurait montré la nécessité de partir des données et d'en déduire des enseignements.

Le Big data

Big data et marketing sont intimement liés ; cette émergence des data résulte naturellement du développement des TIC, des capacités de stockage quasi infinies et d'outils de traitement très puissants.

La définition communément admise par les spécialistes, retient trois points de base :

- grand volume de données ;
- importante variété de données ;
- vitesse de traitement.

Auxquels s'ajoutent deux autres « V » :

- valeur ajoutée, création de valeur ;
- véracité, données fiables (!)

Dans le domaine du marketing trois sources d'approvisionnement en données se complètent et se combinent. Il s'agit de gérer et d'exploiter les données :

- que l'on possède ;
- que l'on achète ;
- que l'on obtient.

Ces trois actions correspondent à un acronyme en vogue dans cet univers : le POE qui signifie Paid, Owned, Earned. Quelques mots pour en donner la signification. Le Paid correspond à des données achetées par l'entité concernée, par exemple des fichiers adresses ; le Owned à ce qu'elle possède, son fichier clients/adhérents et les transactions effectuées avec les membres desdits fichiers ; le Earned est « gagné » sans bourse délier, comme par exemple les données récoltées sur les pages de cette entité sur les réseaux sociaux.

Les « traces » laissées par nos usages de l'internet constituent une source d'approvisionnement considérable en data ; parmi elles, rappelons les informations collectées par cookies, favoris, paramètres et préférences de navigation, passage sur les réseaux sociaux, consultation des sites des entreprises… Pour l'instant ces données sont reliées au numéro IP de la machine connectée, mais pas de soucis, Google cherche une solution et développe un programme permettant de probabiliser les caractéristiques de la personne au clavier !

Heureusement pour nous, actuellement les mobiles et les tablettes n'enregistrent pas de cookies, mais… la « parade » sera bientôt trouvée avec un nouveau numéro d'identification !

Tous les passages de l'internaute sur la toile peuvent être enregistrés et traités en temps réel. Le résultat de ces actions conduit à se voir proposer un produit ou un service dans une publicité en ligne correspondant à la recherche que l'on est en train de faire. De même, et de façon particulièrement insistante, des bannières de publicité sur ce type de produit s'affichent systématiquement dans les heures et les jours qui suivent la recherche initiale, que l'on ait acheté ou non. Plus sophistiqués sont les algorithmes qui savent abandonner le prospect quand un achat a eu lieu !

Tout cela résulte des procédures du big data, comme la vente de la publicité qui arrive sur l'écran, vendue aux enchères en temps réel ! Dans le jargon publicitaire, il s'agit de RTB (Real Time Biding)

Les nouveaux enjeux immédiats tiennent aux « objets connectés » :

- les TV connectées. Elles font le plus souvent double emploi avec la tablette ou le mobile également connectés utilisés simultanément par le téléspectateur. D'autres engins ou dispositifs sont ou seront prochainement connectés, voitures, réfrigérateurs, compteurs et divers produits électro-ménagers.
- les wearables, ou objets connectés personnels, telles les montres connectées et autres capteurs « portés » par les individus.

Les estimations les plus fantaisistes circulent quant au nombre d'objets connectés actuels et futurs, actuellement estimés entre 10 et 20 milliards dans le monde ; les prévisions à 2020 ou 2030 vont jusqu'à 80 milliards. On notera que ces prévisions, d'origine nord-américaine pour la plupart, sont toujours calculées à l'échelle mondiale.

L'exemple des compteurs électriques est intéressant. Aux USA les compagnies installent des compteurs connectés qui permettent au fournisseur et au client final de disposer en temps réel de l'information sur les consommations ; on imagine aisément le profit que peuvent en tirer les deux parties. En France, le test du compteur « intelligent » Linky d'EDF soulève quelques problèmes compte-tenu du fait que la voie de retour vers le consommateur ne semble pas être prévue…

Il est en tout cas très probable que les applications liées au logement vont redonner à la domotique, trop tôt mise en avant, un nouveau contenu. On consultera avec intérêt le site de la CRE (Commission de régulation de l'énergie) *http://www/smartgrids-cre.fr/index.php?p=technologies-emetteur-radio-linky* qui décrit le dispositif transmettant en temps réel les informations du compteur pour d'autres usages que celui réservé à EDF.

Lié aux objets connectés, le Quantified Self est, selon Wikipédia, un mouvement qui regroupe les outils, les principes et les méthodes permettant à chaque personne de mesurer ses données personnelles, de les analyser et de les partager. Les outils du Quantified Self peuvent être des capteurs, des applications mobiles ou des applications Web.

En mars 2013, le mouvement Quantified Self en France s'organise autour de MyDataLabs, association centrée sur la donnée personnelle. Les applications médicales sont en première ligne ; le coaching semble également prendre une place importante dans ce domaine. On peut supposer, ou même espérer, que ce mouvement mondial deviendra un pare-feu face aux débordements du marketing et de la publicité.

Quels que soient l'origine et le donneur d'ordre, marques commerciales, ONG, associations caritatives… le big data implique l'utilisation de puissants algorithmes développés en interne ou proposés par des prestataires spécialisés.

Le terme d'algorithme est sans doute celui qui revient le plus souvent dans la littérature spécialisée aujourd'hui.

Au passage, il semble que ces activités nécessitant des développements informatiques très lourds, soient une voie de recyclage et/ou d'avenir pour les cerveaux français en rupture avec les activités financières de trading haute fréquence !

Notons par exemple, qu'actuellement en Europe, plus de 50 sociétés spécialisées analysent en temps réel ce qui se passe sur les réseaux sociaux.

« En temps réel » est une expression, de plus en plus fréquemment associée au big data. Il s'agit là d'une réelle révolution pour les marketeurs et les publicitaires, qui voient leur pré carré envahi par ces démarches quantitatives et les process informatiques qui les sous-tendent.

Pour bien comprendre la genèse du phénomène il faut remonter une trentaine d'années en arrière.

Depuis le début des années 80, l'informatique réalise des progrès continus en matière de stockage et de capacité de traitement.

Au début de cette période les entreprises qui possédaient des datas, les VPCistes par exemple (La Redoute et les 3 Suisses notamment) eurent l'idée de confier à des prestataires extérieurs l'analyse d'extraits randomisés de leurs fichiers clients et des transactions correspondantes. À cette époque, il était impensable de traiter l'intégralité du fichier. Par ailleurs, personne ne disposait alors des programmes de fidélisation qui permettent aujourd'hui d'ajouter des caractéristiques sociodémographiques aux données brutes.

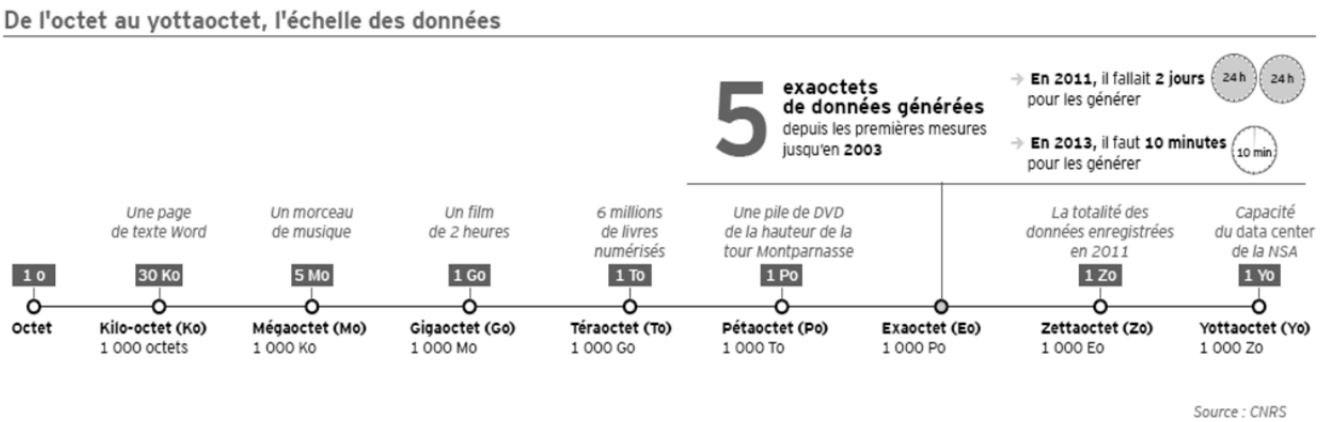
On savait tout sur les achats du client X, stockés sur une longue période, mais au-delà de son nom et de son adresse, on ne savait rien d'autre. Pour pallier ces carences on probabilisait le sexe des clients prénommés Dominique ou Claude, au moyen d'algorithmes travaillant à partir des achats effectués. On commençait à réaliser des fusions avec des fichiers externes sur la base de variables communes.

Cela aboutissait à des segmentations de clientèle, effectuées à partir de données passées, qui restaient opérationnelles un certain temps (plusieurs années le plus souvent) avant de remettre l'ouvrage sur le métier !

Le changement de paradigme opéré par le big data correspond à un mode de traitement de l'information qui tend vers le temps réel permettant un glissement rapide des segmentations statiques décrites ci-dessus vers des segmentations dynamiques mises à jour en continu.

Il est intéressant de souligner que les algorithmes fonctionnent essentiellement sur des calculs de corrélation sans recherche des causalités.

Quelques chiffres issus d'une publication d'Ernst et Young réalisée pour le forum culturel d'Avignon 2013 illustrent l'ampleur du phénomène. On pourra en retrouver le détail sur : http://www.forum-avignon.org/fr/etude-ernst-young-pour-le-forum-davignon



Un petit rappel des unités est sans doute utile :

1 kilooctet (Ko)	= 10^3 octets	= 1 000 octets	= 1 000 octets
1 mégaoctet (Mo)	= 10^6 octets	= 1 000 Ko	= 1 000 000 octets
1 gigaoctet (Go)	= 10^9 octets	= 1 000 Mo	= 1 000 000 000 octets
1 téraoctet (To)	= 10^{12} octets	= 1 000 Go	= 1 000 000 000 000 octets
1 pétaoctet (Po)	= 10^{15} octets	= 1 000 To	= 1 000 000 000 000 000 octets
1 exaoctet (Eo)	= 10^{18} octets	= 1 000 Po	= 1 000 000 000 000 000 000 octets
1 zettaoctet (Zo)	= 10^{21} octets	= 1 000 Eo	= 1 000 000 000 000 000 000 000 octets
1yottaoctet (Yo)	= 10^{24} octets	= 1 000 Zo	= 1 000 000 000 000 000 000 000 000 octets

En mots, un yottaoctet, c'est un million de milliards de milliards d'octets, une friandise pour les pénombriens !

Le big data n'épargne pas la recherche. Marc Lipinski, conseiller régional d'Ile-de-France et directeur de recherche au CNRS a participé le 6 décembre 2013 à une rencontre interdisciplinaire organisée autour de la question de l'ouverture des données massives scientifiques au CNRS. On retrouvera un entretien enregistré à cette occasion sur www.letudiant.fr. Cette rencontre fait le point sur les enjeux du big data. Elle a souligné que la recherche est aujourd'hui confrontée à un nombre croissant de données à traiter, analyser et stocker. Celles-ci ne proviennent pas seulement des chercheurs mais aussi d'un nombre important d'autres contributeurs non scientifiques qui ont la possibilité de participer à leur exploitation pour peu que ces données leurs soient ouvertes (open data).

Depuis, des dizaines de colloques plus ou moins sérieux ont eu lieu sur ce sujet….

L'open data

Trois grands principes fondent l'open data :

- un format ouvert ;
- la gratuité ;
- la liberté de réutilisation.

De fait, l'univers public est présent dans le big data depuis le milieu des années 2000, avec les premiers pas de l'open data, à partir de données détenues par les administrations et services publics et mises à disposition du public dans un souci de transparence et d'une recherche d'une plus grande efficacité de l'action publique. Les compagnies de transport anglo-saxonnes et japonaises ont été les premières à mettre à disposition du public des données de fonctionnement de leurs services, permettant à des développeurs de proposer des applications mobiles utilisables au fil des déplacements.

Les transports publics et la circulation sont les domaines de prédilection de l'orientation temps réel dans l'open data.

L'extrait de *La Gazette des communes* (30-10-2013) cité ci-après montre que la France n'était pas très en avance en ce domaine. On consultera avec intérêt le site de *La Gazette* bien documenté à ce jour ainsi que le site d'Etalab pour plus d'information sur l'Open Data.

« Selon le classement Open Data Index de l'Open Knowledge Foundation, présenté lundi 28 octobre, la France n'arrive qu'en 16 ème position sur 70 pays évalués. Cartographie, transports, dépenses publiques..., il reste des efforts à fournir pour se hisser au niveau de la Grande-Bretagne ou des États-Unis.

Ce premier classement d'envergure a été établi de façon collaborative par l'Open Knowledge Foundation (http://okfn.org/), une association qui milite pour la culture libre et promeut à ce titre la gouvernance ouverte.

Il a été établi selon l'ouverture de dix sets de données majeurs : résultat des élections, dépenses publiques, émission de pollution, etc. Chaque item est lui-même noté selon plusieurs critères : la licence, le format, la gratuité…, pour juger s'il respecte bien les principes fondamentaux de l'open data. http://sunlightfoundation.com/policy/documents/ten-open-data-principles/

« Si des progrès indéniables ont été réalisés ces dernières années, beaucoup reste à faire », regrette l'OKF. Sur les 700 jeux de données évalués, seuls 86 obtiennent le score maximal, soit 12%. (…)

En tête de ce classement, on retrouve les pays anglo-saxons et les pays nordiques, pionniers sur l'open data, Grande-Bretagne, États-Unis, Danemark, Norvège. La France n'arrive donc qu'en 16 ème place, derrière la Moldavie ou la Bulgarie, qui n'ont pas la réputation d'être particulièrement transparents. Bien que le pays ait signé la charte Open Data du G8 et soit engagé dans l'ouverture des données avec data.gouv.fr, trop de données publiques fondamentales restent indisponibles, déplore le chapitre français de l'OKF, ouvert cette année. La réutilisation et le partage des données sur les entreprises et sur les textes de lois restent soumis à redevance. Les cartes de faible résolution fournies par l'IGN en open data limitent les possibilités de réutilisation. Dans le secteur du transport, la SNCF ne publie toujours pas les horaires détaillés de ses trains à grande vitesse. Enfin, le détail des dépenses publiques reste hors de portée des citoyens. Les codes postaux ne sont également pas accessibles, alors qu'ils sont très utiles pour éviter des erreurs de géolocalisation dues à des communes homonymes et donc un nettoyage fastidieux des données.

Décrire une réalité complexe avec des métriques unifiées », Henri Verdier, qui dirige la mission Etalab en France, nuance ce résultat, quitte à tordre la définition de l'open data : « Un peu comme le classement de Shanghai pour les universités, l'Open Data Index a les inconvénients de ses avantages : il essaye de décrire une réalité complexe avec des métriques unifiées. Il adopte par exemple une définition très stricte de l'open data (format ouvert, gratuité et liberté de réutilisation). De ce fait, il récuse la qualité de « données ouvertes » à certaines données qui sont publiées par la France, gratuites et numériquement exploitables mais ne donnant pas lieu à droits de réutilisation. De même, il se concentre sur un cœur de jeux de données, et ne tient pas compte de toutes les autres données qui ont été partagées. »

Depuis la nocturne, la loi du 7 octobre 2016 pour une République numérique a été votée et promulguée. Le site *vie-publique.fr* en résume le contenu. *La loi a été rédigée à l'issue d'un long processus de concertation. Une consultation a d'abord été menée par le Conseil national du numérique, entre octobre 2014 et février 2015. À partir des contributions recueillies (plus de 4 000), une stratégie numérique a été présentée le 18 juin 2015 par le gouvernement. Puis une consultation publique a été organisée du 26 septembre 2015 au 18 octobre 2015 sur le texte de l'avant-projet de loi. Cinq nouveaux articles ont été retenus au terme de la consultation. Elle comporte trois volets :*

- *Le premier volet concerne la circulation des données et du savoir. Il comprend des mesures sur l'ouverture des données publiques, la création d'un service public de la donnée. Il introduit la notion de données d'intérêt général, pour optimiser l'utilisation des données aux fins d'intérêt général. Une partie est également dédiée au développement de l'économie du savoir, avec la possibilité pour les chercheurs de publier librement leurs articles scientifiques dans un délai de six à douze mois. Le Sénat a voté en faveur de la facilitation de l'ouverture et de la réutilisation des données des administrations ainsi que des décisions des juridictions administratives et judiciaires. La diffusion de ces données sera circonscrite aux données dont la publication présente un intérêt économique, social, sanitaire ou environnemental.*
- *Le deuxième volet traite de la protection des citoyens dans la société numérique. Il affirme le principe de neutralité des réseaux et de portabilité des données. Il établit un principe de loyauté des plateformes de services numériques. Le consommateur dispose en toutes circonstances d'un droit de récupération de ses données. Le texte introduit également de nouveaux droits pour les individus en matière de données personnelles (droit à l'oubli numérique pour les mineurs, testament numérique pour donner des directives aux plateformes numériques, confidentialité des correspondances privées). Un amendement adopté par l'Assemblée nationale prévoit une peine de deux ans d'emprisonnement et une amende pouvant aller jusqu'à 60 000 euros pour le fait de transmettre ou de diffuser sans le consentement exprès de la personne l'image ou la voix de celle-ci, prise dans un lieu public ou privé, dès lors qu'elle présente un caractère sexuel (phénomène dit revanche pornographique " revenge porn ").*
- *Le troisième volet est consacré à l'accès au numérique pour tous avec notamment la couverture mobile, l'accessibilité aux services numériques publics, l'accès des personnes handicapées aux services téléphoniques et aux sites internet. Il prévoit aussi le maintien de la connexion internet pour les personnes les plus démunies. Le Sénat a adopté en première lecture un amendement qui oblige les opérateurs de télécommunications à s'engager, via des conventions avec les collectivités, pour l'installation du très haut débit.*

Un amendement de l'Assemblée nationale prévoit la remise au Parlement par le gouvernement d'un rapport sur la possibilité de créer un Commissariat à la souveraineté numérique et sur les conditions de mise en place d'un système d'exploitation souverain et de protocoles de chiffrement des données.

Un amendement adopté à l'Assemblée nationale prévoit que les propriétaires ou locataires qui louent leur logement de façon ponctuelle via des sites comme Airbnb, devront fournir à ces services la preuve qu'ils en ont l'autorisation, pour empêcher les sous-locations illégales. Les sites qui loueraient des logements sans l'autorisation adéquate pourront être sanctionnés. De même, les plateformes ayant pour objet des prestations de services proposées par des professions réglementées devront recevoir un avis conforme de l'institution chargée de l'application des règles déontologiques de ladite profession. Concernant les plateformes collaboratives, le Sénat a ajouté, pour les locations de logement, l'obligation de vérifier que les utilisateurs ne louent pas leur résidence principale plus de 120 jours par an.

