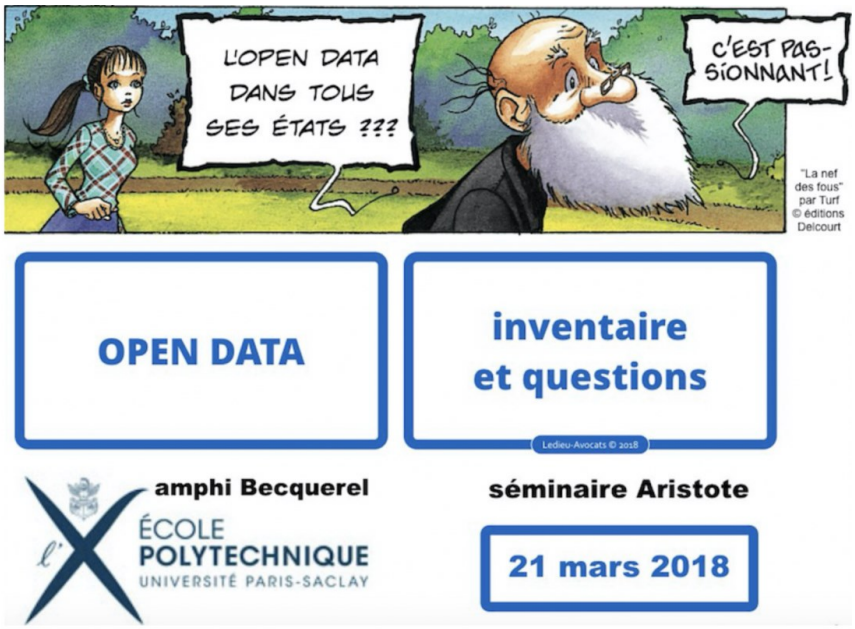
L'Open data dans tous ses états, Open science, RGPD

École Polytechnique - Palaiseau

Mercredi 21 mars 2018



Coordination Scientifique

Bernard Monnier (MIM), Valérie Masson-Patrimonio (Ecole Polytechnique)
Christophe Calvin (CEA)

Partenaire

Editorial Board

L'Open data dans tous ses états, Open science, RGPD

Séminaire Aristote, 21/03/2018 à l'École Polytechnique

Coordination scientifique

Bernard Monnier (MIM), Valérie Masson-Patrimonio (Ecole Polytechnique)
Christophe Calvin (CEA)

Table des matières

Compte-rendu des interventions

Introduction

Bernard Monnier a introduit la journée en mentionnant que « Open », ce mot clé, est devenu l'un des plus utilisés dans la plupart des journaux, revues ou documents sur Internet, tant utilisé qu'il en devient galvaudé : open source, open data, open hardware, open innovation, open science, … quelle réalité derrière ces expressions ?

Il a précisé qu'on est passé d'un monde fermé avant les années 2000, règne de la confidentialité, à des organisations ouvertes. L'ouverture est devenue une obligation pour innover, pour partager les risques, compléter les offres par des éléments différenciant provenant de partenaires complémentaires pour partager ensuite les bénéfices engendrés par ce mode collaboratif devenu incontournable.

Cette ouverture est aujourd'hui prônée au plus haut niveau de l'état. On parle d'open data dans les différents ministères pour permettre aux startups de créer de la valeur grâce à ce trésor précieux que représente la data et que les GAFA ont su utiliser pour réussir à devenir des leaders. On parle d'open science pour rendre accessible la

connaissance aux entrepreneurs et leur permettre de créer de la valeur grâce aux résultats de la recherche.
Cette stratégie d'ouverture est louable mais il n'en demeure pas moins qu'elle est délicate à appréhender, tant au niveau de la production des nouvelles idées que de l'utilisation des résultats. La journée sera consacrée à proposer un éclairage sur ces deux points.
Les questions liées à la production et la diffusion de données issues de différents domaines seront présentées par des responsables d'administrations disposant de données importantes à mettre à disposition.
Les questions d'utilisation seront abordées par des professionnels du secteur privé, des entrepreneurs, des juristes, des avocats…
Compte tenu de l'actualité du sujet, un focus sera mis sur le cas particulier d'utilisation de données personnelles, afin d'évoquer la mise en application le 26 mai 2018 du Règlement Général européen sur la Protection des Données (RGPD, en anglais GDPR, General Data Protection Regulation). La CNIL, les juristes et les professionnels touchés par cette nouvelle réglementation nous parleront de leur solution pour satisfaire aux exigences de ce nouveau règlement.

1. Les enjeux de l'open data, du RGPD, le cahier innovation et prospective de la CNIL

Régis Chatellier (CNIL)

Régis Chatellier est chargé d'études innovations et prospectives au laboratoire d'innovation numérique de la CNIL

Dans un cadre légal en évolution, de la loi Numérique au RGPD, les acteurs publics comme les acteurs privés souhaitant mettre en place une politique d'open data doivent respecter certains cadres protecteurs de la vie privée et des libertés des individus. Au-delà de l'open data et de sa définition première, sans son cahier Innovation et prospective consacré à la ville numérique, LINC explore dans une matrice à quatre entrées des scénarios de rééquilibrage public/privé par la donnée, une matrice mobilisable dans d'autres projets de partage de données.

Comment faire de l'open data tout en respectant la protection des données personnelles ? Pour cela il s'appuie sur une étude menée dans le cadre de la ville intelligente et qui est publiée sur le site du Laboratoire d'innovation numérique de la CNIL (LINC). La Cnil est un régulateur complet, qui agit sur toute la chaine, la seule administration à avoir pour mission, à la fois, d'informer et conseiller, mais aussi de contrôler et de sanctionner. Son rôle est aussi de produire des avis sur les textes notamment en 2018 l'adaptation de loi informatique et libertés au RGPD. Mais c'est bien le législateur qui écrit la loi, non la Cnil, qui a pour mission de l'appliquer.

Le RGPD en vigueur depuis avril 2016, et dont l'entrée en application lest fixée au 25 mai 2018, s'inscrit dans le prolongement de la loi informatique et libertés, reprenant 80 à 90 % de ses principes. La Cnil c'est 200 personnes : 70 % de juristes et 30 % ingénieur, docteur en informatique et chiffrement, qui travaillent le plus souvent en binôme. La Cnil produit aussi des études, des cahiers sur des sujets innovants afin d'éclairer certains sujets. Par exemple, avec le cahier « Corps : nouvel objet connecté », publié en 2014, il s'agissait de mesurer à quel moment les données de bien-être deviennent des données de santé.

L'open data sur les données publiques sont gratuites, fournies dans un format ouvert et permettant la réutilisation des jeux de données suivant un régime de licences, sans principe de finalité. Faire de l'open data implique de ne pas avoir de droit de regard sur leurs utilisateurs. Les données sont réutilisables sans qu'il y ait de problématique juridique pour ce qu'il voudrait en faire, ce qui interdit la publication de données personnelles, régies par un principe de finalité : des données collectées pour certains usages ne peuvent être utilisées pour d'autres finalités. La définition des données personnelles est très large. Une donnée personnelle peut être directe (plaque d'immatriculation de véhicule, carte bancaire ou de sécurité sociale, etc., ou indirecte (géolocalisation, identifiants en ligne, éléments propres à l'identité physique, psychologique, économique, culturelle etc.). Dès lors que l'on peut retrouver par identification direct, ou par recoupement ou par inférences, on entre dans le champ des données à caractère personnel et de leur protection.

Quand on collecte des données à caractère personnel, on doit le faire pour des finalités déterminées, définies, explicites et légitimes et ne pas les traiter d'une manière incompatible avec les finalités première. Il est difficile de faire coïncider le « faites-en ce que vous voulez », et la partie donnée personnelle « garantissez aux personnes dont vous avez collecté les données de ne pas les réutiliser d'une manière sur laquelle, il n'aura pas de contrôle ». On parle aussi du droit à la portabilité qui offre aux personnes la possibilité de récupérer une partie de leurs données dans un format ouvert et lisible par machine. Elles peuvent ainsi les stocker ou les transmettre facilement d'un système d'information à un autre, en vue de leur réutilisation à des fins personnelles.