Le Sénat a adopté un dispositif favorable au développement du jeu vidéo en ligne et créé un contrat de travail spécifique pour les joueurs professionnels de jeu vidéo. Enfin, il a réservé le bénéfice de l'exception au droit d'auteur pour liberté de panorama (qui permet de reproduire ou de diffuser l'image d'une œuvre protégée se trouvant dans l'espace public), aux seules personnes physiques à l'exclusion de tout usage à caractère directement ou indirectement commercial.

Pour mémoire

À ce point de la description de cet univers plus ou moins étrange pour nombre d'entre nous, il convient, d'une part, de résumer les chapitres précédents et, d'autre part, d'ajouter un éclairage technique afin que certains termes ne vous soient pas totalement étrangers !

Big data

Comme nous l'avons vu la définition initiale, donnée par le cabinet McKinsey and Company en 2011, présentait la célèbre règle des 3V : un grand Volume de données, une importante Variété de ces mêmes données et une Vitesse de traitement s'apparentant parfois à du temps réel. Le Big data correspondait à l'explosion des données dans le paysage numérique (le « data deluge »). Cette description a été complétée avec une vision davantage économique portée par le 4ème V de la définition, celui de Valeur, et une notion qualitative véhiculée par le 5ème V, celui de Véracité des données (disposer de données fiables pour le traitement).

Depuis 2011, les technologies capables de traiter en un temps limité de grands volumes de données évoluent sans cesse.

Open data

L'Open Data est un mouvement qui préconise une libre disponibilité pour tous et chacun, sans restriction de copyright, brevets ou d'autres mécanismes de contrôle de données jusque-là non disponibles. Ces données ouvertes et non sensibles, en anglais open data, deviennent alors des informations publiques brutes utilisables par tous.

On appelle données non sensibles celles qui ne portent pas atteinte à la vie privée ou à la sécurité de l'État et celles non soumises au droit d'auteur.

La libéralisation des données est encadrée par des licences qui fixent les conditions de leur diffusion et de leur réutilisation.

Quelques fondamentaux

1) Explosion du volume des données disponibles ou accessibles :

- structurées (données chiffrées, transactionnelles, fichiers classiques, Ascii, Excel, Access...) ;
- semi structurées (provenant de capteurs, de logs...) ;
- non structurées (texte libre, image, vidéo).

2) Mutation des outils de stockage et de traitement, sans limite de capacité et orientés temps réel.

Les outils de base

Hadoop **:** emblème par excellence du Big Data, Hadoop est une architecture spécifique de bases de données, permettant de traiter en grand nombre tous types de données (y compris les données non structurées). Hadoop autorisant, grâce à son architecture distribuée en clusters (HDFS pour Hadoop Distributed File System) le stockage de très gros volumes, permet à des applications de travailler sur des pétaoctets de données.

MapReduce **:** couplé à Hadoop, MapReduce est le mode de calcul permettant de traiter les big data. Il présente une fonction Map (distribution des données sur plusieurs clusters parallèles où les calculs intermédiaires seront effectués) et une fonction Reduce (les résultats des calculs intermédiaires distribués sont re-centralisés en vue du calcul final). MapReduce est issu de la recherche Google.

YARN : depuis 2013, Hadoop, initialement orienté batch, s'est équipé de YARN, solution qui lui permet, en plus du traitement massif de données, de faire du streaming et du temps réel.

Spark : s'intègre facilement dans l'écosystème Hadoop, avec lequel il est entièrement compatible. Il fait appel non pas au MapReduce sur disques, mais à de l'in-memory. Il autorise des temps d'exécution beaucoup plus courts (jusqu'à 100 fois).

Les outils d'analyse

- des outils dédiés big data chez les géants IT (IBM, SAP, ORACLE...) ;
- des logiciels statistiques usuels comme SAS ou R ;
- des développements de solutions « sur mesure », solutions métier, proposées par des SSII de toutes tailles ;
- un foisonnement de modèles statistiques et d'algorithmes d'intelligence artificielle, de machine learning ;
- des outils de visualisation et de présentation des résultats ;
- des promesses toujours plus fortes « d'analytics » automatique ;
- mais malgré la puissance des outils disponibles et en constante évolution (logiciels libres), le data mining reste une étape capitale pour extraire les bonnes données et éviter que les machines, via le machine learning, se retrouvent à apprendre du bruit ;
- l'intervention humaine pour sélectionner les données en entrée reste nécessaire pour éviter que la machine ne se perde !

Historiquement, deux communautés scientifiques différentes se sont lancées à l'assaut du Big Data : d'un côté, des personnes faisant surtout de l'algorithmie, de l'autre, celles qui faisaient principalement de la statistique. Les premières utilisent essentiellement Python, langage open source qui couvre le traitement, la visualisation et l'apprentissage machine. Les secondes, utilisent souvent « R », logiciel libre également, orienté traitement des données et analyses statistiques. Aujourd'hui la porosité est de mise entre ces communautés, ce qui a nécessité de faire évoluer ces langages initiaux.

2. BIG DATA BROTHER ET LOGICIELS LIBRES

Stéphane Laurière

Pourquoi un tel succès du logiciel libre dans le domaine des big data ? Quels en sont les enjeux pour l'innovation et pour la démocratie ?

Une couverture du roman 1984

Dans un télescopage des dates et des symboles dont l'histoire a le secret, c'est l'année même de son avènement annoncé que Big Brother a vu naître un petit frère porteur de principes d'émancipation qui allaient contribuer à solidement armer la société contre la dystopie du contrôle : le logiciel libre.

Trente ans plus tard néanmoins, le monstre d'Orwell reste menaçant sous les traits réels ou imaginaires de PRISM, Matrix ou Cerclon. Et les big data, si elles donnent l'espoir d'avancées majeures, font aussi courir le risque de big dégâts. À ce titre, le logiciel libre, dans tous les domaines du numérique et en particulier celui des big data, reste un enjeu de société majeur.

Quelques (big) data du logiciel libre

En trente ans, le logiciel libre est passé du statut de curiosité à celui de pilier de l'économie du numérique. Loin d'être le « cancer » dénoncé par Steve Ballmer, directeur général de Microsoft de 2000 à 2014, « le libre », à l'instar d'un de ses succès les plus fameux – le système d'exploitation Linux, s'est avéré porteur de principes redoutablement efficaces sur le terrain de l'innovation, conduisant certaines des entreprises parmi les plus florissantes du monde à en devenir aujourd'hui les faire-valoir pour accélérer leur croissance.

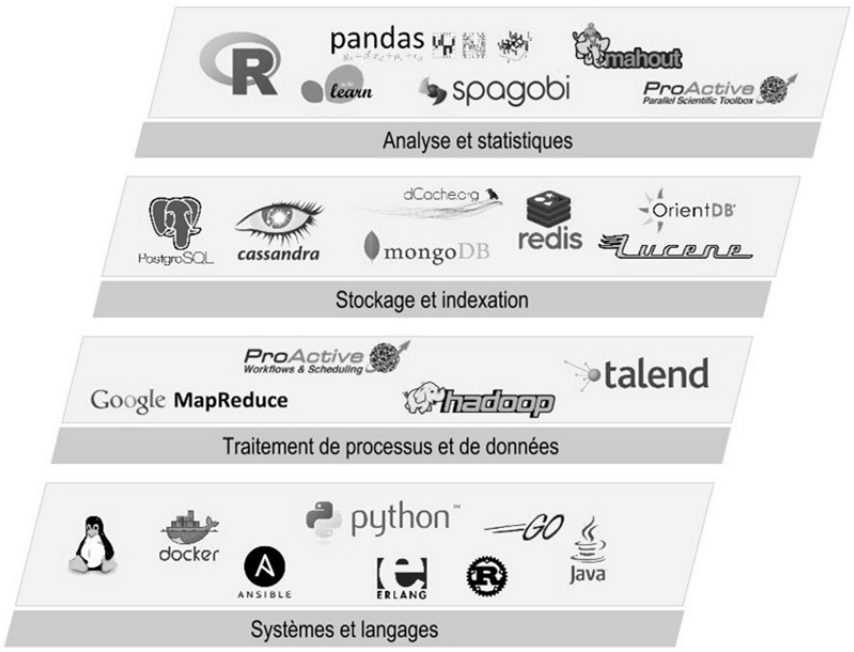
« Les entreprises rivalisent de plus en plus sur le front de l'ouverture maximale de leurs technologies » fait ainsi remarquer un journaliste de *Wired* au sujet de l'initiative Open Artificial Intelligence lancée par Elon Musk (Tesla Motors, SpaceX), et Sam Altman (Y Combinator) en 2015. Ainsi voit-on aujourd'hui par exemple Microsoft sponsoriser la conférence Debian, et Google publier sous licence libre son noyau de machine learning Tensor Flow parmi de nombreux autres projets dont Android qui équipe chaque jour plus d'un million de nouveaux appareils connectés.

Le secteur public n'est pas en reste en France comme le souligne Henri Verdier, à la tête de la DINSIC, ni aux États-Unis. Les principes du logiciel libre ont par ailleurs été précurseurs de l'open data, des licences Creative Commons dans le monde culturel, et de la dynamique de l'open hardware.

La prévalence du logiciel libre dans l'économie se traduit en 2016 par les faits suivants :

- en France, le marché du logiciel libre représentait 50 000 emplois en 2015, et croît deux à trois fois plus rapidement que celui des technologies de l'information dans leur ensemble selon une étude menée par le cabinet Pierre Audoin Consultants pour le Conseil National du Logiciel Libre ;
- le nombre de projets de logiciels libres actifs se compte en millions. GitHub, une des plateformes de social coding les plus actives, recense plusieurs millions de projets libres actifs à elle-seule ;
- une étude SAP Research estimait en 2008 que le volume de code source et de projets libres doublait tous les 14 mois ;
- les investissements en capital-risque dans le logiciel libre aux États-Unis se sont élevés à 1,3 milliard d'euros en 2014, à un rythme de croissance annuelle d'environ 30 % selon le cabinet spécialisé Black Duck Software ;
- 98 % des 500 ordinateurs les plus puissants du monde tournent sous Linux.

Quelques logiciels libres des big data



Parmi les nombreuses branches de l'informatique, les big data, en plus de soulever d'épineuses questions éthiques, ont la particularité de nécessiter des innovations dans toutes les strates du numérique, depuis les infrastructures matérielles et système jusqu'aux interfaces graphiques en passant par les langages et les outils de stockage et d'analyse.

La complexité des logiciels du domaine nécessite, pour chacun d'eux, la mobilisation de communautés importantes de développeurs, de chercheurs, de fournisseurs de services, capables de continuellement faire évoluer ces systèmes dans un environnement de compétition darwinienne. Cela suppose, en premier lieu, de leur assurer une viabilité économique.

Un logiciel est aujourd'hui moins un produit qu'un processus continu de production, de maintenance et de mise à jour. C'est un processus social (du point de vue de l'organisation des communautés et de leur gouvernance), scientifique, technique et économique. Il se trouve que le « modèle du libre » s'est progressivement affirmé comme bien adapté aux besoins d'amélioration continue exigés par le marché. Ainsi des écosystèmes mêlant laboratoires de recherche, grands groupes, sociétés de services, administrations, programmeurs indépendants se sont-ils structurés autour des outils phares des big data.

Pour faciliter la mise en place de tels écosystèmes et la coopération entre leurs acteurs, des regroupements de projets logiciels se sont opérés ces 20 dernières années au sein d'organismes à but non lucratif dont les plus célèbres sont les fondations Apache, Linux, Eclipse, toutes trois de droit américain. Et, en Europe, on a l'association OW2, créée en 2007 à l'initiative de l'Inria, Bull et Orange. Chacun de ces organismes anime et promeut des projets dans la plupart des domaines clefs de l'IT moderne, et en particulier celui des big data. La Fondation Apache héberge de nombreux projets big data dont Hadoop devenu en quelques années une plateforme de référence du domaine. Au-dessus de cette plateforme se sont développés de très nombreuses solutions métiers et des services dont récemment Warp10 dans le domaine des objets connectés, sous licence libre, à l'initiative de la société française Cityzen Data. La fondation OW2 regroupe elle aussi des logiciels big data d'envergure internationale tels que Talend, ProActive et SpagoBI (*NB: l'auteur de cet article est le directeur technique d'OW2*).

Plus récemment créée, l'organisation américaine à but non lucratif Bayes Impact se propose quant à elle de concevoir des logiciels libres dans le secteur des big data pour contribuer à la résolution de problèmes de société, à commencer par celui du chômage. Son fondateur franco-chinois Paul Duan déclare dans une intervention en 2015 *« pour mes parents le levier c'était de défiler sur la place Tiananmen, pour moi ce sont les algorithmes des big data »*, soulignant l'importance qu'ont les big data et le logiciel libre dans le nouveau squelette en silicium de nos sociétés.

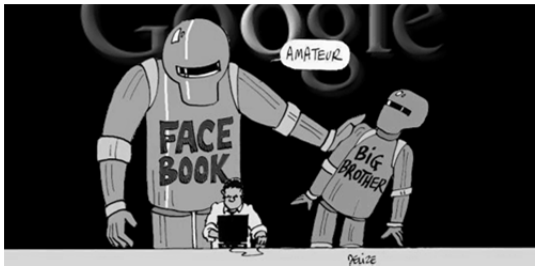
Les raisons d'un succès relatif

Le succès du logiciel libre dans le secteur des big data s'explique d'abord par le succès du libre dans l'ensemble des infrastructures d'Internet, succès qui découle d'abord de la genèse et de l'identité d'Internet lui-même. Comme le rappelle Fred Turner, professeur en sciences de la communication à l'université de Stanford, dans son livre *Aux sources de l'utopie numérique*, Internet est né de la rencontre de la contre-culture américaine et de la culture militaire. La majeure partie des infrastructures clefs d'Internet – protocoles réseaux, serveurs de noms (DNS), serveurs web, bases de données – repose sur du logiciel libre. L'influence de la contre-culture sur les grandes universités américaines qui forment les entrepreneurs de la Silicon Valley a perduré, comme l'indique Stéfane Fermigier, président du Groupe thématique Logiciels libres du pôle de compétitivité System@tic, auteur du rapport *Big Data et Open Source : une convergence inévitable ?* (2012).

Ensuite, le modèle d'innovation du libre a prouvé sa pertinence économique. Comme le dit Simon Phipps, ex-directeur de l'Open Source Initiative, le libre permet en particulier *« d'innover sans avoir à demander la permission préalable* », ce qui donne lieu sur le modèle du *fork* et de la composition à des mécanismes de compétition que nombre d'économistes jugent plus efficaces que ceux reposant sur les brevets pour le progrès collectif. De multiples manières de tirer des profits du logiciel libre se sont développées. Elles font l'objet d'études de synthèse comme celle du Groupe thématique logiciels libres du pôle System@tic Paris Région.

Enfin, les big data étant un enjeu technique et économique majeur pour les géants du Web que sont les *GAFAM* (Google, Amazon, Facebook, Apple, Microsoft), ceux-ci ont investi massivement dans la R&D des big data ce qui a donné lieu à de nouveaux algorithmes et méthodes de traitement, pour partie sous licence libre.

L'ouverture ne suffit pas



Why Privacy is Under Attack *?*

Pourquoi promouvoir, et promouvoir davantage le logiciel libre dans le domaine des big data ? Parce qu'il est un moyen puissant de garantir les libertés individuelles. Dans un monde qui est contrôlé et régulé de plus en plus par du code informatique, le logiciel libre donne au citoyen la garantie de pouvoir comprendre comment les infrastructures de la cité fonctionnent, de pouvoir en analyser les règles, et de participer à leur évolution. Imagine-t-on une société démocratique dont le code civil ne serait accessible qu'à un cercle de privilégiés ?

Néanmoins, comme le font remarquer les chercheurs Daniel Le Métayer et Antoinette Rouvroy, l'ouverture ne suffit pas. Les algorithmes et les programmes des big data posent à la démocratie l'exigence d'une éducation à l'informatique, et celle du débat. Comme le faisait remarquer récemment Pierre Rosanvallon appelant à *« un nouvel âge de l'émancipation »*, *« Internet donne à l'opinion publique une forme matérielle »*. Le code est ce par quoi cette matière s'érigera soit en nouveaux biens communs, soit en armes de contrôle et de manipulation s'appuyant sur les tendances générales détectables dans les big data de l'opinion à l'insu des individus. L'exigence de l'analyse critique de la finalité des algorithmes et celle de l'ouverture du code sont à cet égard une manière de peut-être protéger les hommes de devenir, *« aussi bien oppresseurs qu'opprimés, le simple jouet des instruments de domination qu'ils ont fabriqués eux-mêmes »*, selon la formule de Simone Weil.

Références

- Big Data et Open Source : une convergence inévitable ? – Stéfane Fermigier, 2012
- 15 ans de politiques publiques du logiciel libre en France – Stéfane Fermigier, 2014
- Algorithmes et responsabilités – Hubert Guillaud, InternetActu, 2016
- Réflexions sur les causes de la liberté et de l'oppression sociale – Simone Weil, 1934
- Big Data Is Watching You – Evgeny Morozov, NY Times, 2013
- La bataille du logiciel libre – Thierry Noisette, 2004
- Totalement inhumaine – Jean-Michel Truong, 2001

3. BIG DATA, TRANSPORTS ET SMART CITY

Alain Tripier, en remplacement de *Jacques Bonnot*, absent excusé !

Je vais vous faire une petite présentation de la mobilité dans le cadre de ce qu'on appelle la smart city ou la ville intelligente. La mobilité est un des points qui intéresse beaucoup les collectivités territoriales pour des raisons sur lesquelles il est inutile de s'appesantir.

Qui sommes-nous ? Des usagers, des clients, des citoyens, des contributeurs aussi, parce qu'à chaque fois qu'on apporte quelque chose, on laisse une trace et cela va contribuer à alimenter des data. Donc, si on parle de la mobilité pour les collectivités territoriales dites intelligentes, en fait, il y a deux grands objectifs :

- premièrement, fournir une information multimodale, c'est-à-dire mettre en commun l'information de base, les problèmes de contacts entre les modes de transport, les horaires, les infos trafic. En fait, ce sont des bases d'open data qui sont proposées par la SNCF, la RATP et de nombreuses et diverses compagnies. L'idée que beaucoup de collectivités territoriales ont envie de mettre en application, c'est d'avoir de l'info trafic et des données en temps réel. Quand il y a une panne, que les gens le sachent ; quand il y a des perturbations, que ce soit immédiatement annoncé et que cela passe par les smartphones et les tablettes ;
- second point plus sophistiqué, proposer aux usagers, aux clients, aux citoyens des applications qui leur permettent de choisir à tout moment le mode de transport le plus efficace et éventuellement de changer de mode en cours de déplacement. Ça, c'est plus compliqué. Regardez par exemple sur votre smartphone l'application RATP. En premier lieu si vous êtes dans le métro, très souvent vous n'avez pas de réseau ! Pour prendre le métro ou le bus, vous avez des tas d'informations RATP, mais on ne vous dit pas qu'à la sortie X de la station Y, vous avez aussi une station de Vélib ou d'Autolib.