L'anonymisation des données est un processus loin d'être figé. Si Lorsque l'on parle de l'open data des données publiques, il faut aussi tenir compte des données des grands acteurs privés qui peuvent intéresser la sphère publique. Le législateur est placé en face du besoin de définir l'obligation de produire de l'open data, qui peut être utilisé également par des acteurs privés, des chercheurs, voire des concurrents. Il faut aussi prendre en compte la qualité des données qu'ils fourniront, leur niveau de granularité vs agrégation, voire les techniques de floutage de données « Où mettre le curseur ? ».
Par ailleurs, il annonce la possible création d'un guichet unique, entre Cada (Commission d'accès aux documents administratifs) et Etalab (mission créée en 2011 chargée de la politique d'ouverture et de partage des données publiques du gouvernement français) qui fournirait une boîte à outils destinée à tous les acteurs. Se pose aussi la question de la certification de l'anonymisation des données, de l'obligation de moyens, de contrôle a posteriori, et de l'ouverture du marché du conseil et de l'audit de la pertinence du mécanisme d'anonymisation (la désanonymisation n'existe pas). Ce marché est en cours de création. La CNIL devrait respecter un temps d'apprentissage. Régis Chatellier pointe aussi les stratégies défensives de certains acteurs vis-à-vis de la portabilité (« rien ne se

passe, personne ne va rien demander, faire a minima et régler les contentieux »), avec le risque de perdre des opportunités, ou offensif en tirant avantages des opportunités réelles de la portabilité, ou de l'open data.

2. Massive Data Analytics for Strategic Intelligence

Thanh-Long Huynh (Quantcube Technology)

QuantCube Technology utilise des données open data publiques, et pratique l'analyse simultanée de multiples sources de données permettant des prises de décision dans le domaine de l'économie et de la finance. Ces sources de données incluent notamment les réseaux sociaux, les données satellites, les données météorologiques, les données océanographiques. Mais outre les données publiques, de plus en plus des grands acteurs de la donnée la sollicitent pour exploiter leurs données en termes d'application.

Prédire l'avenir, cela passe par l'analyse de tous types de données des blog, des réseaux sociaux, des offres d'emploi, du climat et océanographiques pour anticiper les rapports agricoles, du mouvement des navires dans le monde, de la hauteur des vagues, des images des satellites, des cold cases ou du stress de l'herbe, autant de phénomènes qui impactent la croissance économique), les donnée satellites systématiques (analyse d'image) , analyses des cold cases de façon artificielle à partir de textes, toutes provenant de l'open data publique.

Thanh-Long Huynh a présenté de nombreux cas d'usage. Ainsi dès 2013, la start up a su prédire les gagnants de l'émission the Voice, trois minutes avant le résultat officiel, pour

ensuite s'orienter vers la prévision des élections politiques, et monétiser ce type d'application. Sur 19 élections qui ont suivies, un score de 89 % de succès de prédiction a été obtenu sur des élections y compris très serrées. Ainsi, le Brexit a été prédit un mois à l'avance en analysant les inflexions de tendances. Idem pour l'élection de D. Trump, dont la prédiction a été publiée dans la presse française avec 4 jours d'avance. La start up reconnait deux erreurs à propos d'élections au Pakistan, dont il a été montré ensuite qu'elles avaient été faussées. Elle a également suivi les changements en Arabie Saoudite, en analysant, en temps réel, ce que pensait la population très connectée. QuantCube est intervenue en macro économie, en s'intéressant à l'emploi ; en suivant les secteurs qui recrutent plus ou moins, avec trois mois d'avance sur les organismes et les banques américaines. Dans le secteur maritime, elle est capable suivre les mouvements plus de 70 000 navires exportateurs de matières premières (énergie, minerais) autant d'indicateurs avancés pour les industries manufacturières. De même en utilisant les données de Rungis elle a pu suivre la variance inflationniste des fruits et légumes, ou, pour d'autres secteurs encore, estimer les différences entre chiffres officiels et non officiels, à partir d'images satellites en open data. De la même façon, dans des territoires où l'on ne dispose pas de chiffres, voir à partir des nouvelles constructions routières et de bâtiments, et en effectuant une analyse systématique des images, des taux de remplissage de l'hôtellerie corporate, pour ainsi prévoir es crises immobilières. Ses techniques de croisement de données sont également utilisées dans le secteur pétrolier, avec outre la surveillance des tankers, ou étudier le ratio d'ombre entre la hauteur des toits amovibles des cuves rapporté à leur taille. A un niveau plus granulaire, QuantCube est capable de mesurer le « sentiment index » des entreprises qui est considérer comme un actif immatériel. Ses études englobent aussi bien les legal cases des 98 cours de justice américaines, que l'étude de l'influence d'El Nino sur la sécheresse en Indonésie qui va impacter la production d'huile de palme, et son impact sur les pluies diluviennes de l'Amérique Latine. Connaissant la taille des navires, leurs caractéristiques, leurs incidents, elle s'est aperçue des corrélations entre la taille des navires et la taille des vagues, alors que les primes d'assurance sont les mêmes pour les navires qui circulent en janvier ou en juillet, alors que les vagues sont plus hautes en juillet. Cela a permis d'élaborer un nouveau modèle pricing d'assurance.