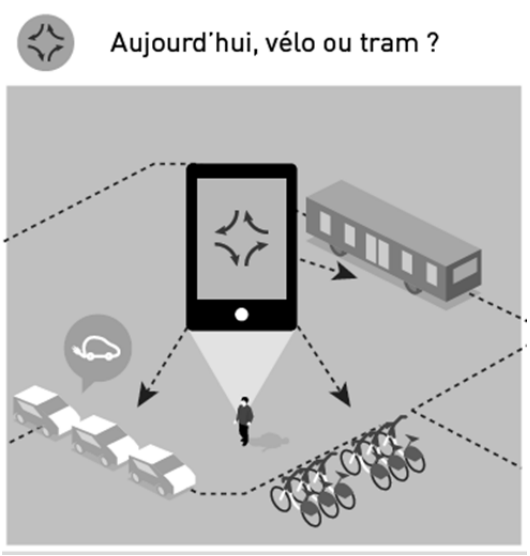
Donc, l'idée, c'est de fournir des applications qui soient complètes et qui soient transversales à tous les modes de transport. Un des meilleurs exemples en ce moment, c'est Copenhague. Ce n'est pas une très grande ville. Il y a 3 000 voitures électriques, un peu comme Autolib, qui sont à disposition. L'information sur tous les systèmes, que ce soit les vélos en libre-service, les voitures, le métro, le bus sont sur la même application. Si vous êtes, par exemple, dans une voiture électrique et que la circulation se congestionne, l'application vous dit « *tu peux poser ta voiture là, tu as une station de métro ici, donc tu as intérêt à prendre le métro* ». Cette approche est assez ubiquitaire à Copenhague puisque leurs applications smartcity prennent également en compte l'écologie (la pollution) ainsi que les questions de l'énergie.

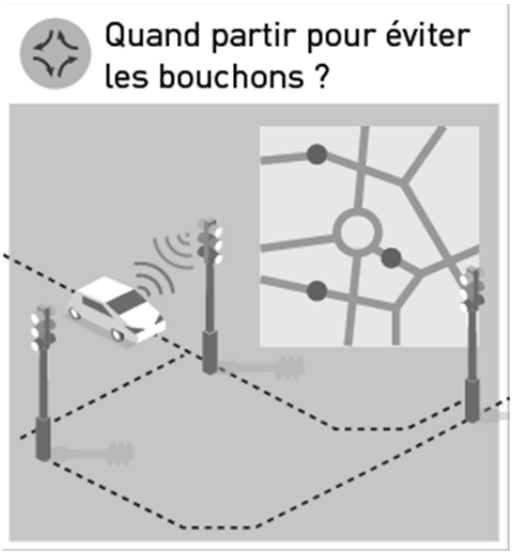
En France, il y a des exemples qui sont plus que des expérimentations, qui sont arrivés à maturité. Vous avez des choses très intéressantes qui se passent dans la communauté urbaine de Lyon, le grand Lyon ; à Nice, à Issy-les-Moulineaux (qui fait du big data depuis très longtemps), on n'en est plus aux expérimentations, on est déjà entré dans la complexité du sujet avec tout de même un peu de retard par rapport à des villes comme Copenhague.

Une autre application du big data transport, beaucoup moins proche de l'usager mais qui lui rend également un service en amont, est la régulation du trafic. À partir du moment où on dispose des data des opérateurs de transports, des opérateurs de télécoms et des réseaux de capteurs, on arrive à réguler le trafic avec des logiciels très sophistiqués. Par exemple, IBM s'est spécialisé dans ce domaine depuis un certain nombre d'années et régule la circulation de Londres et d'autres grandes villes du monde. Le système va faire fonctionner les feux rouges et la régulation du trafic plus en amont que ce qu'on a décrit avant.

Voilà en gros les objectifs en ce qui concerne l'information **à** la mobilité, l'information dans la mobilité et la régulation du trafic au niveau des villes.

L'infographie ci-dessous est empruntée au grand Lyon. Vous avez sur votre smartphone le choix entre différents moyens, les Vélov, les voitures électriques en autopartage, les bus ; vous avez également le choix de prendre éventuellement votre voiture pour vous déplacer. Tout ça à partir d'une application unique qui permet d'avoir un choix plus circonstancié, plus intéressant pour vous.





Dans le domaine de la mobilité, on a un déluge de données. Si vous prenez le périphérique parisien et tous les téléphones mobiles qui sont allumés dans les voitures, cela vous donne une idée parfaite de ce qui circule. On a cette récolte de données par les opérateurs télécom mais on a aussi évidemment toute la récolte de données par la billettique, c'est-à-dire par les systèmes de facturation des opérateurs. Autre point : les plateformes d'intermédiation comme Uber et Blablacar peuvent également jouer dans l'apport de data sur la mobilité, de même les applications interactives et collaboratives dont la plus performante en ce moment est Waze, rachetée par Google. Elles offrent sur vos smartphones, l'information en temps réel sur la circulation pour votre itinéraire (c'est réservé aux voitures) et parallèlement elles peuvent contribuer à nourrir les systèmes d'information.

Quand vous êtes dans un endroit embouteillé, vous envoyez l'info volontairement, ce n'est pas seulement votre smartphone qui le fait, c'est vous qui tapez quelque chose dessus pour dire « *faut pas venir circuler dans ce quartier-là* ». Évidemment, toutes ces traces numériques sont utilisées. Idéalement, l'ensemble des données devraient être à disposition des collectivités (voir plus bas la référence à l'article de B. Marzloff et B. de Fos). Cet aspect positif est souvent un vœu pieux car les opérateurs, téléphoniques notamment, n'ont pas forcément envie de donner leurs data, mais plutôt de les vendre !

Maintenant, je vais vous parler, en quelques minutes, du programme qui aurait dû être présenté par mon ami Jacques Bonnot.

C'est un gros programme qui a été mis en place par l'ADEME (Agence de l'environnement et de la maîtrise de l'énergie) qui a lancé beaucoup d'appels à projets ces dernières années sur les problématiques de mobilité. Ce programme-là est très complet, c'est une première ! Il est en cours d'exécution aujourd'hui avec trois objectifs :

- recueillir les données numériques à partir d'une application mise à disposition sur les smartphones des habitants ;
- agréger ces nouvelles données au fur et à mesure (là, on est bien dans le big data, ça rentre en data déluge !) avec ce qu'on connaît, notamment les enquêtes sur les déplacements des ménages qui sont récurrentes ;
- adapter au fur et à mesure de l'évolution du programme l'offre de déplacement et aider les collectivités territoriales avec un outil d'aide à la décision. L'intérêt de ce programme, c'est que, au lieu de faire quelque chose dont on aura le résultat dans dix ans, on fait quelque chose en avançant en même temps. Ça se fait dans deux endroits : sur le territoire de Reims-métropole et dans une autre communauté de communes pas très loin de Reims. Pourquoi ont-elles été choisies ? Vraisemblablement, on a choisi celles qui étaient volontaires et qui étaient déjà très avancées dans la réflexion big data, notamment Reims-métropole.

En fait, les bénéfices attendus sont de développer une stratégie de marketing territorial, c'est-à-dire que les usagers saisissent leurs informations et donc donnent en continu des data qui vont bien au-delà d'informations du type : « là, c'est congestionné, devant la cathédrale de Reims, faut pas y aller ». Ils nourrissent le système et ils en sont eux-mêmes acteurs. Pour les collectivités, s'impose la nécessité d'adapter les politiques de mobilité, au fur et à mesure, au fil de l'expérience. Voici une diapo reprise de leur programme :

À gauche de cette diapo, on a ce qui se passe pour le citoyen ; il bénéficie d'un service qui permet de mieux connaitre les ressources. Ça prend un peu la forme d'un réseau social, d'un micro-blogging qui est très contributif et ça marche très bien. L'expérience est en route maintenant depuis un an et montre que les citoyens rémois et ardennais jouent parfaitement le jeu. Il y a une interface de visualisation des déplacements. Que ce soit un conducteur ou un piéton, la personne qui se déplace trouve une information qui lui permet de changer de comportement, c'est-à-dire de changer de mode de transport. Ça, c'est du côté du citoyen. Et, évidemment, à droite de la diapo du côté de la collectivité, on anime le territoire en fonction de ces outils, à destination des différentes communautés d'acteurs qui sont intéressées et qui ne sont pas forcément que des usagers. Il y a de nombreuses communautés territoriales qui sont intéressées à mieux connaître les pratiques pour mieux agir et gérer les infrastructures.

Le cas du Transilien de la banlieue parisienne fournit un autre exemple d'application big data dans les transports, avec des nombres impressionnants : ce sont 6 200 trains, 3 200 000 voyageurs par jour, 180 rames connectées. Depuis 2009, chacune de ces rames produit 70 000 informations par mois, 40 000 variables qui sont transmises toutes les 30 minutes.

Jusqu'à présent, ces informations étaient quasiment exploitées à la main, enfin, pas vraiment à la main, mais il y avait des gens qui traitaient ces fichiers avec des plans statistiques habituels. C'était un peu lourd quand même ; le projet, qui est en cours de développement est d'automatiser ces analyses et d'avoir, en temps réel (on trouve ici encore cette notion de temps réel très importante dans le big data), une information précise sur l'état du matériel. Ce qui est intéressant, c'est que les différents acteurs sont orientés vers la maintenance préventive avec un système de machine learning. Il leur permet de croiser les données de fonctionnement qui arrivent toutes les 30 minutes avec leurs données d'exploitation. C'est une démarche correspondant bien à l'esprit du big data avec quelque chose qui est très en amont, très technique, par rapport à tout ce qu'on a pu raconter sur la vie des citoyens et des usagers.

Un article publié en Novembre 2016 par Bruno Marzloff dans *La Gazette des Communes* apporte un éclairage intéressant sur les évolutions vers la smart city et ouvre une réflexion sur les biens communs : http://www.lagazettedescommunes.com/469440/datacites-inventer-linteret-general-de-la-smart-city/.

4. DATA CENTERS EN SCÈNE

Une équipe pénombrienne se penche sur le sujet très chaud des data centers !

Par ordre d'entrée en scène : **Béatrice Beaufils, Françoise Dixmier, Michelle Folco, Marion Selz**

Béatrice : Eh les filles, lâchez vos écrans et regardez plutôt ça. C'est quoi, à votre avis ?

Françoise : C'est beau ! On dirait un tableau de Zao Wou-Ki !
Michelle : Un bout du nuage de Tchernobyl ?
Marion : Mais non, arrête, le nucléaire, c'était la nocturne d'il y a 3 ans. Ce soir, c'est big data.
Michelle : Alors, c'est sans doute le cloud qui stocke nos mails.
Béatrice : Non, perdu ! C'est ce qui sort des onze centrales à charbon de Caroline du Nord.
Michelle : Quel rapport avec les big data ?
Béatrice : Le rapport c'est que c'est l'énergie pas vraiment propre qui alimente en partie les installations de Google, Apple, Facebook, Amazon.
Marion : Mais Facebook et tout ça, c'est pas dans la Silicon Valley ?

Béatrice : Ah oui, leurs "green centers", leurs centres de recherche tout ce qu'il y a d'écolo. Mais pour le stockage et le traitement des données, les data centers, ils en ont plein en Caroline du Nord ; là-bas, l'énergie est très bon marché, mais pleine de charbon.
Françoise : Attends, moi je suis sur internet. Là, ils disent que ça, c'était avant, et que maintenant ils ont des grandes fermes de panneaux solaires et d'éoliennes. Même chez Apple, ils affichent « *100 % renouvelables »*.
Béatrice: Oui, c'est leur pub ! Et c'est vrai qu'ils font des efforts. Mais bon, le soleil et le vent, ça ne marche pas tout le temps et les data centers ça doit marcher 24h sur 24. Alors, le propre, ils le vendent, ça leur fait plein de crédits carbone, et puis ils en achètent encore plein et comme ça ils compensent, ils compensent. Mais ce qu'ils consomment, c'est l'énergie standard, plus noire que verte. Tout ça c'est légal, mais de là à dire qu'ils sont "*100 % renouvelables*".
Michelle : Tu veux dire qu'ils sont plutôt 100 % filous, donc. Remarque, s'ils se sentent obligés de mentir sur le renouvelable, c'est déjà un début, non ? Mais c'est quoi exactement un data center ?
Marion : Ben, c'est un peu du cloud, justement.
Michelle : Le cloud, c'est pas dans le ciel ? Mes données, elles ne se baladent pas dans le ciel ?
Françoise : Tiens, sur internet, je vois qu'elles se baladent plutôt dans des câbles, même des câbles sous-marins.

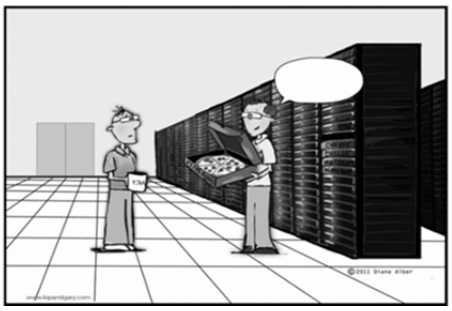
Et puis elles vont dans des tas de data centers, c'est tout ça ton cloud.

Béatrice : Tiens, je t'en montre un, data center. C'est là-dedans qu'on stocke nos mails, nos photos, toutes nos données de banque et d'assurance, mais aussi toutes les données volées par tous tes capteurs qui te sonnent tout le temps, les big data quoi !

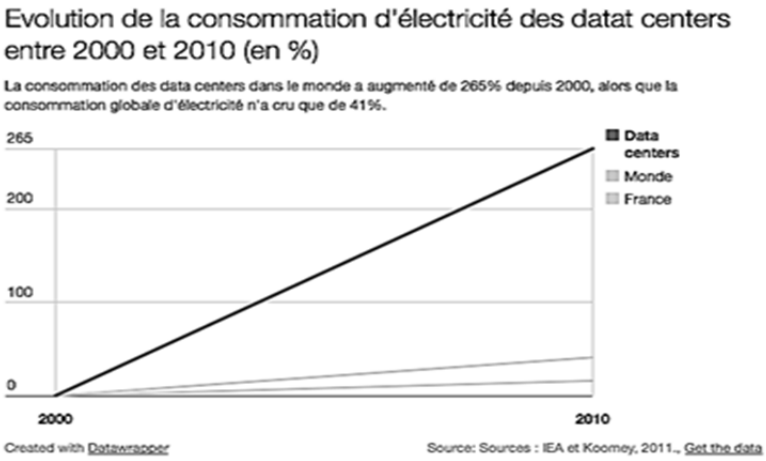
SONNERIE

Béatrice : Tiens, tu vois !
Françoise : C'est l'hôpital, faut que j'y aille dès ce soir, il paraît que mon bracelet connecté indique une forte dépression. Pourtant, je ne me sens pas si mal que ça.
Michelle : Bon, ben fais leur confiance, c'est pour ton bien ! Mais reste encore un peu avec nous.
Béatrice : Oui, reste ! Et regarde plutôt ça.

Béatrice : Il faut bien que ces big data soient physiquement stockées quelque part.
Michelle : C'est beau, c'est clean.
Béatrice : Oui mais pas tant que ça. Tu vois tous ces serveurs empilés, ils appellent ça des pizza-box, des boîtes à pizza. Il faut bien les alimenter et ils sont gloutons ; et puis, ça chauffe, ça chauffe alors faut refroidir et tout ça, ça consomme.



Marion : Et ça consomme combien ?
Béatrice : En Caroline du Nord, il paraît que ça représente 5 % de l'électricité consommée dans tout l'état.
Marion : Oui, mais c'est parce que ces data centers sont concentrés là. Si tu te places à l'échelle de la commune où ils sont implantés, tu dois même pouvoir dire qu'ils en consomment 95 %. Mais c'est à l'échelle mondiale qu'il faut voir, non ?
Béatrice : Tu as raison. On dit qu'ils pompent 2 % de l'électricité mondiale.
Françoise : Bon 2 %, ça va… Mais, remarque, quand même, il y en a de plus en plus des données, donc il va falloir qu'ils en fassent pleins d'autres des centres et ils vont pomper de plus en plus.
Marion : D'autant que pour la sécurité, ces données sont dupliquées dans plusieurs endroits, donc évidemment ça augmente encore les volumes stockés et aussi la consommation d'énergie.
Michelle : Tiens, justement, je vois un graphique qui parle de ça, regardez, ça fait peur !



« La consommation des data centers dans le monde a augmenté de 265 % depuis 2000, alors que la consommation globale d'électricité n'a crû que de 41 % » C'est débile, le truc démarrait à peine en 2000, alors cette comparaison n'a aucun sens ! Et pourtant, c'est l'AFP.
Marion : Tiens, tu pourrais leur écrire à l'AFP. C'est vrai, tu t'en prends toujours aux infographies du *Monde*, ça te changerait.
Béatrice : Bonne idée, mais même si ce graphique est bidon, c'est sûr que ça augmente quand même énormément.

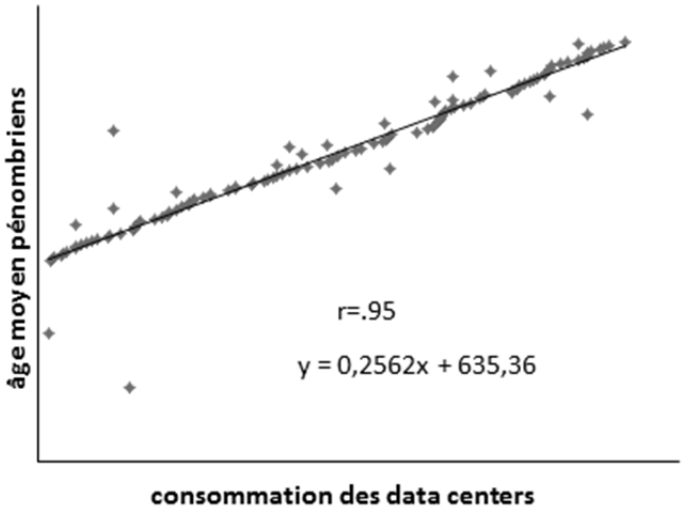
SONNERIE

Françoise : Tiens, une pub pour un antidépresseur.

Béatrice : Bref avec toutes nos données, on consomme, on consomme, et on réchauffe la planète.
Françoise : D'accord. Mais admettons qu'en traitant intelligemment toutes ces données, on améliore énormément le déplacement des voitures. Du coup, on arrive à réduire la consommation mondiale d'énergie.
Michelle : Et la chaleur des boîtes à pizzas ? On peut peut-être la récupérer, chauffer des maisons, des piscines, des tas de trucs comme ça, non ? J'ai entendu dire qu'on installait un data center à Paris dans le quinzième, et qu'avec ça ils chaufferaient l'immeuble d'à côté ; ce n'est pas bête.
Marion : Ben oui, tu as raison. Et certains commencent à avoir des idées révolutionnaires. Jeremy Rifkin dit qu'Internet, on peut aussi l'utiliser pour partager de l'énergie, comme à Montdidier, dans la Somme, par exemple. On reçoit un message qui conseille de recharger sa voiture au moment où la production d'électricité est à un pic…
Michelle : Et puis l'énergie, elle n'est peut-être pas partout aussi sale qu'en Caroline du Nord. Il paraît qu'il y a des data centers en Islande, là-bas l'énergie sort du sol en veux-tu en voilà. Et pour refroidir, hop, on ouvre la fenêtre.
Marion : Bon, on est peut-être un peu optimistes, là.

SONNERIE

Françoise : Encore !!! C'est mon assurance santé cette fois. Ils sont déjà au courant pour ma dépression, ils m'augmentent ma prime, zut alors.
Marion : Eh regardez. En moulinant dans mes big data, j'ai fait une découverte !

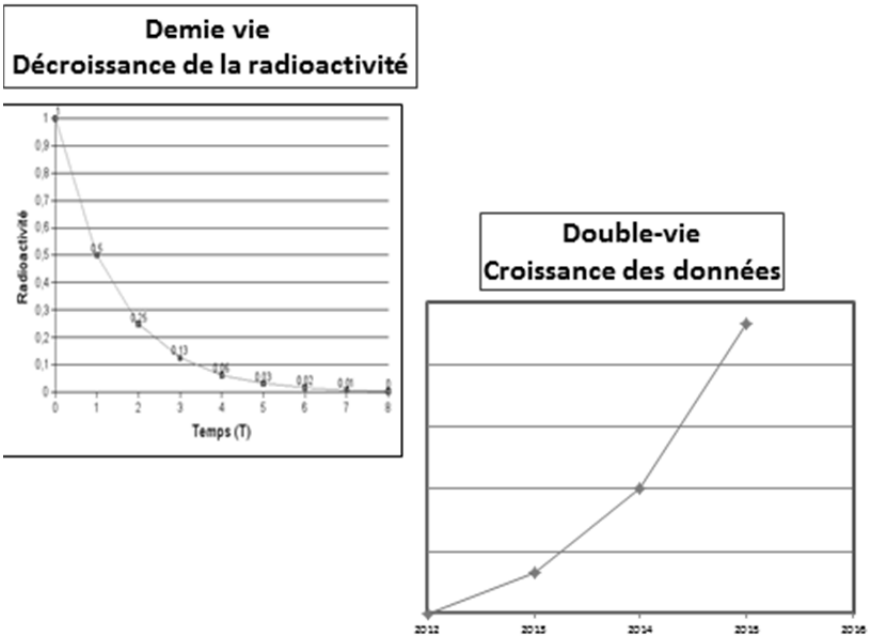


Plus la consommation des data centers augmente, plus l'âge moyen des adhérents à Pénombre augmente. Vivement que la consommation diminue, plus besoin de liftings, on rajeunit tous. On peut aussi faire le truc à l'envers. On recrute plein de jeunes adhérents à Pénombre, et comme ça, on fait diminuer la consommation des data centers… C'est vraiment génial ces big data.

SONNERIE

Françoise : Oh là, là. C'est mon employeur, cette fois. Ils me renvoient mes affaires, ils ne peuvent pas me garder avec ma dépression. Finalement, là je commence à plus me sentir bien du tout.
Michelle : Ma pauvre… mais tu vois, ils avaient raison à l'hôpital.
Françoise : Donc on réchauffe la planète, et j'ai plus de boulot. On ne pourrait pas rembobiner le film ?
Michelle : On pourrait peut-être commencer par se débarrasser un peu de certaines données. Il doit y avoir plein de trucs inutiles, non ? Comment on s'y prend ? Déjà que ce n'est pas facile de se débarrasser des données sur son propre ordinateur, alors on fait comment quand c'est parti dans le cloud ? Surtout si tu dis que c'est dupliqué ailleurs…

Françoise : Tiens, d'ailleurs, c'est combien de siècles, leur demi-vie ? Pire que les déchets nucléaires à Haute activité vie longue…



Béatrice : Ben pour l'instant, les données, on parle plutôt de leur double-vie que de leur demi-vie. Et en plus, ça change vite. En 2013, on disait que ça doublait tous les trois ans, en 2014 tous les deux ans, maintenant ce serait tous les dix-huit mois. Il y a peut-être une demi-vie de la double-vie ? En tout cas, pour le moment, on ne cherche pas du tout à s'en débarrasser, au contraire, ça se vend même très, très, cher, les données.
Michelle : Mais qu'est-ce que tu veux qu'on fasse avec les mails que je t'envoie ?
Béatrice : Ce qui a de la valeur, ce n'est pas seulement ce que tu racontes dans tes mails, ce sont tes métadonnées. Tu sais, l'endroit où tu es, à quelle heure, tes contacts…
Michelle : Mais, on n'a pas le droit de me piquer mes données comme ça.
Béatrice : Le droit, le droit… Quel droit ? Tu as bien vu que tes données se baladent partout et tu crois que le droit, il est le même partout ?
Marion : Ah non, c'est vrai, ce n'est pas du tout le même partout, pas pour le moment. En fait, en France, ça fait longtemps qu'on se préoccupe de la protection des données personnelles et qu'on a même contaminé l'Europe. Mais maintenant, on essaie de s'entendre avec les États-Unis, et là on n'est pas au bout de nos peines !
Françoise : Moi, je ne les trouve pas tellement protégées, mes données personnelles. Ça veut dire quoi, cette protection ?
Marion : Ben, en gros, ça veut dire que tu dois être informée à chaque fois que tes données sont recueillies et que tu es supposée donner ton accord pour qu'elles soient utilisées. Et puis aussi que tu dois pouvoir obtenir qu'elles soient détruites si tu le souhaites.
Michelle : Ah bon, et tu crois que ça se passe vraiment comme ça ?
Marion : Bof, non pas pour le moment en tout cas…
Françoise : Et pourquoi on a tant de peine avec les US ?
Marion : Ils sont moins attentifs que nous au respect de la vie privée. Mais comme il y a des intérêts économiques à ce que les données traversent l'Atlantique on a essayé quand même de s'entendre.
Béatrice : Si je comprends bien ce que tu dis, on n'y est pas encore !
Marion : Non, et dans bien des cas il y a un vide juridique… On est encore vraiment dans la fabrique du droit sur toutes ces questions, on va sûrement voir ça tout à l'heure…

SONNERIE

Françoise : Ah ben, finalement, il y a eu une erreur dans mes données, la dépression, c'était celle de Bretagne !
Béatrice : Non !!! Tu crois qu'on a mélangé tes données de santé avec celles de la météo marine ?
Michelle : Au moins tu vas pouvoir profiter tranquillement de la nocturne avec nous…

REDEVENONS SÉRIEUX…
DIALOGUE IMPROMPTU AVEC LA SALLE

Jean-René Brunetière

Nous allons enchaîner. Nous avons parmi nous un expert de Telecom Paris Tech. Antonio tu prends la parole quand tu veux sur nos affaires.

Antonio Casilli

Je reviens sur la question des transports et du big data car cela nous permet de faire un lien, une passerelle abstraite avec la question des grèves et des transports.

Sur les transports et le big data on pense tout de suite à UBER. C'est une manière aussi de mettre l'accent sur un aspect particulier qu'on espère traiter plus tard dans la soirée, le fait que ces données-là sont toujours liées au travail de quelqu'un. C'est à mon avis un aspect à souligner. On va le voir dans la rhétorique des big data.

On peut toujours imaginer la production des big data comme un processus alchimique dans lequel on arrive dans des conditions opaques plus ou moins magiques à la captation passive de données. Le fait est qu'il y a derrière un énorme travail de préparation, de nettoyage et finalement de mise en forme de ces données même non structurées. C'est le cas pour les données qui sont présentées de manière passive à partir de nos mouvements ou tout simplement à partir des dispositifs que nous gardons dans nos poches comme ce traceur-là qu'on appelle, faute de définition meilleure, un téléphone.

Toutes ces données produites sont les nôtres ; il y a des gestes qui sont les nôtres mais aussi ceux d'une énorme quantité de chaînes productives faites de personnes qui sont embauchées pour capter, ranger, nettoyer et préparer ces données. C'est important de le rappeler à ce stade-là pour ne pas à chaque fois présenter les big data comme le triomphe de la machine tandis qu'il s'agit du triomphe du travail humain. Dès qu'il y a un quelqu'un qui nous suggère le parcours le plus rapide pour aller de A à B en situation de grève, c'est quelqu'un qui a décidé pour nous que nous allons emprunter un certain chemin et pas un autre.

En gros – vous avez déjà entendu la blague selon laquelle le cloud n'existe pas et que c'est seulement sur l'ordinateur de quelqu'un d'autre que vos données sont stockées – je dis aussi que les big data et surtout l'algorithmique des big data n'existent pas. C'est la décision de quelqu'un d'autre qui nous impacte et en plus c'est une décision qui nous sollicite, qui est capable d'extraire un travail et une valeur produite par un travail. J'espère qu'on va y revenir. Merci.

Jean-René Brunetière
Est-ce qu'il y a des questions ou des observations ?

Question de la salle
Est-ce qu'il y a des outils qui permettent de savoir combien de personnes il y a dans les voitures ?

Chantal Cases
Le téléphone !

Jean-René Brunetière
Oui mais il y a des personnes sans téléphone, d'autres qui en ont deux ! On compte les téléphones mais on ne compte pas les gens. Il y a une période – moi je suis ingénieur des ponts – où on comptait les gens dans les voitures de visu en mettant un clampin sur le bord de la route. Je ne sais pas si ça se fait encore.

Intervention de la salle
Certains véhicules connectés sont en mesure de savoir s'il y a quelqu'un sur les sièges ou pas.

Jean-René Brunetière
Donc il y a des capteurs sous les fesses qui permettent de compter ?

Intervention de la salle
Il faut vérifier que les voitures sont assez récentes, pour l'instant je ne suis pas certain que ce soit extrêmement utilisé.

Jean-René Brunetière
Si les capteurs enregistrent un poids excessif on conclut qu'il y a quelqu'un sur les genoux !

Question de la salle
Il y a le bouclage des ceintures de sécurité ?

Jean-René Brunetière
La réponse à la question est un peu floue mais on a essayé d'y répondre !

Question de la salle
Une question toute simple sur le plan pratique. Est ce qu'il suffit d'éteindre son iPhone pour retomber dans l'anonymat, ou bien, est-ce que même éteint on continue à émettre des ondes et à fournir les endroits où on se trouve ?

Jean-René Brunetière
S'il est détruit je pense que ça s'arrête. S'il casse qu'est ce qui se passe ?

Antonio Casilli
Ça continue à émettre au moins certaines informations.

Jean-René Brunetière
Ça continue à émettre au moins certaines informations a dit Antonio. Tu peux peut-être préciser ?

Jean-René Brunetière
Il va avoir la parole tout à l'heure.

Geoffrey Delcroix
Ça dépend ce que vous entendez par éteint. Sauf pour des questions d'espionnage par des agences qui ont des moyens un peu particuliers, s'il est éteint, vous ne risquez pas grand-chose mais je ne serai pas catégorique non plus. Mais si par éteint vous voulez dire que vous n'êtes pas en train de l'utiliser, je peux vous garantir qu'il collecte énormément d'informations. On a fait quelques tests sur les téléphones pas spécifiquement éteints. Si ça vous intéresse on a participé à un projet avec Muriel, MOBILITICS, si vous tapez ça dans votre moteur de recherche préféré vous devriez voir quelques résultats. Par exemple les applications mobiles peuvent récupérer de manière extraordinairement intensive la localisation, etc. Tout dépend de la signification que vous donnez à téléphone éteint. En tout cas un téléphone éteint collecte beaucoup de données.

En termes de modalités pratiques dans certaines entreprises maintenant on vous demande d'ôter la batterie et dans certains pays maintenant c'est tombé dans les usages de laisser son téléphone à l'extérieur parce qu'il n'y a pas que les agences américaines qui sont capables de récupérer des données sur internet !

Jean-René Brunetière
Oter la batterie… mais il y a de plus en plus de téléphones où la batterie n'est plus amovible. On est sûr de toujours émettre. Antonio un dernier mot puis on passe à la santé.

Antonio Casilli
Ce n'est pas un dernier mot. Edward Snowden suggère de mettre les smartphones dans le frigo parce qu'il y aurait un effet de brouillage du signal, c'est l'effet des cages de Faraday.

Intervention de la salle
Dans le four à micro-ondes ça marche aussi.

Jean-René Brunetière
Merci, on passe à la santé.

5. BIG DATA ET SANTÉ


Chantal Cases et Guillaume Jeunot

Chantal Cases

Pour me présenter brièvement, je dirai que je suis tombée dans les big data il y a environ quinze ans. À l'époque, Pénombre avait travaillé sur le trou de la sécu, les plus « anciens » s'en souviennent.

Je dois aussi dire que je suis depuis dix jours présidente de l'Institut des données de santé, groupement d'intérêt public, chargé de faciliter et d'organiser l'accès aux grandes bases de données administratives en matière de santé à des fins de bien-être public et de gestion du risque. Mais, c'est juste moi qui parle ce soir, ils ne savent même pas que je suis là... (Rires)

Donc, de quelles données vais-je parler ? Je vais parler des données classiques, de la génération d'avant le big data dont vient de parler Alain. Ce n'est pas des péta-machins, mais plutôt des tera-trucs : c'est tout ce qui transite par votre carte vitale. Une ligne par soin, une ligne par médicament, tout ce qui est remboursable est dedans et ce qui ne l'est pas est aussi dans votre feuille de soins. Si vous vous faites rembourser des choses, vous êtes dedans.

1,2 milliards de feuilles de soins pour l'ensemble de la population vivant en France.

Je vais parler aussi des données des séjours hospitaliers (autre grand succès de Pénombre, la nocturne sur le PMSI, c'est comme ça que ça s'appelle. En l'occurrence, c'était plutôt le PMSI (psy)). Là on est plutôt dans les soins vécus. Un enregistrement à chaque fois que vous êtes hospitalisé.e ou que vous allez vous faire traiter à l'hôpital en ambulatoire. Là aussi, ça sert à plein de choses. Ça sert aussi à tarifer, et ça donne plein d'informations, encore plus que les autres données de la sécu parce que, en plus de savoir tout ce qu'on vous a fait pendant ce séjour, on sait pourquoi. On a aussi des diagnostics.

Ce qui m'intéresse moi c'est donc plutôt cela, ce qu'on appelle les données « médico-administratives ».

Au départ, il y a l'identité des gens ou le numéro de sécu. Cela permet d'apparier les fichiers entre eux et d'en faire des choses que certains peuvent considérer comme dangereuses, mais qui peuvent aussi être fichtrement utiles si on les utilise avec éthique. La CNIL reparlera de tout ça...

Pour quoi faire ? Le premier exemple auquel on pense, qui a beaucoup agité la blogosphère et les médias classiques ces dernières années, c'est la crise du Médiator, un antidiabétique qui a été beaucoup utilisé comme coupe-faim en dehors de ses indications. Il présentait des risques qui ont été maintes fois révélés par des lanceurs d'alerte (d'abord une lanceuse d'alerte, en l'occurrence). Il a été retiré du marché en 2009. Une des raisons de ce retrait du marché : les résultats d'études sur de très grands fichiers de données, ceux de l'assurance maladie. Même s'il n'y a pas beaucoup d'utilisateurs, on a retrouvé un nombre supérieur de personnes par rapport à ce qu'on attendait qui souffraient de problèmes de valvulopathie cardiaque. Je ne suis pas médecin, je n'en dirai donc pas plus, sauf que c'était très embêtant.

Voici des articles publiés par des personnes qui travaillent pour la CNAM, pour la sécu et pour l'INSERM et qui ont travaillé sur un million d'enregistrements de remboursements de l'assurance maladie.

Toutes les personnes sélectionnées étaient traitées pour diabète ; on les repère grâce à leurs médicaments ou grâce à des prises en charge pour affections de longue durée qui bénéficient d'un remboursement à 100 %. Parmi toutes celles-là, ils ont cherché celles qui, dans les années précédentes avaient été traitées par le Médiator. On a essayé de regarder combien d'entre elles avaient été hospitalisées pour les fameuses valvulopathies. Sur plus d'un million de personnes, ils en ont trouvé 43 000 traitées par le Médiator. Pour faire ça, il fallait avoir les données de remboursement des médicaments et les données d'hospitalisation afin d'apparier les deux et ainsi les exploiter. C'est de l'épidémiologie traditionnelle en faisant des comparaisons entre ceux qui étaient traités et ceux qui ne l'étaient pas et en calculant des probabilités.

Ça a été utile. C'est bien dommage qu'on l'ait fait si tard. Mais à vrai dire, on n'aurait pas pu le faire bien avant, parce que finalement ces données appariées de façon industrielle, on en dispose depuis très, très, peu de temps.

Les données hospitalières ont vraiment été mises en place à la fin des années 1990. Le SNIIRAM (fichier qui regroupe l'ensemble des remboursements de la sécu) a été de fait créé au milieu des années 2000. Les deux ont été appariés à la fin des années 2000. En fait, 2009 c'est le tout début de l'utilisation de ces données appariées. Il y avait d'autres expériences sur de tous petits échantillons et qui sans doute étaient moins convaincantes que celle-là. Voilà à quoi peuvent servir ces données.

Je voudrais ajouter que l'utilisation de ce genre de données n'est pas nouvelle. Dans ma vie passée, j'ai aussi été directrice d'un institut de recherche, qui s'appelle l'IRDES et où, depuis la fin des années 80, il y avait des statisticiens qui ont monté une enquête annuelle pour laquelle ils tiraient des échantillons dans les fichiers de la sécu. Ils enquêtaient des gens pour savoir comment ils se portaient et se comportaient : s'ils fumaient, s'ils avaient une complémentaire santé, etc. Et, après coup, ils ré-appariaient cette enquête avec les consommations de soins telles qu'on les connaît dans les fichiers de l'assurance maladie.

C'est un travail qui depuis plus de 20 ans a nourri toute la réflexion sur la protection sociale. Donc, le big data en santé, cela existe depuis longtemps. Ce qui est nouveau, c'est qu'on apparie beaucoup plus de fichiers et qu'on traite des données de plus en plus massives. L'enquête, elle, n'aurait pas permis de repérer le médiator : l'échantillon était trop petit !

Autre exemple : on tire un gros échantillon de diabétiques, on prend les données de l'assurance maladie, on ré-enquête les gens. On essaye de comprendre quelle est leur qualité de vie, quelle est la diversité des soins qu'ils reçoivent, si ces soins sont bien conformes aux recommandations que l'on donne aux médecins… On s'aperçoit que ce n'est pas toujours le cas, que les analyses ne sont pas faites dans les temps. Il y a parfois des interactions médicamenteuses pas terribles. On voit aussi avec les enquêtes que ça a l'air de s'améliorer au cours du temps. Ces constats peuvent avoir un impact positif sur les prescriptions des médecins… Autre élément, c'est à partir de ces fichiers qu'on a analysé les risques spécifiques liés aux pilules de 3ème génération, toujours avec la même méthode : on suit les femmes traitées avec ces pilules et on observe si elles ont subi des embolies pulmonaires ou des décès dans les années suivantes. Il y a aussi eu une étude célèbre sur les parcours de soin des personnes à qui on a enlevé la thyroïde ou un morceau de thyroïde. On s'est alors rendu compte que deux tiers d'entre elles, je crois, n'avaient pas eu les analyses préalables recommandées pour éviter éventuellement cette intervention… On a également, encore à l'IRDES, apparié ces données de santé avec les données de carrières collectées par la CNAV, ce qui permet d'analyser la santé selon le parcours professionnel ou les effets sur le parcours professionnel des arrêts maladie…

Tout ça fait progresser la connaissance et peut avoir des effets positifs sur la santé publique, et moi je trouve que c'est plutôt pas mal. Je ne suis pas la seule à trouver ça pas mal, et notamment les associations de patients et d'usagers se sont beaucoup mobilisées après l'affaire du Médiator pour dire que c'est inadmissible, ces choses-là on devrait les savoir depuis longtemps. Il y a plein de données et on ne les utilise pas, alors il faut qu'elles soient ouvertes.

Cette ouverture s'est faite aussi parce que les chercheurs se sont mobilisés : la récente loi santé, celle que vous connaissez à cause du tiers payant, a aussi, dans son article anciennement 47 et maintenant 193 organisé l'ouverture des données de santé.

C'est là qu'on passe du moyen big au presque open !

Pour préparer cela, le ministère a réuni une commission dite « open data en santé » à laquelle j'ai aussi participé, et qui a essayé de réfléchir sur ce qu'il fallait faire de ces données et comment il fallait les utiliser, avec quelles précautions.

On est arrivé à un certain nombre de conclusions suffisamment partagées pour que ça puisse s'écrire dans un rapport. En gros, ce rapport dit que quand il n'y a pas de danger pour les personnes et leur anonymat, il faut que les données publiques soient ouvertes. Quand il y a un risque, il faut les ouvrir aussi, mais en contrôlant ce qui se fait, en posant des conditions de sécurité et en autorisant des projets qui ont un sens, qui sont des projets d'intérêt général, pas liés aux intérêts commerciaux, des projets qui sont bien ficelés, qui utilisent les bonnes méthodes. C'est ce que fera le successeur de l'Institut des données de santé qui s'appellera l'Institut national des données de santé. On voit que c'est tout de suite beaucoup plus important ! J'espère que ça fera progresser la connaissance et l'information de chacun sur ce qui se passe dans le monde de la santé et des soins.

Je vais bientôt passer la parole à mon camarade Guillaume Jeunot, le Geek de service. Mais, avant j'ai un certain nombre de questions à lui poser. Je voudrais lui parler un peu de mes objets connectés à moi. Il y a quelque temps, mes enfants m'ont offert un petit bracelet qui se connectait à mon téléphone parce que je fais du footing le dimanche et ils trouvaient que c'était bien et qu'il fallait qu'ils me soutiennent dans mon effort. Donc, quand je rentrais du footing, je pouvais connecter le bracelet ; ça me disait combien de kilomètres j'avais parcourus, et j'avais un beau graphique sur ma fréquence cardiaque ! Ce qui m'a un peu embêtée, c'est que le bracelet disait que je courais dix kilomètres et que je courais avec une copine qui de temps en temps venait avec son fils qui était à vélo et dont le compteur disait plutôt 7,5...

Je me suis dit que peut-être c'est parce que je fais des trop petits pas ! Pour tout vous dire, je trottine...

Ce bracelet me disait aussi combien de temps j'avais dormi, la nuit, si c'était du sommeil profond ou pas. Je pouvais même partager tout ça avec mes copains et mes enfants !

Prudente, je ne l'ai pas fait, mais l'info n'était pas perdue pour tout le monde. De temps en temps, le fabricant de l'appli me disait : « *vous n'avez rien fait là* ». Forcément, je ne portais pas toujours le bracelet ! Donc je manquais de pas, ça c'est sûr ! Et puis, de temps en temps, il me disait : « *c'est bien, vous dormez plus que la moyenne des gens de votre âge* ».

Je me suis dit, c'est qui les gens de mon âge ? Ceux qui utilisent le bracelet ? Ou d'autres ? Tout cela était quand-même assez mystérieux !

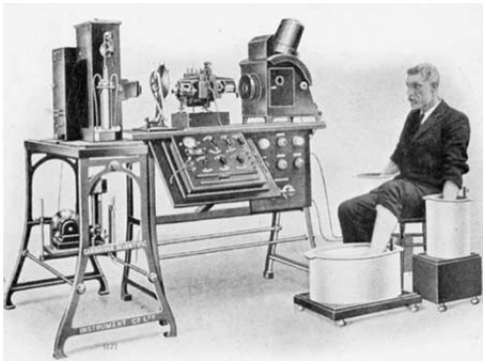
Il y a deuxième truc qui me tracasse un peu, sans doute parce que je travaille dans le milieu de la recherche, et que je voudrais partager avec Guillaume. Google inquiète beaucoup les chercheurs et les épidémiologistes, parce que Google s'est mis à dire : « *vos enquêtes, c'est rien du tout, nous, rien qu'avec les requêtes Google et les échanges sur les réseaux sociaux, on est capable de vous prédire les épidémies mieux que personne. On va pouvoir œuvrer pour la santé publique, prévenir le suicide, observer les épidémies de grippe etc.* ». Alors, ça va peut-être finir par marcher, mais pour l'instant, cela n'est pas terrible, si je lis les bonnes feuilles et les travaux un peu sérieux. Ça ne prédit pas la grippe et le suicide. Ça prédit l'écho médiatique de la grippe et du suicide, ça n'est pas tout à fait pareil, si j'ai bien compris les articles que j'ai lus. Mais, on ne sait jamais, comme le disait notre président en introduction - malheureusement non retranscrite mais peut-être partie dans le cloud - l'intelligence artificielle a battu le meilleur joueur de go, donc soyons prudents !

Guillaume Jeunot va nous expliquer tout ça.

Guillaume Jeunot

Ça risque de marcher, mais ça dépend de ce qu'on attend comme résultat et de qui est à l'origine de la production.

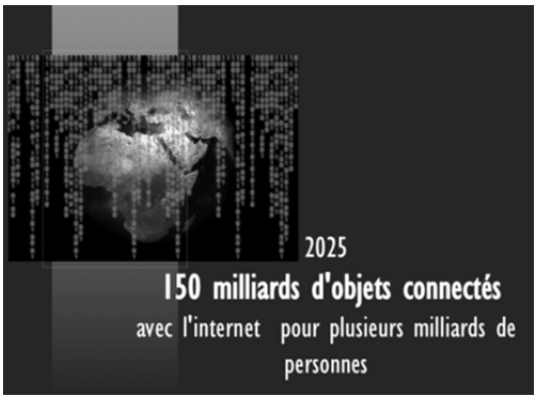
Voici l'ancêtre du bracelet connecté, qui présente Willem Einthoven, l'inventeur du Galvanomètre à cordes, premier instrument de détection de l'activité cardio-électrique au début du vingtième siècle.



Il faut reconnaître qu'il y a une sacrée évolution technologique entre le début de l'utilisation de cette invention faite pour capter le signal électrique humain et votre bracelet électronique qui va permettre de voir, de manière extrêmement subtile, vos paramètres.

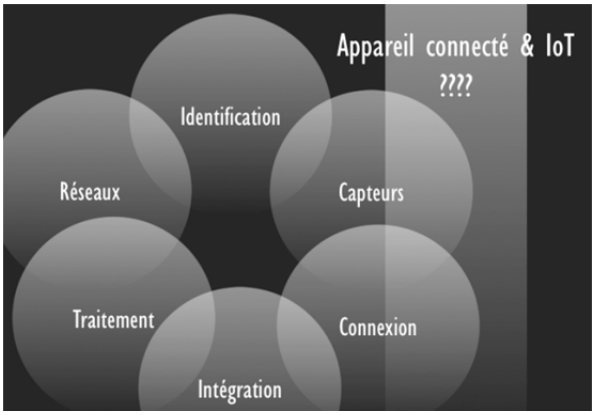
J'ai un profil d'ingénieur biomédical qui a surtout travaillé sur les problèmes d'étalonnage des dispositifs médicaux. C'est pourquoi je suis assez sensible à la fiabilité de certains types d'outils. Pour moi, la notion de santé pose beaucoup d'ambiguïtés (on y reviendra tout à l'heure) par rapport à ce qu'on peut connaître entre le monde médical et, par exemple, le bien-être. On ne peut mélanger les deux dans le même panier.

Un autre aspect de mon travail concerne la standardisation des échanges de données de santé. Ce qui me préoccupe en ce moment, c'est d'éviter les ambiguïtés entre la dépression médicale et météorologique. La confusion peut, on l'a vu, provoquer beaucoup de désastres. On en reparlera peut être avec la CNIL et avec la réglementation. Il y a un flou aujourd'hui autour de ces données...



Les cabinets d'audit estiment qu'en 2025 il y aura 150 milliards d'objets connectés. La plupart de ces objets sont capables de faire des mesures en temps réel (quelques millièmes de secondes, pour la plupart). On peut imaginer la quantité d'informations que ça peut échanger. On peut se poser la question : est-ce le volume de données qui importe ou l'information qu'elles portent ?

Se pose d'abord la question de définir l'objet connecté.



On parle souvent d'internet des objets, souvent les deux sont associés, on est sur un nouvel environnement. Je vais vous donner une définition qui me satisfait, qui s'appuie sur 6 composantes.

Un appareil connecté doit être identifiable. Si on n'est pas capable de préciser qui a fait la mesure, l'information qu'on pourra agréger dans un autre système n'aura pas de sens. Si on parle d'identification, aujourd'hui ce qu'on voit principalement sur les appareils ce sont des codes-barres, des puces RFID (Radio Frequency Identification) qui permettent une identification par utilisation d'une puce contenant une information sur un produit ou bien vous-même, reliée à une antenne miniature et chargée de communiquer cette information à un lecteur. Le passeport biométrique fonctionne sur ce modèle. Ce qu'il faut savoir c'est qu'il y a de nouvelles technologies qui avancent, qui se mettent en place. Certains labos de recherche commencent à utiliser l'ADN pour pouvoir identifier des dispositifs. On arrive sur les nanotechnologies donc la porte est grande ouverte à de nouvelles innovations. On peut s'en inquiéter mais ça peut avoir des impacts considérables sur la santé de nos concitoyens...

Le deuxième élément qui caractérise un appareil connecté c'est la notion de capteur. Je pense qu'à la maison, tout le monde a eu un thermomètre. Aujourd'hui il est même numérique, avec un affichage digital, il est connecté et on peut récupérer l'information sur son PC ou sur son smartphone. On a une multiplication des capteurs. Sur un smartphone, déjà vous avez un micro qui peut être utilisé pour autre chose que la conversation. On a des accéléromètres, des gyroscopes, qui permettent de l'utiliser comme une boussole. Au niveau des nanotechnologies, les capteurs de petite taille, dans l'industrie par exemple, peuvent être implantés dans des fraiseuses capables de mesurer la température, la pression, qui font que la machine numérique va pouvoir s'arrêter avant la rupture, la casse, la neige, etc.

La troisième composante importante c'est la connectique. Avant on avait tous la prise RJ45 sur un PC. Et puis on a été sur des technologies sans fil, Bluetooth, etc. Il y a tout un ensemble de connexions qui existent qui sont toutes plus ou moins adaptées par rapport à la distance et à la consommation. Les puces RFID dont je parlais tout à l'heure sont aussi « connectantes ». Elles sont par exemple implantées dans les câbles électriques afin de détecter les ruptures liées aux vols de cuivre et informer sur les réparations nécessaires.

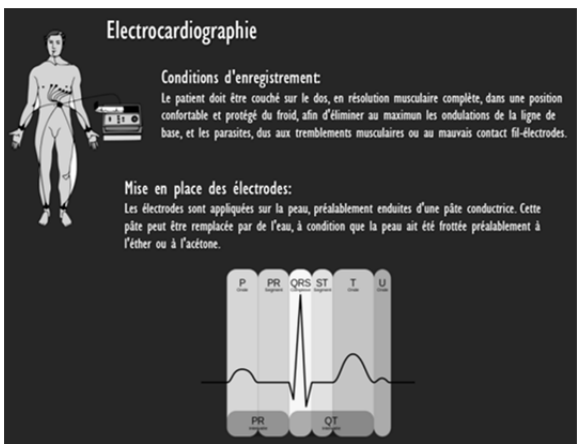
La quatrième composante concerne la notion d'intégration. Une information seule n’a en général pas de valeur si elle est isolée, donc il faut la mettre en regard d'autres informations et on a des systèmes qui nous permettent de concentrer ces données.

En cinquième vient la phase de traitement de l'information par les bases de données. On a les outils de Business intelligence et puis des entrepôts de données qui vont devenir de plus en plus intelligents. On dispose aussi des applications liées au web sémantique qui sont en train de se mettre en place.

La dernière composante c'est le réseau parce que pour que l'information se mette à circuler, pour qu'elle s'échange, il faut du réseau !

On n’en est qu'au début...

Plus spécifiquement, on m'a demandé de parler des objets connectés en santé. On parle des constantes des patients. L'utilisateur peut suivre la santé de sa famille et de ses proches. Dans le domaine de la santé, on a un certain nombre de pratiques qui sont issues d'un savoir-faire et d’avancées technologiques qui étaient liées à des usages bien particuliers. Dans la diapo suivante on retrouve le lien avec le bracelet connecté :



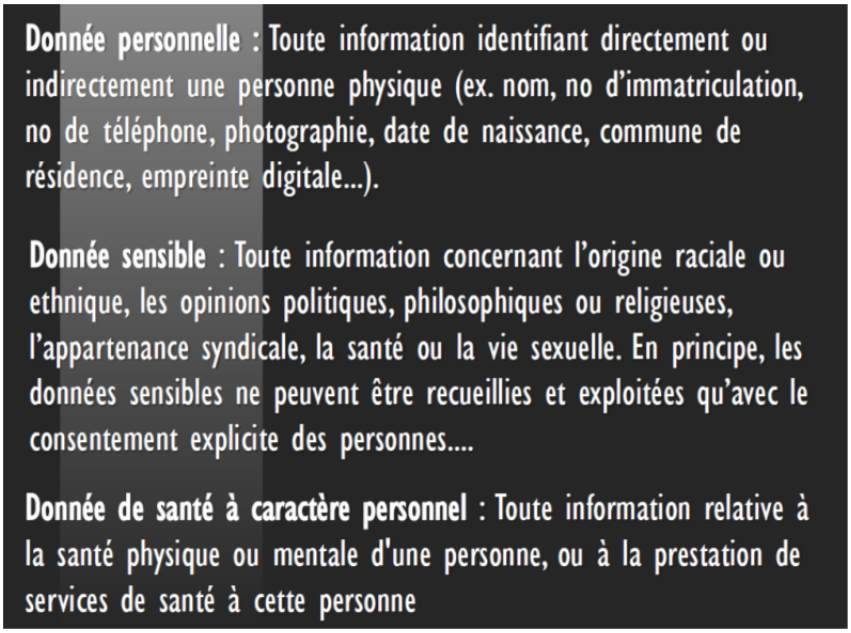
Un élément important concerne les conditions d'enregistrement, ça ne se fait pas sur un coin de table surtout s'il y des pieds métalliques, ça ne va pas se faire non plus après un footing quand on a bien transpiré. Les électrodes ne se placent pas n'importe comment, mais en fonction du type de signal qu'on veut mesurer.

Imaginons, Chantal, qu'on veuille utiliser ton bracelet pour t'informer pas seulement sur ton rythme cardiaque, on peut s'interroger sur la fiabilité de l'information qu'on en retirera.

Au niveau réglementaire quand on parle d'objet connecté on est dans le flou comme on l'a déjà souligné. En revanche, quand on parle de dispositif médical, avec les prothèses connectées, il y a eu un certain nombre de scandales qui ont conduit à améliorer la qualité.

On a des classifications des dispositifs qui existent mais qui ne prennent pas du tout en compte ces nouveautés.

On a un retard au niveau de la réglementation, j'espère qu'on n'aura pas à attendre un scandale suite à l'utilisation d'un appareil dont ce n'était pas l'usage pour que ça se mette en place. On retrouve ce besoin de classification, par exemple si on mesure la température de la pièce chez soi. Est-ce ou non une donnée personnelle ?



Pour la mesure du rythme cardiaque c'est bien sûr des données à caractère personnel et là on a des enjeux. Il y a des précautions à prendre, que tout le monde ne respecte pas. Or, si on prend des appareils connectés on est dans un contexte de mondialisation dont on a parlé tout à l'heure avec des data centers et on ne sait pas où vont ces données ! Les données peuvent tourner, par exemple, parce qu'on préfère que les traitements aient lieu la nuit pour des raisons de coût. Comme il fait toujours nuit quelque part, les données tournent avec la terre.

Du fait de la mondialisation, il est difficile de protéger ses données personnelles. Les Américains sont moins sensibles sur ce point. Nike avait posé des capteurs sur ses baskets et des sociétés d'assurance ont commencé à récupérer les infos, mais les protocoles n'étaient pas fiables et comme on attaquait le porte-monnaie, les Américains se sont dit « on va faire un texte pour protéger nos données et on va enlever les capteurs des baskets en attendant que le protocole soit plus sécurisé ».

Ceci m'amène à une réflexion sur la réglementation sur laquelle va reposer la sécurité de nos informations. Google qui collecte nos données est également quelqu'un qui fait de la publicité et, avant tout, c'est un modèle économique. Quand on voit que certains politiques commencent à se tourner vers ce type de fournisseur pour essayer d'imaginer les modèles de demain on peut avoir de réelles inquiétudes sur l'usage de nos données si elles ne sont pas correctement qualifiées.

Au niveau des travaux qu'on mène, on essaie de s'assurer de la signification des données, c'est-à-dire de qualifier correctement l'information. En gros, une information a toujours une finalité si elle a été recueillie. Si on l'utilise pour autre chose elle perd de sons sens. Ton capteur a du sens par rapport à ton activité sportive de particulier. Mais, si la science commence à la traiter, elle n'aura plus aucun sens. Le marketing donne des outils formidables pour nous faire oublier que ces instruments sont compliqués. Aujourd'hui quand on prend un appareil connecté les données peuvent être revendues à l'autre bout du monde.

Par rapport aux enjeux, il y a un site qui me plait bien : Shodan.

NDLR : Selon wikipédia, Shodan est un site web spécialisé dans la recherche d'objets connectés à Interne et ayant donc une adresse IP visible sur le réseau. Il permet ainsi de trouver une variété de serveurs web, de routeurs ainsi que de nombreux périphériques tels que des imprimantes ou des caméras. Shodan est également un outil utilisé par des chercheurs en sécurité et des pirates pour rechercher des dispositifs mal sécurisés et en prendre le contrôle à l'aide d'un seul navigateur Web.

Question de la salle
Est-ce qu'on sait combien il y a d'objet connectés actuellement ? On a prédit pour le futur mais maintenant ?

Quelqu'un dans la salle
80 milliards !

Alain Tripier
Là, on est dans la pénombritude totale. Vous avez des cabinets américains qui font sans arrêt des études diffusées sur le net gratuitement. Elles vous disent des tas de trucs qui changent tout le temps. En ce moment, les évaluations mondiales d'objets connectés varient entre 20 et 80 milliards.

Quelqu'un dans la salle à nouveau :
Je vous rappelle que nous sommes 6 milliards. Ça en ferait une petite dizaine chacun... Mais il y a aussi l'industrie et les capteurs pollution.

Jean-René Brunetière
Pénombre aime particulièrement les évaluations précises sur les dossiers !

Guillaume Jeunot
Actuellement on vit une transition en ce qui concerne les adresses IP : aujourd'hui on va passer à des IP6 (on est en IP4), ce qui va multiplier le nombre de connecteurs. Ce qui est sûr c'est qu'on n'en est qu'au début.

Jean-René Brunetière
Autre question ?

Alexandre Léchenet
Deux observations :

Sur les assurances, en France, Axa offre gratuitement à certains de ses usagers des montres connectées pour compter les pas. Pour l'instant c'est juste positif, si on a marché suffisamment on aura des petits cadeaux. Mais il y a derrière ça l'idée d'habituer les gens à avoir un mouchard sur le bras. Il y a aussi Allianz qui met des petits objets connectés dans les voitures de ses assurés pour observer leur conduite. Et pas du tout, pour l'instant, pour surveiller ce qu'ils font mais c'est simplement pour encourager les bonnes pratiques.

Sur un autre sujet, il y a un truc intéressant : toutes ces données déposées sur le cloud sont accessibles par la police. Il y a notamment des sites qui permettent d'envoyer son ADN pour savoir qui sont nos ancêtres ou quels risques on a par rapport à telle ou telle maladie. Ces sites, qui stockent une copie de notre ADN reçoivent des demandes d'accès de la police et donc même si on n'est pas dans le fichier on finit par y être parce qu'on a décidé de le laisser à une entreprise. On peut imaginer que ma montre soit invitée à témoigner contre moi pour démonter mon alibi.

6. PASNET.NET
UNE NOUVELLE INTERVIEW DU PROFESSEUR STATONE

Dans un sketch conçu et réalisé par **Fabrice Leturcq**, on découvre avec quelque inquiétude les applications du big data dans un domaine très sensible. Afin de restituer ce moment de bravoure Fabrice a transposé le film en bande dessinée.

Présentatrice : Professeur Statone, Pénombre connaissait vos travaux de criminologie consacrés aux interactions entre activités policière et criminelle, mais ignorait votre investissement dans une startup du secteur des NTIC : la société PASNET.NET®

Statone : C'est exact, je suis executive manager de cette société depuis maintenant 5 ans.

P : Et PASNET.NET® fait aussi dans le big data...

S : Évidemment ! Le traitement des données est aujourd'hui un gisement important de ressources pour les services chargés de la lutte contre le crime, certains outils aujourd'hui sur le marché prétendent même prédire le lieu, la date et la nature des crimes à venir...

P : Et ça marche ?

S : Bien sûr que non ! Mais ça rend beaucoup plus facile la prévision de l'activité policière.

P : ... Vous voulez dire criminelle ?

S : Non, non, policière ! PASNET.NET® propose à ses clients cambrioleurs, pickpockets, escrocs, trafiquants, une prédiction de l'activité policière basée sur leurs propres statistiques d'activité.

À partir de statistiques de faits constatés, les logiciels disponibles établissent ce qui est vendu aux policiers sous l'appellation « hot spot ».

P : Points chauds ?

S : Exactement, c'est à dire les lieux supposés des prochains crimes. Or ces faits, ce sont nos clients qui les ont commis, nous les connaissons donc mieux que personne.

Une petite surveillance des lieux de patrouille nous a rapidement permis de décrypter l'algorithme vendu à prix d'or au ministère de l'Intérieur.

P : Mais alors, vous êtes en mesure de prédire...

S : ... Dès la fin du weekend, les lieux vers lesquels les policiers et gendarmes seront orientés par la machine au cours de la semaine suivante...

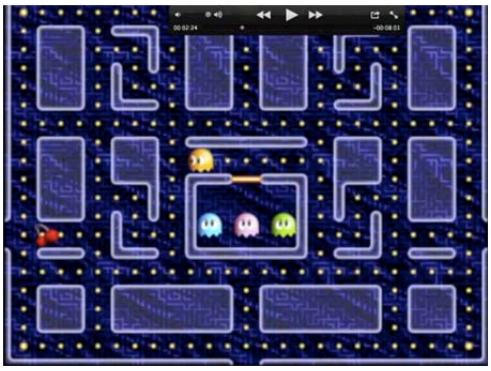
P : Mais les statistiques policières vont s'effondrer !

S : Mais non ! Un bon policier parvient toujours à interpeller un clandestin, un dealer de drogue, ou un motard sans casque. Et puis il y a les indépendants, ceux qui ne cotisent pas à notre organisation…

P : L'essentiel est que vos clients soient certains de ne pas avoir de surprise !

S : C'est ça !

S : Regardez ! Voici l'outil de positionnement des inspecteurs. Ils sont regroupés ici et attendent les événements prévus par le logiciel...

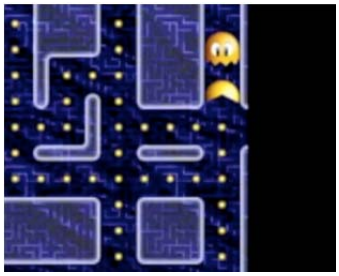


P : Et ça, c'est votre outil !

S : C'est ça ! Pendant que les pandores attendent bien au chaud, nos équipes circulent tranquillement vers leur prochaine cible et peuvent travailler en toute tranquillité. Nous les accompagnons en temps réel, d'un QG...

... à partir duquel nous les avertissons d'un éventuel mouvement des policiers...

Nous les aidons ensuite dans leur fuite grâce au meilleur détecteur de bouchons disponible sur le marché.



P : C'est extraordinairement moderne !

S : Oui n'est-ce pas?

S : Mais ça n'est pas tout, nous sommes aujourd'hui en mesure de fournir des services à tous les moments de l'activité délinquante... Nous attendons beaucoup de l'open data...

P : C'est à dire l'ouverture, au public, des données de l'administration !

S : C'est cela... L'un des services les plus populaires de PASNET.NET®. Le Gogo® est né de la diffusion il y a quelque temps des données carroyées de l'INSEE, vous en avez entendu parler ?

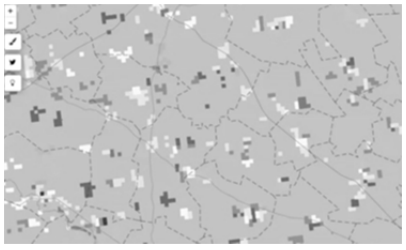
P: Oui, bien sûr, le découpage en carrés de 200 mètres de côté du territoire national, permettant une description extrêmement fine des caractéristiques socio-économiques des territoires.

S : Oui, eh bien c'est une mine pour qui cherche à préparer un mauvais coup !

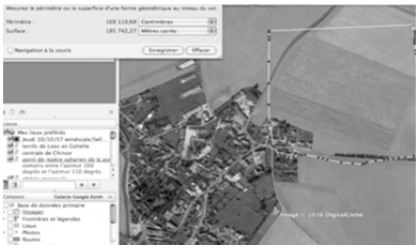
Regardez cette démo préparée par Gégé-les data, mon plus ancien associé**.**

Devant se rendre au baptême de sa nièce du côté de Chartres, il prépare son trajet, comme à son habitude, en visionnant notre carte carroyée...

Il recherche d'abord les carrés bleus. Plus c'est foncé, plus le revenu moyen par unité de conso est élevé, En voilà un !



Zoom! un petit repérage des lieux, un saut sur Google Earth lui permet de localiser la rue de la Prairie



S : Un petit tour sur cadastre.gouv.fr lui apprend que plusieurs belles parcelles appartiennent à la même adresse....

Enfin, Google Street View lui permet de confirmer l'intérêt du coin et de choisir ses cibles sans risque d'erreur et de repérer les points faibles...

P : ... et même, coup de bol, la qualité des serrures !

S : Le beauceron est âpre au gain, et il le sait !

Il a lu Zola mon Gégé !

P : Évidemment, à la campagne, ça marche, mais en ville ?

S : En effet, la Beauce se prête bien à l'exercice, mais l'avantage c'est la mobilité. Nos clients urbains trouvent avantage à repérer sans risque des lieux où ils ne font que passer et demeurent absolument inconnus. Mais pour l'espace urbain, nous développons actuellement un projet de crowd sourcing qui nous permettra de cuber les données INSEE afin d'offrir à nos clients une vision étage par étage des revenus...

P : Décidément, big data, open data et maintenant crowd sourcing, vous êtes à la pointe !

S : Oui, vous avez vu nos services premium, mais à la base, PASNET.NET® est d'abord un service de chacun pour tous

P : Une sorte d'Uber pop ?

S : Oui, Uber et ... l'argent d'Uber...comme dirait Gégé !

S : Grâce à nos clients emprisonnés, nous avons constitué mazone.pasnet® une base de données gigantesque, véritable mine pour le chercheur en peinologie que je suis ! Nous possédons en effet dans le détail toutes les peines prononcées par tous les tribunaux et cours depuis plus de cinq ans. Les habitudes et sensibilités de tous les magistrats, siège et parquet, nous sont parfaitement connues, ainsi que les résultats des avocats pénalistes de tous les barreaux de France ! Quand on sait que la justice, c'est ce que mangent les juges au petit déjeuner vous imaginez les services que nous sommes en mesure de rendre à nos clients !

P : Et dites-moi, sur les objets connectés, vous avez quelque chose, sur les objets connectés ?

S : C'est l'avenir, bien sûr.

Nous avons deux petits projets déjà bien actifs actuellement...

S : Le premier, "JOINT-Venture®"...

c'est la cigarette électronique au Cannabis,

le produit est au point, il vous offre un service de livraison, dès que vos réserves sont épuisées par simple connexion à votre smartphone!

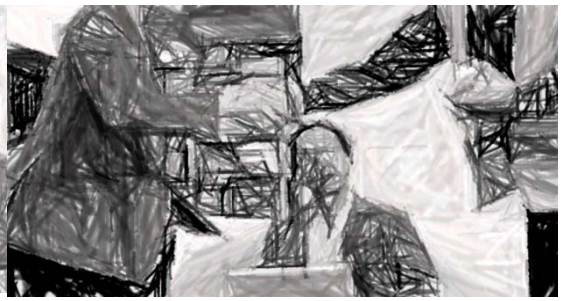
S : L'autre, c'est BIBIcourt® ...c'est Gégé qui s'en charge

Bruits de pas ...

S : Tiens d'ailleurs le voilà !

P : Bonjour Monsieur Gégé

Gégé : Bonjour Madame, qu'est-ce qui vous amène ?



P : Mais quel est donc BIBIcourt® ce projet dont vous avez la charge?

G : C'est très simple, je fais courir les smartphones,...

… les montres électroniques …et les lunettes Wi-Fi...

G : J'ai même un joint...électronique !

G : Vous comprenez les caves, aujourd'hui, quand y négocient un contrat avec leur beurrier ou leur assureur, y s'engagent à faire du sport, pour la santé, ça rassure l'assureur.

Et ces machins là, ça sait tout, ça voit tout et ça dit tout au banquier ! Alors pas question d'rater un footing !

…Seulement y z'ont pas toujours envie d'aller s'cailler les miches à tourner autour du pâté d'maison. Alors, pour pas cher, c'est BIBI qui s'y colle !

P : Et cet objet, là à votre cheville ?

G : Ça ? C'est mon boulet électronique mobile, je suis en liberté conditionnelle !

FIN

Une production Violette Ultra, visible en image animée sur :
www.dailymotion.com/video/x4rboj

7. CROWD SOURCING, IT COUNTS

Jean-René Brunetière

Un grand merci au professeur Statone qui apporte toujours des éléments scientifiques dans des débats qui pourraient être un peu plan-plan. C'est le domaine de la sécurité.

Les big data permettent aussi d'observer des faits sociaux, on va en avoir un exemple avec un élément important qui est la parité. Qu'est que les big data peuvent apporter à la parité ? Margaux Calon qui est de *Wax Science* va nous en parler. D'abord c'est quoi *Wax Science* ? Racontez-nous ça.

Margaux Calon

C'est une association qui fait la promotion des sciences, qui s'intéresse à la parité dans les sciences, pour les jeunes. Je m'appelle Margaux Calon. Je travaille à l'Institut des Systèmes Complexes et je suis surtout bénévole au *Wax Science* ça fait maintenant deux ans et demi. Comment ça se fait qu'on travaille depuis deux ans et demi sur la problématique de la parité ? On devient pour l'entourage le référent de la parité et des questions féministes, voir l'article du *Monde* sur la parité. En soirée des gens viennent nous voir « *j'ai pensé à toi, à la conférence, il y avait trois hommes* ». C'est plutôt bien, on voit qu'il y a des petites prises de conscience ponctuelles, mais cela reste individuel, ponctuel, ce n'est pas inscrit dans des initiatives.

La question que l'on a commencé à se poser : comment fédérer cette prise de conscience et ces petites initiatives individuelles pour faire quelque chose. Un autre problème, c'est que quand on va plus loin et qu'on s'intéresse à ces questions de parité, on va à des conférences, des workshops, des formations et on se rend compte qu'on est tous un petit peu d'accord, on est tous déjà informés depuis longtemps, on se met à parler des mêmes choses avec les mêmes gens. Comment fait-on pour sortir de cet entre-soi et sensibiliser les gens qui ne seraient pas allés vers ces sujets-là ? Ce qui nous a amenés à sensibiliser des gens qui ne sont pas du tout au courant, c'est en effet très difficile de parler de parité, très souvent on a des débats très enflammés – on l'a vu ces derniers temps au niveau politique –. Au CRI, Centre d'Études Pluridisciplinaires qui nous héberge, on a fait une campagne de posters qui ont été tagués de façon assez virulente par les élèves. Comment éteint-on le feu et comment discuter de tout cela d'une manière sereine ?

Alors on s'est dit, bien ce sont les chiffres – on ne dit pas cela parce qu'on est des scientifiques mais parce que c'est le seul moyen pour avoir des discussions objectives. Et ne serait-ce que pour savoir s'il y a un problème ou pas. On a eu des gens qui nous ont dit « *tiens c'est bien votre application (sur laquelle je vais revenir), je vais l'utiliser pour vous montrer que le problème de la parité c'est quelque chose du passé* », et nous ça nous va très bien, on a besoin de ces chiffres-là pour pouvoir commencer un débat serein. Ces chiffres parlons-en.

En 2012, je ne sais pas si vous savez que la Commission européenne a dit qu'il y avait 33 % de femmes dans la communauté scientifique, qu'il y a 21 % de garçons en terminale L, que 16 % des maires sont des femmes. Est-ce que vous savez combien d'universités portent le nom d'une femme ? 2 ? Non 0,5, Pierre et Marie Curie ! Tous ces chiffres c'est la preuve qu'on manque de visibilité. Alors, « pourquoi ce serait un problème ? ». Il n'y a pas de parité, il n'y pas de mixité, est ce que c'est vraiment un souci ? Oui c'est vraiment un souci ; on donne toujours deux exemples. Ce qui nous préoccupe le plus au niveau de notre association, c'est l'orientation professionnelle : la jeune fille qui grandit sans modèle, sans « rôle model », ne va pas forcément s'orienter vers l'univers scientifique, ce qui marche aussi pour les garçons. Et au-delà des stéréotypes il y a la question de la performance ; une étude a montré qu'une entreprise ou une organisation où il y a moins de diversité serait moins innovante. On a tout intérêt à aller vers plus de parité.

Ce sont les problématiques auxquelles on s'est intéressé et qui ont donné naissance à ItCounts [*NDLR : Afin que nul n'ignore voici l'adresse du site : http://itcounts-app.org/*], une application qui est censée y répondre. On veut faire quelque chose à l'échelle individuelle, c'est-à-dire qu'on n'est pas obligé d'aller à des conférences, c'est vraiment soi à son petit niveau avec une méthode innovante, le meilleur moyen d'apprendre à un grand nombre de personnes – de 17 à 77 ans !

ItCounts, c'est un projet de crowdsourcing, comme on en a parlé tout à l'heure, qui va compter des ratios hommes/femmes à n'importe quel moment, n'importe où. Il y a un site qui permet de collecter, de visualiser, d'analyser les données et de les partager. Nous attachons de l'importance au sentiment d'appartenance à une communauté. L'élément le plus important est la carte, ce qui vous permet de vous situer et de voir aussi les gens qui sont proches de votre problématique. L'application est gratuite, en open source téléchargeable sur vos smartphones où vous pouvez visiter le site internet pour compter et partager vos ratios hommes/femmes quand vous voulez et où que vous soyez. Il y a 6 étapes. Pour faire vite, vous téléchargez l'application, vous renseignez une catégorie selon que vous participez à une conférence, ou que vous êtes dans une classe. Êtes-vous en train de compter des enseignants ou des élèves ? Ensuite vous entrez les données comptées, vous entrez le ratio. C'est aussi simple que cela.

Jean-René Brunetière
On va faire l'expérience, on va faire l'essai en vraie grandeur.

Margaux Calon
Il y a six étapes dont trois principales, ça prend moins de deux minutes.

Jean-René Brunetière
Pour faire une démonstration, pas de 4G, il y a une barrière technologique, une faille du système, c'est clair !

Question de la salle
Les données sont-elles géolocalisées ?

Margaux Calon
Pourquoi on a choisi de géolocaliser ? C'est une question qu'on s'est posée, une question d'appartenance, à la base on voulait même les nommer clairement, « *là je fais une conférence à l'Unesco, j'ai tel pourcentage d'hommes et de femmes* ».

Antonio Casilli
C'est du déclaratif ?

Margaux Calon
C'est du déclaratif. Comment on implique les gens et comment on en fait une communauté, si c'est anonyme ? Tu peux être géolocalisé mais toutes les données sont anonymes ; elles ne sont pas du tout exploitées au niveau individuel, et c'est pourquoi on était plutôt à l'aise avec cette idée de géolocalisation qui a plutôt vocation à être pédagogique pour avoir une vue d'ensemble du phénomène et pour permettre des comparaisons. Ensuite quand on a les données, c'est très bien mais se pose la question : collecter des données pour quoi faire ? Notre premier intérêt, c'est de se sensibiliser et de sensibiliser les autres. Une fois qu'on a rentré les données qu'est-ce qu'on fait ?.On va les partager sur Facebook, ça parait anodin, ça attire l'attention surtout auprès des jeunes. Quand on a un gros volume de données on peut comparer, regarder ce qui se passe dans les quartiers d'à côté.

En biologie, il y a plusieurs sites qui nous permettent de comparer, « *j'ai 30 % de femmes comment je me positionne par rapport à la moyenne* ».

Ensuite il y a un deuxième point c'est se poser des questions et essayer d'y répondre soi-même. C'est de la science citoyenne, il y a beaucoup d'applications comme ça, qui se passe entre le capteur et l'émetteur, surtout que l'utilisateur est le récepteur, il faut qu'il puisse analyser ses propres données lui-même, qu'il ne se contente pas de les collecter et ensuite de les voir analyser par des chercheurs. On a les moyens d'analyser, de visualiser en générant des graphes, de comprendre ce qui se passe en changeant les filtres.

Le troisième niveau c'est de passer de la donnée au réel.

Ce qui nous intéresse c'est que les gens s'impliquent concrètement sur le terrain. Pour ça on a le site, avec une toolbox, on va leur donner des outils, des tutoriels, par exemple « La parité en 180 secondes » ce qui est plutôt concret.

Ça c'est pour les trois premiers niveaux, mais on n'a pas envie de s'arrêter là.

On a envie de travailler avec les entreprises et les centres de recherche, faire des campagnes de sensibilisation pendant une semaine ; ou en imaginant avec le CNRS, un institut, que tout le monde utilise l'application pendant une semaine ce qui nous permettrait de sortir des rapports plus rapidement qu'on ne le fait, et aussi d'agir plus rapidement avec les responsables. On pourrait faire des labels sur les conférences, c'est quelque chose qui nous plairait.

Le plus gros projet bien sûr, c'est d'utiliser cette application à but de recherche en se posant des questions beaucoup plus méta. Comment quantifie-t-on la prise de conscience individuelle ? Ça serait dans un second temps et pour ça on aura besoin d'être appuyés par des centres de recherche ; donc si vous voulez en parler avec nous, c'est avec grand plaisir. Ça vient juste de commencer, d'être lancé chez Google en février, parce qu'on a reçu le soutien logistique de Google. Et nous avons été soutenus financièrement par L'Oréal, on n'est plus en phase de construction puisque vous pouvez charger l'application mais on prépare une deuxième version. Si vous avez envie de travailler avec nous dans les entreprises, les centres de recherche. On a tout un toolkit, on est une association, c'est gratuit ! N'hésitez pas à vous adresser à nous.

REPRISE DU DIALOGUE AVEC LA SALLE

Margaux Calon
Je voulais porter à votre connaissance des cartes qui utilisent les données d'Openstreetmap pour montrer les rues qui portent des noms d'hommes ou de femmes. On est très loin de 0,5.

Intervention de la salle
On peut se poser la question de la représentativité des données recueillies par cette application. Il me semble que les gens qui envoient des données seraient plutôt tentés de le faire en cas de non-parité qu'en cas de parité. Dans cette salle par exemple, il y a une certaine parité, il n'y a rien de choquant. Mais si on est dans un lieu où on voit 3 femmes pour 80 hommes ou l'inverse, c'est dans ce cas qu'on va envoyer des données. Il y a une question de représentativité.

Margaux Calon
En effet, il y a beaucoup de biais à partir du moment où l'on fait appel à un grand nombre de personnes. On n'a pas la prétention d'utiliser scientifiquement ces données. On s'est demandé à un moment si on pouvait travailler avec le gouvernement pour leur transmettre des résultats ou écrire des rapports, mais on n'a pas les moyens de le faire. Quant à la représentativité, on compte surtout sur notre communauté. On ne pense pas pouvoir toucher énormément d'utilisateurs, mais on veut des utilisateurs récurrents. On insiste sur le côté communautaire. Quant au comptage, ça devient un réflexe. Ce soir, je l'ai fait. On le fait systématiquement, qu'il y ait parité ou non, on ne se pose pas la question.

Question de la salle
Il y a des parités un peu mises en scène. À propos du gouvernement, le dernier qui vient d'être annoncé, par exemple. Il est strictement paritaire. Mais dans l'ordre protocolaire, les hommes sont d'abord nommés, puis viennent les femmes. La proportion d'hommes est plus importante en début de liste. Ce sont des modifications délibérées qui permettent de dire "On a une parité parfaitement respectée". D'ailleurs, cette manipulation n'a pas du tout été commentée dans les journaux alors qu'avec quelques petits graphiques, ça saute aux yeux.

Antonio Casilli
J'ai voulu télécharger l'application. Je l'ai fait. Et je l'ai désinstallée tout de suite, dès que vous avez prononcé le mot Google. Mais dans l'intervalle, j'ai remarqué que vous avez une interface intelligente pour faire réfléchir non seulement à un comptage basique, mais aussi à d'autres dimensions :

- Dans le public, qui a été la première personne à poser une question ?
- Ou parmi ceux qui posent des questions, est-ce plutôt des hommes ou des femmes ?

Sur la question de la représentativité statistique, ce n'est peut-être pas ce qui est le plus important. Ce qui est important, c'est que l'application nous oblige à regarder des situations qui ne sont pas des situations que nous souhaitons comme étant la normalité. Une manière de positionner cette application, c'est de la comparer à des applications mises en place par des activistes dans des pays qui ont des démocraties un peu plus troubles que la nôtre – quoique la nôtre soit un peu troublée. Ils ont mis en place des dispositifs pour dénoncer des situations d'irrégularités ou de fraude électorale. Des observateurs se rendaient avec des smartphones et l'application dans des bureaux de vote et enregistraient s'il y avait un, deux ou trois accidents liés au vote. Est-ce représentatif d'un point de vue statistique ? Non. Est-ce important de signaler des situations que nous ne souhaitons pas être normales ? Oui.

Jean-René Brunetière
Ça nous rappelle une situation qu'on avait connue dans une précédente Nocturne de Pénombre sur les suicides à France Télécom. Les lanceurs d'alerte avaient très facilement convenu que statistiquement, ça ne valait pas grand-chose. Mais de parler des suicides était pour eux le moyen le plus efficace pour faire prendre conscience des problèmes d'ambiance de travail dans l'entreprise. Les statistiques très relatives, voire pourries, peuvent être des éléments efficaces de prise de conscience.

Margaux Calon
Sur Google, pour revenir rapidement dessus. Nous avons été suivies dans le cadre d'un programme sur l'entreprenariat féminin. On a été accompagné par trois salariés de Google, notamment sur la communication.

Chantal Cases
J'ai chargé l'application, je ne l'utilise pas encore beaucoup. Je vous promets que je vais le faire. À Pénombre, on a pas mal réfléchi et contesté le chiffre militant. À juste titre. Mais là, c'est une autre initiative. Ce n'est pas le chiffre qui compte, c'est le fait de compter. Je voudrais saluer cette initiative. Elles ont la pêche : je vous invite à l'installer.

Alain Tripier
Après les avoir rencontrées chez Google, j'ai présenté l'application et la démarche à David Lacombled, qui est président de l'IAB et qui travaille chez Orange. Il est très préoccupé par la parité et visiblement, il pense que compter c'est très important. Je suis sûr que vous allez trouver avec Orange des choses qui pourront vous aider un peu. Il faut faire tout ce qu'on peut pour vous aider parce que la démarche est intéressante.

Intervention de la salle
J'ai une remarque qui n'a rien à voir avec la présentation. En règle générale, par rapport aux données de santé, on se préoccupe des pays développés, mais il y a aussi les pays émergents. Des situations dans lesquelles les gens n'ont rien. Soit vous avez votre petit *device* à 200 dollars, soit vous mourrez. Le *big data*, ça permet malgré tout d'avancer et de bouleverser la vie de millions de gens dans des pays émergents. La révolution *big data* va être plus importante dans les pays émergents que dans les pays développés.

Question de la salle
Je voulais savoir si l'application ItCounts s'applique seulement au domaine scientifique ou plus largement.

Margaux Calon
Initialement, je trouvais qu'il fallait qu'on s'intéresse aux sciences, qu'on commence par là. Parce que nous sommes des scientifiques. On a décidé de développer l'application pour les sciences, mais rapidement, on a eu des demandes. Alors on a décidé de l'étendre à de nouveaux domaines et disciplines. Quitte plus tard à faire plusieurs applications.

Jean-René Brunetière
Merci beaucoup.

8. ENJEUX DÉMOCRATIQUES ET SOCIÉTAUX

En prélude, une réflexion d'un pénombrien, **Bernard Aubry**, sur un sujet absent de cette nocturne mais qui touche aux enjeux démocratiques.

«Brave New World »: Full Speed Ahead !

Soumettre en direct les hommes politiques à l'épreuve du *fact-checking* est assurément une idée intéressante (cf. Lettre blanche n°62 - janvier 2016).

Rappelons en effet que l'auteur « *rêve de la diffusion d'un télétexte, sous les déclarations des politiciens à la télévision qui repérerait les informations d'un organisme tel que l'Insee sur le sujet traité* ».

Néanmoins il n'est pas certain que l'on y gagnerait en clarté car l'expérience nous montre que pour un même thème, il existe toujours une profusion d'indicateurs et au moins autant de chiffres.

Il faudrait aussi, en toute rigueur, obliger le politicien à revenir sur son propos afin qu'il s'explique sur le champ couvert par l'indicateur. Ce serait fastidieux ! À titre d'exemple, sur le chômage : quelle source ? Avec ou sans son halo ? Pour quel territoire ? La France ou la seule métropole ? À quelle date (en moyenne annuelle, trimestrielle, au 31.12, auquel cas il faudrait savoir si le chiffre est corrigé, ou non, des variations saisonnières, etc.) ?

Quant à savoir si l'augmentation d'un mois à l'autre est plus ou moins forte qu'en Allemagne, les choses se compliqueraient encore davantage…

Je proposerais plutôt une autre démarche qui aurait au moins l'avantage de rehausser le crédit des hommes politiques aujourd'hui tant décriés, trop souvent à tort. Le logiciel fournirait à l'orateur le chiffre « officiel » le mieux à même de soutenir son argumentation. Ainsi dans le cadre d'un débat télévisé en direct, l'homme politique lirait sur sa propre tablette le chiffre le plus flatteur qui soit à la défense de sa cause, l'adversaire présentant de son côté d'autres chiffres, évidemment favorables à sa propre argumentation.

D'autres innovations pourraient être imaginées dans le genre : un logiciel qui afficherait en temps réel les contradictions du discours avec ceux prononcés dans des interventions antérieures. Un autre calculerait un indicateur de « plasticité » pour mesurer la vitesse d'adaptation aux changements de l'opinion (quelque chose comme une girouette numérique). Et pourquoi pas un indicateur de « normalité », l'homme politique « normal » étant défini comme celui dont le discours s'adapte le mieux à l'attente moyenne des citoyens mesurée par un indicateur synthétique établi à partir de l'ensemble des sondages du moment ?

En notre temps, post-moderne, la femme n'est plus l'avenir de l'homme : le couple traditionnel est condamné. Place au couple du statisticien et de l'informaticien qui, dans une passion commune pour le big data, assurera à l'humanité un avenir radieux... à moins, bien sûr, qu'il n'engendre des monstres.

Jean-René Brunetière

Essayons de faire une petite revue des problèmes que tout ça pose du point de vue éthique, moral, etc.

La CNIL qui a été créée en 1978 dans un contexte assez différent de ce qu'on connaît aujourd'hui est aux prises avec des phénomènes assez nouveaux ; j'ai appris que la CNIL faisait de la recherche. Geoffrey Delcroix travaille dans la recherche à la CNIL. Il va nous raconter tout ça avec Antonio Casilli. Entrons dans tous ces problèmes.

Antonio Casilli

À moi d'introduire Geoffrey et de démarrer cette dernière partie. Elle est consacrée à certains des aspects sociaux, politiques, ludiques qui sont présentés et mis en évidence par les big data et de manière plus générale, par ce dont on a parlé ce soir. On va chercher à se recentrer un peu.

On va commencer avec une petite présentation de Geoffrey Delcroix. Déjà, on va préciser que lui, ce n'est pas la CNIL, mais un aspect particulier de la CNIL. Car c'est une institution qui a plusieurs divisions. Il y a les juristes, qui s'occupent du droit, les ingénieurs qui s'occupent de développer des solutions logicielles pour certains des enjeux qu'on va développer ce soir, il y a une arme politique de la CNIL qui rend des avis sur les projets du gouvernement. Et il y a une quatrième arme, qui est celle de l'innovation et prospective. D'ailleurs, je vous invite à suivre une des initiatives de cette partie de la CNIL : les cahiers IP, *Innovation et Prospective* publiés tous les ans. Il y a eu plusieurs livraisons, dont l'une était consacrée aux objets connectés (*Le corps nouvel objet connecté*), une autre sur la place des données dans les industries culturelles ainsi que sur le partage dans le monde numérique.

Geoffrey Delcroix

Merci pour cette introduction Antonio c'est parfait. Je m'appelle Geoffrey Delcroix. Je suis chargé d'études au sein du pôle innovation études et prospective de la CNIL. Je ne suis pas la CNIL, je suis un morceau et je n'en représente pas une voix officielle.

Je vais commencer par deux mots sur la CNIL et reprendre un certain nombre d'éléments pour que le cadre soit un peu clair. La CNIL a été créée à la suite de la loi Informatique et liberté en 1978 et encadrée par une directive européenne en 1995. Aujourd'hui, on est dans une phase un peu particulière puisqu'il y a un règlement européen en cours d'adoption qui changera pas mal de choses dans le pouvoir des autorités ainsi que les droits et les règles qui s'appliquent. On est dans une phase avec une forte internationalisation de nos activités, question de mondialisation.

Rassurez-vous, ça fait un petit moment qu'on a compris que les acteurs du numérique n'étaient pas des acteurs nationaux. Le droit essaie de prendre ces éléments en compte. Forcément, ce n'est pas facile, c'est un long combat. Parce que quand on fait certaines interprétations c'est normal, elles sont contestées devant la justice qui ensuite décide qui a raison. Parmi les exemples d'internationalisation, il y a un groupe qui s'appelle le G29, et qui regroupe l'ensemble des CNIL européennes qui essaient d'agir ensemble, de donner des avis, d'avoir des positions communes ou de poursuivre certains acteurs internationaux.

Je vais faire une petite présentation, je vais essayer d'aller assez vite.

Source : Open Data Institute, Ulrich Atz
http://theodi.github.io/data-definitions/

Déjà, c'est intéressant de voir que les trois sujets big data, open data et données personnelles sont trois sujets pour lesquels les définitions sont un peu difficiles. Parfois on parle de choses qui existent depuis longtemps, parfois, on parle de choses complètement nouvelles.

Sur l'open data, je crois que c'est un mot qui est très à la mode et du coup, il est utilisé pour des choses qui ne sont pas du même niveau. Ça a été très bien montré avec le «presque open data» dans le domaine de la santé. Je pense que c'est exactement ça : l'open data pur, c'est l'idée qu'il n'y a aucune limite à la réutilisation. Mais dans l'open data santé, on n'appelait pas toujours à la mise à disposition publique de données sans aucun contrôle, on appelait plutôt à ce qu'elles soient un peu plus ouvertes par rapport à la situation actuelle qui est très fermée.

L'expression « données personnelles » est aussi assez difficile à définir. Et quand on commence à s'interroger sur les intersections éventuelles entre ces différents domaines c'est encore plus compliqué. Même l'Open Data Institute, quand ils ont essayé de donner une définition et qu'ils ont été mis face à du «big-open-personal data», la seule chose qu'ils ont été capables de faire a été de mettre un point d'interrogation !

Sur le sujet des données personnelles, la définition est compliquée à comprendre parce que ce n'est pas juste une question de données directement identifiantes (nom, prénom, etc.) mais c'est un ensemble donné d'informations qui peuvent être liées de manière directe ou indirecte à une personne qui est identifiée ou identifiable. Est-ce qu'un objet connecté dans la maison est un producteur de données personnelles ? Ce n'est pas une question de pureté de la donnée. Mais, si c'est utilisé dans un contexte qui permet d'apprendre des choses sur le comportement et le mode de vie d'un foyer ou d'une personne, dans ce cas, ça va être considéré comme des données qui méritent une forme ou une autre de protection. Ce n'est pas facile à saisir ; on peut avoir un champ extrêmement large de données concerné par ces questions de protection de données. De même, également parce que spontanément – et c'est normal – on pense tout d'abord aux identifiants directs et aux choses assez traditionnelles, mais on peut avoir des informations identifiantes de matching (agrégation de données) qui sont extrêmement efficaces pour vous reconnaître et parfois plus efficaces qu'un nom ou un prénom. Il peut y avoir beaucoup d'homonymes, alors que votre téléphone a une adresse MAC unique, une carte Wi-Fi a un identifiant d'abonné et un IMEI : tout un tas de choses qui peuvent être plus efficaces pour vous tracer que des identifiants très primaires.

La CNIL réfléchit à ces questions-là. On est au cœur de ces sujets et ce sont des choses sur lesquelles on travaille pas mal. Il y a quelque chose qui est très important à dire, j'insiste toujours beaucoup là-dessus, c'est le premier article de la loi fondant la CNIL.

«L'informatique doit être au service de chaque citoyen. Son développement doit s'opérer dans le cadre de la coopération internationale. Elle ne doit porter atteinte ni à l'identité humaine, ni aux droits de l'Homme, ni à la vie privée, ni aux libertés individuelles ou publiques.

Toute personne dispose du droit de décider et de contrôler les usages qui sont faits des données à caractère personnel la concernant, dans les conditions fixées par la présente loi.»

Ce qui est intéressant, c'est que ce n'est pas une question de protection des données : on ne protège pas des données, ce n'est pas un droit patrimonial. Ce n'est pas une question de simple protection. C'est une question de faire en sorte que les technologies numériques se développent sans porter atteinte à la dignité humaine, à la vie privée, aux droits de l'Homme et aux libertés individuelles et publiques. C'est un article plutôt intéressant dans le contexte actuel. Il faut le prendre de plus en plus comme une question qui n'est pas une question technico-technique de droit, de protection de données et de choses qui peuvent se rapprocher des droits des bases de données mais bien comme un sujet relié aux libertés publiques, aux droits de l'Homme, aux libertés individuelles, à l'autonomie des individus dans la société.

Ce qui me permet de réagir à autre chose : c'est vrai qu'il y a de vraies différences entre l'Europe et les États-Unis de ce point de vue-là. Ça c'est vraiment la vision européenne de la protection des données comme un élément constitutif des droits humains, des droits civiques. Aux États-Unis, il y a, *a priori*, une logique de protection des droits des consommateurs. Mais il ne faut pas exagérer, il y a des règles très strictes aussi aux États-Unis. Elles sont plus souvent mises en exercice par des juges – parce que c'est plutôt la culture américaine, avec des juges qui peuvent faire beaucoup de dégâts aussi – qui ne sont pas toujours aussi coulants que les régulateurs européens. Ce sont eux qui ont inventé le droit à la vie privée comme le droit à être laissé tranquille : le « *right to be left alone* », c'est un juge de la Cour suprême qui l'a défini comme ça.

Nous, on se retrouve dans une situation qui a effectivement beaucoup changé depuis 1978 par rapport à l'utilisation des fichiers. On est plongé dans ce lac de données, ce tsunami des données et on a cette idée qui tourne autour de nous que les données sont le pétrole de l'économie numérique. Cette expression-là, elle est intéressante, car finalement elle est toujours utilisée comme quelque chose de très positif alors que je ne suis pas certain que ce soit l'exemple d'une ressource utilisée de manière raisonnable et durable. Si les données sont le pétrole de l'économie numérique, il faut peut-être réfléchir sur comment consommer les ressources de manière responsable. C'est là que notre activité en tant que pôle d'innovation existe pour incarner le fait que le rôle d'un régulateur n'est pas celui d'un empêcheur d'innover en rond. Le but n'est pas d'empêcher les choses de se faire mais de créer un cadre d'exploitation des données qui soit respectueux des droits des individus et qui ne les laisse pas comme des objets qu'on traite, mais les considère comme des adultes, des citoyens, dignes de respect.

Notre activité n'est pas de piloter vraiment la recherche, mais d'avoir une approche pluridisciplinaire, d'explorer des sujets et d'être un point de contact des écosystèmes. En gros, on discute avec des gens qui spontanément n'ont pas envie de discuter avec nous : des startups, des équipes d'innovation dans les grands groupes industriels.

Par rapport au sujet de ce soir, j'ai effectivement eu envie de vous parler du sujet de notre deuxième cahier *« Le corps nouvel objet connecté »*. On a voulu s'intéresser au sujet de la santé et du bien-être dans le monde numérique. Plutôt que de partir des sujets traditionnels qui sont extrêmement importants – par exemple les données de santé –, on a voulu partir des usages qui sont de l'ordre du signal faible de quantification des activités : les petits bracelets qui comptent le nombre de pas ou qui peuvent mesurer les cycles du sommeil. On a publié un document sur le sujet. Il y a une chose qui nous a intéressés : pourquoi les gens se mettent à utiliser des choses comme ça ?

Anne-Sylvie Farapo a fait une typologie des selfs quantifiers qui mesurent leur activité :

- logique de surveillance d'une constante ;
- logique de routinisation, essayer de marcher un peu plus ;
- logique de performance, est-ce que j'ai couru plus vite.

Nous aussi on rajoute quelque chose qui manque un peu : un effet de curiosité qui s'incarne dans le fait qu'un bracelet finit dans un tiroir. Il y avait différentes questions qui nous ont intéressées. La première question, c'est par rapport à la nature des données : le nombre de pas, ce ne sont pas des données de santé. La loi informatique et libertés, elle, définit des données sensibles par nature (appartenance religieuse, ethnique, données de santé, sensibles par nature, des données prises dans le contexte médical) pour lesquelles des règles très strictes vont se poser. Là, ce ne sont pas vraiment des données de santé. Ce sont des données annexes, d'une activité anodine. En revanche, si j'ai votre nombre de pas par jour sur trois ans, peut-être que je peux faire des inférences plus intéressantes sur votre risque futur de maladies cardio-vasculaires. Si en même temps, j'ai votre courbe de poids sur la même période peut-être que je peux deviner des choses qui sont en revanche très sensibles. Sur les cycles du sommeil, on n'est plus dans l'ordre de la recherche, mais il y a un certain nombre de pathologies qui peuvent apparaître par le biais de troubles du sommeil bien avant qu'il y ait des symptômes plus visibles.

Je ne suis pas du tout spécialiste. Pourtant à la base le rythme du sommeil, ça peut paraître anodin, il y a une logique qui est complètement déconnectée de la sensibilité potentielle de ce genre de choses. Vraiment c'est un point qui est important pour nous : c'est cette notion de données sensibles par la manière dont elles sont utilisées bien plus que par leur nature initiale, nature de la donnée en elle-même.

Du coup, on s'est posé la question de savoir quels sont les usages qu'on peut avoir.

L'exemple des assureurs est pris très régulièrement, donc je ne vais pas revenir dessus, mais quand les assureurs le font pour réduire la sinistralité globale, si les gens font plus attention, ce n'est pas forcément gênant. Si en revanche ça sert à faire de l'individualisation du risque, ce n'est pas du tout la même chose. Il faut une régulation qui soit sur la manière d'utiliser ces informations.

Cédric Hutchings, le patron de Withings, nous avait dit : *« Finalement, dans le futur, on trouvera tous très étonnant que dans le passé on n'ait pas eu de tableaux de bord de la santé au quotidien. Ça paraîtra comme une bizarrerie ».* J'ai trouvé que c'était une prédiction, un scénario plutôt intéressant. Dans le futur, on aura un tableau de bord en permanence de notre santé au quotidien. Du coup on a fait effectivement cette comparaison – quelqu'un évoquait le *« pay as you drive »* tout à l'heure, ce sont ces systèmes assurantiels fonctionnant par rapport à la manière dont les gens conduisent –, on est dans le *«pay as you walk»* pour avoir votre activité.

Antonio Casilli

Entretemps, ils ont vraiment lancé la startup, qui s'appelle Bitwalk.

Geoffrey Delcroix

Aujourd'hui aux États-Unis, il existe des choses dans les entreprises de complémentaire santé. Une startup va voir les entreprises et leur propose : « *Vous donnez des Apple Watch à vos employés, les Apple Watch comptent le nombre de pas par jour. Et, si en un mois ils n'ont pas atteint l'objectif, ils doivent payer 15 dollars pendant 2 ans* ». Si vous n'avez pas fait le nombre de pas, ça va vous coûter bien plus cher qu'une Apple Watch. C'est amusant parce qu'on essaie toujours de trouver des choses qui sont indirectes pour donner l'impression qu'on n'est pas en train de toucher à la prime d'assurance. À la fin ça revient au même, c'est assez étonnant !

Nous avons fait un petit scénario qu'on a appelé « Léa et ses capteurs » ; ça ressemble à ce qu'on a vu tout à l'heure dans les petits sketches. Les gens accrochaient les bracelets à leurs chiens pour faire le nombre de pas, et donc ça nous a interpellés par rapport à la question de la norme, en particulier de normes chiffrées qui sont très étranges, du type « 5 fruits et légumes par jour », « 10 000 pas par jour ». On a dans notre document un avant-propos d'Antoinette Rouvroy, où elle évoque cette idée de maladie de la norme, de l'Homme normal. Il faut qu'on atteigne cette idée qu'on soit normé, normal. Il y a un aspect très normatif dans ces quantifications d'activités qui mérite qu'on se pose la question. Dans cet avant-propos, Antoinette Rouvroy mettait en avant l'idée que les pratiques de quantification et de mesure font des individus des entrepreneurs de leur propre santé. On remet au niveau individuel des sujets qui sont aussi des questions de société.

L'autre élément – et je renvoie au cahier qu'on vient de publier en fin d'année et qui se tourne vers les plateformes de vidéos et de musique en ligne – c'est la question de la recommandation. En ce moment, c'est le mot algorithme qui est à la mode. Ce qu'on trouve extraordinaire, c'est d'entendre à la radio, le patron de Deezer, quand on lui demandait comment marchent ses algorithmes de recommandation, répondre « c'est la magie des algorithmes ». Peut-être n'avait-il pas envie de rentrer dans le détail à la radio ou peut-être ne comprenait-il pas lui-même comment ça fonctionne.

Nous sommes face à cette situation où on se rend compte qu'il y a un discours très magique autour de ces systèmes algorithmiques. Ça a déjà été évoqué. Et quelque part ça nous interpelle, parce que l'idée c'est de faire disparaître les technologies et qu'à force de faire disparaître la technologie, l'utilisateur n'a plus les moyens de savoir comment il se retrouve dans cette situation, ni pourquoi on lui a recommandé ça. On infantilise beaucoup l'utilisateur, par ces systèmes, sous prétexte de transformer l'expérience d'utilisation en quelque chose de sans couture et de très invisible. Du point de vue d'une autorité de régulation qui pose comme principe de base de mettre l'individu et ses choix au centre du système, c'est difficile à appréhender et c'est important de réfléchir autour de ça.

Je voulais juste vous montrer un petit dessin de Vidberg où un homme dit : « *Les données de vos clients sont des biens que vous pouvez vendre, c'est totalement éthique puisque vos clients feraient la même chose s'ils le pouvaient.* » Le patron répond « *Ça semble équitable* », ce à quoi l'homme ajoute : « *La première étape, nous allons déshumaniser l'adversaire en l'appelant "data".* »

Ça renvoie à ce qu'on essaie de faire. Ces éléments de loi s'appliquent moins aux développements des technologies modernes, qu'ils n'essaient de renvoyer à la question : comment replace-t-on l'individu au cœur du système ? Et en quoi « de grands pouvoirs impliquent de grandes responsabilités » comme le dit un grand philosophe.

Quand des entreprises ou des responsables de traitement accumulent autant de données, ce n'est pas forcément un problème mais ils doivent être mis en face de leurs responsabilités, qui sont importantes.

Antonio Casilli prend la parole pour l'ultime communication.

C'était vraiment intéressant, surtout d'avoir choisi un angle d'attaque qui est très « micro ». Regarder la mise en chiffres de soi, à travers le quantified self ou les objets connectés qu'on emmène avec soi, à nos poignets dans nos poches ou parfois dans notre corps...

La question est que les big data et open data ont aussi une dimension macrosociale et macro-économique. Je pense que l'illustration de ce croisement entre big data et open data, peut-être pas des plus rassurant, se trouve chez les data brokers, un type d'opérateurs et d'entreprises bien particulier. Il y en a un nombre limité dans le monde. Un rapport récent de la FTC (Federal Trade Commission aux États-Unis) parlait de six grands data brokers qui normalement achètent, ou parfois captent, des données produites soit par les utilisateurs d'une plateforme, soit par des objets connectés. Parfois même, ils achètent des données qui vous concernent aussi. Ces data peuvent provenir de données mises en ligne par les gouvernements, d'achats, des enquêtes traditionnelles ou encore, de manière un peu grise, des données tirées de la statistique publique. Les data brokers sont capables de créer et de collecter une masse de données. Ce qui était, il y a quelques années, inconcevable du point de vue strictement du chiffrage c'est à dire du nombre de données, mais aussi du chiffre d'affaires.

La chose la plus importante qu'il faut souligner c'est que les data brokers, et en général toutes les plateformes et les entreprises qui produisent de la donnée aujourd'hui, font partie d'une économie de la donnée personnelle. Et pas seulement personnelle d'ailleurs, mais de la donnée en général.

C'est important de le souligner parce que derrière la question des données aujourd'hui il y a la question des valeurs extraites à partir de nos données personnelles. Je ne vais pas vous faire le coup de vous poser la question « *Combien évaluez-vous votre date de naissance ou le nom de votre chien* ? » qui serait pourtant cruciale pour récupérer votre mot de passe.

On revient à la question : qu'est-ce qu'une donnée personnelle et quelle est sa valeur ? Si on vous pose la question, normalement c'est pour opérer une négociation à la baisse. Or, cette question nous est posée de manière assez régulière. Parce qu'en gros la plupart des plateformes numériques qui gèrent ou occupent notre vie ou nos pensées sont des plateformes basées sur une économie des data qui est une économie qui ne dit pas son nom. Elle est toujours présentée comme une invitation à participer ou à avoir des amis, ou encore à mettre en ligne, ou à s'exprimer.

On parle de partage, de « sharing economy » mais il s'agit surtout de production de données brutes. On est face à une énorme usine à data et surtout au besoin, qui est un besoin politique aujourd'hui, de reconnaître que les données sont une forme de valeur et surtout une valeur qui est produite par notre activité humaine.

Même une activité passive à nos yeux, comme se déplacer d'un point à l'autre avec un objet connecté, est une activité qui produit des données. Ou alors, pour reprendre l'exemple des données de santé : au moment où on a des données de santé, quelqu'un doit les saisir, que ce soit votre pharmacien avec la carte Vitale, votre médecin quand il remplit un formulaire ou vous-même sur le site d'Ameli ou d'autres plateformes, vous êtes en train de faire un travail de production de données. Quelque part, il y a quelqu'un qui, dans la meilleure des hypothèses, est capable de les utiliser pour des finalités de bien public ou de progrès de la société. Mais dans la plupart des cas, il y a des manières, pour les intermédiaires, de récupérer ces productions. Et dans ce cas de figure, cela concerne aussi bien les données qui sont émises que les données qui sont publiées ou partagées en faisant appel à notre volonté. C'est très important, surtout dans une économie de plateformes sociales, à laquelle on s'est habitué depuis 10 ans, avec Facebook, Twitter...

L'envie de s'exprimer cache finalement toujours une économie de données. Chaque photo de chaton que vous mettez en ligne est à vos yeux une simple photo de votre animal de compagnie, mais pour la plateforme ce sont avant tout des métadonnées (appareil photo, date, lieu, adresse IP) et tout ça a plus de valeur que la simple photo à vos yeux. Il faut s'habituer à faire une prise de conscience politique et à admettre que la data, c'est du travail. On est face à une nouvelle forme de travail. Une manière de protéger et de reconnaître ces données serait d'admettre qu'elles sont liées à une activité travaillée. Et peut-être que ce n'est pas le droit privé qui va protéger nos données personnelles mais le droit du travail.

Si je vous le dis comme ça, ça peut paraître une provocation, mais sachez aussi qu'il y a une réflexion qui devient de plus en plus importante et qui circule dans certains milieux activistes et politiques : la réflexion autour du *digital labor.* On le dit en anglais, parce qu'on n'est pas en train de parler des gens qui travaillent dans le numérique, mais des usagers de ces plateformes.

On peut les passer en revue rapidement, il y a toujours des productions de données :

- la première, c'est à travers la circulation de contenus partagés sur des plateformes sociales. On en a déjà parlé ;
- le deuxième exemple, ce sont les données qui viennent de l'internet des objets, liés à la domotique (thermostat, poubelles, frigos...). On peut imaginer que ces outils complotent entre eux et avec le réseau Wi-Fi pour envoyer toutes leurs données à Amazon, qui ensuite essaiera de vous vendre des livres de recettes de cuisine pakistanaise parce qu'il a repéré que vous aimiez ça ;
- sans parler encore de ce que je suggérais : cette idée de l'économie du partage qui est un secteur émergent de l'économie, pour l'instant limité, la sharing economy ; c'est aussi une économie des data. Prenons l'exemple d'Uber, qui se présente comme une entreprise de transports urbains mais qui est en fait une entreprise de données. Elle fait commerce des données. Il y a les données des passagers (dans quel quartier, à quelle heure, vous y allez ?), mais aussi les données des chauffeurs eux-mêmes. Si vous êtes un chauffeur Uber – je vous invite à tester l'application –, une application qui est ludifiée, qui vous donne des scores en tant que chauffeur, en tant que passager et qui devient addictive. Et si vous êtes un chauffeur, vous passez votre temps à vérifier votre activité sur un tableau de bord. Puis vous consacrez un temps vraiment important à choisir la bonne photo, à renseigner votre profil. Les passagers évaluent ensuite les chauffeurs et les chauffeurs évaluent également les passagers. Si vous avez un score de moins de 4,7, bonne chance pour trouver la voiture à 3 heures 30 du matin dans le 3ème arrondissement !

Cela amène une nouvelle question, celle que certains juristes appellent la sur-subordination. C'est-à-dire une subordination caractérisée par une forme de surveillance. Si les données peuvent être envisagées comme une production et donc une forme de travail, est-ce que nous ne sommes pas face à une subordination ? Nous n'avons pas de contrat de travail avec Uber, pas plus que les chauffeurs qui n'ont pas un contrat de travail avec Uber. Juridiquement, il n'en est pas question. À certains moments, vous êtes poussés à ouvrir l'application. Si vous êtes un chauffeur, vous devez être présent à certains endroits et heures. Vous êtes contraints de produire une prestation. Des formes de prestations qui passent par la production de données, qu'on pense à Uber ou Fitbit. Si c'est votre employeur ou ssureur qui vous demande de produire, il y a une forme d'injonction.

La question peut se poser alors : est-ce que nous sommes face à des formes de subordination qui s'expriment à travers une mise en chiffres ? Une forme qui ne vient non pas de soi, qui n'est pas choisie mais subie. Une mise en chiffres qui ne relève pas de ma propre surveillance, mais d'une surveillance que d'autres essaieraient de m'imposer.

On revient alors à une question qui a été trop peu évoquée : la question de la surveillance de masse. C'est une question qu'on ne peut plus évacuer depuis le 6 juin 2013, le jour où Edward Snowden a fait sa première révélation. Pardonnez ce langage un peu christique, mais c'est quelque chose qui a changé notre manière d'envisager les technologies et les enjeux liés à ces technologies. Le débat qu'il faut avoir au niveau social, au niveau politique, c'est un débat qui porte sur la question du lien de subordination avec les plateformes. C'est un débat que nous devons avoir d'abord avec nos hommes et femmes politiques, en nous posant la question de la surveillance de masse. Depuis qu'Edward Snowden a commencé à révéler l'étendue du pouvoir de la NSA, la France a décidé que la NSA c'était mal et a donc pensé qu'il fallait faire la même chose chez nous. Elle a enchaîné une procession interminable de lois liberticides qui en plus sont basées sur une exploitation de données, un traitement algorithmique de données – ce que l'on a appelé les « boîtes noires » – qui codent certains comportements mais aussi identifient des signaux faibles. C'est un débat que nous devons avoir tous ensemble, aujourd'hui, demain et pour les années à venir malheureusement.

Jean-René Brunetière
Merci beaucoup. On termine sur l'ouverture d'un débat. On ne va pas le tenir cette nuit, on a déjà très largement dépassé l'horaire, comme c'est la tradition.

Intervention de la salle
Snowden c'est nouveau, mais ce qui est derrière Snowden ce n'est pas du tout nouveau. Est-ce que la CNIL va changer sa politique pour correspondre à la politique américaine ? J'ai fait deux demandes à la CNIL concernant mes dossiers au ministère de l'Intérieur et de la Défense, et à chaque fois, j'ai reçu la même réponse : « *Nous avons mis en conformité vos données* ». Je n'ai pas le droit de les voir.

Geoffrey Delcroix
Ça s'appelle le droit d'accès indirect et dans le domaine des fichiers étatiques régaliens, les gestionnaires des fichiers en question ont la possibilité de ne pas transmettre les informations. Si vous voulez modifier ça, il faut faire changer la loi, et pas la CNIL.

Intervention de la salle
Il y a peut-être une autre composante que nous aurions pu introduire dans une troisième partie. L'ONU s'est préoccupée récemment de ce qu'on appelle la data-revolution. Ils ont produit un rapport qui s'appelle « *Data-revolution: a world that counts* ». Je trouve que c'est aussi un sujet sur lequel il faut qu'on introduise un débat.

L'essentiel de l'argumentation est de dire que le monde a besoin de politiques publiques et que ces politiques reposent sur des données qui sont essentiellement des données du monde privé. Il est suggéré un accord entre la puissance publique et les entreprises privées pour qu'elles donnent leurs données pour le bien public.

Cela ne correspond ni au raisonnement des grandes boîtes pour qui la donnée est la matière première de leur création de valeur. Faut-il donc créer l'impôt « en données »... C'est peut-être un enjeu aussi central que celui des libertés individuelles.

Geoffrey Delcroix
Ça mériterait un débat plus complet, mais j'attire votre attention sur la notion de données d'intérêt général et données de références.

Jean-René Brunetière
Il y a des tas de choses intéressantes à dire. Et après, ça continue autour du buffet.

Et pour compléter cette intervention, vous pouvez consulter les documents suivants :

Paola Tubaro, Antonio Casilli. Enjeux sociaux des Big Data. Mokrane Bouzeghoub, Remy Mosseri. *Les Big Data à découvert*, CNRS Éditions, pp.292-293, 2017. <hal-01456369>

« L'exploitation du moindre clic par l'industrie numérique », entretien avec Antonio Casilli, sur le site www.JEFKLAK.ORG/JANV.2015/MARABOUT.

Ce volume contient :

Nul ne peut se prévaloir de sa propre turpitude

Adresse postale : Pénombre, 32 rue de la Clef, F 75005 Paris **Courriel** : redaction@penombre.org
Site internet : http://www.penombre.org