La participation de QuantCube à des challenges internationaux, lui ont permis d'appliquer sa technique pour différencier camions, auto, motos, et ainsi faire du comptage dans les entrepôts d'automobile, ou de compter les camions dans les centres logistiques, toujours en temps réel. D'autres challenges lui ont permis à partir d'image satellites de distinguer hôpitaux, commerces, terrains militaires, en obtenant 85 % d'accuracy, etc. Utilisant des réseaux de neurones, QuantCube a noué des collaborations avec des sociétés telles que HP, Amazon, OVH, l'Agence spatiale européenne, etc., qui ont beaucoup de puissance technologique mais ne savent que faire des données. Elle s'évertue à ce que les données restent dans les pays dont elles sont issues. Son modèle économique est basé sur le SaaS, en utilisant toujours des users profils anonymes. Elle annonce pouvoir proposer une cartographie active internationale.

QuantCube Technologies est une startup « pépite » FinTech, créée en 2013. Elle emploie 30 personnes (moyenne d'âge 25/26 ans), des data scientistes multilingues, et est spécialisée dans l'analyse de données big data.

3. Tendances et enjeux du Big Data en santé : focus sur la question des modèles économiques

Dr Christophe Richard (Santeos -Atos Worldline- Syntech Numérique)

Une donnée de santé est une donnée personnelle qui a une sensibilité particulière, et le RGPD en général et de santé en particulier, va unifier les choses. Le RGPD ne concerne pas les éléments d'information sur la santé physique ou mentale d'une personne, il concerne plutôt les prestations de services qui leur sont associées.

L'e-santé, la télémédecine, forment un écosystème très riche qui sous-tend un certain niveau d'exigence, de sécurité, de qualité, d'où une grande complexité, avec la question : « Comment générer de la valeur ». La loi de 2016 de modernisation du système de santé a modifié les règles de gestion des données et fournie des éléments pour « l'open data en santé ». L'accès aux données a été facilité, les délais raccourcis, avec la création d'un guichet unique : l'Institut national des données de santé (INDS), doté d'un comité d'éthique qui vérifie la conformité à des règles éthiques concernant la viabilité et la pertinence des demandes de données, avec avis de la Cnil. Si l'accès est facilité, les règles sont devenues plus strictes. Ainsi, les données ne doivent pas être réidentifiantes (concept jugé moins flou que l'anonymisation). L'INDS explique sur son site comment accéder aux données de santé. L'idéal, pour le patient comme pour le praticien, serait, à partir des données, de faire, bien sûr, du curatif mais surtout du prédictif et de la prévention, en temps réel, pour une meilleure prise en charge, en utilisant toute la puissance du big data.

L'enjeu n'est pas que celui du patient, mais qui aussi l'amélioration de la recherche, pour mieux cibler, optimiser les soins, mais aussi réduire les dépenses de santé publique, inventer de nouveaux processus de vigilance, disposer de remontées d'information sur les effets indésirables de médicaments. Autant que promesses qui tiennent Amazon, dans le domaine de la logistique. On parle beaucoup du big data en santé mais le plus souvent, il s'agit d'expérimentations de modèles et d'algorithmes, voire de communication et de marketing. Actuellement le médecin n'utilise pas le big data pour vous prendre en charge, et nombreux sont les cabinets qui ne disposent pas d'Internet même si les médecins l'utilisent chez eux à titre privé). Il existe des systèmes d'échange, disons de smart data basique, entre les médecins les établissements de santé (dossier médical personnel, groupements hospitaliers de territoire, territoires de soins numériques), et où l'on essaie un peu d'acculturer les personnels de santé au numérique. En réalité, on fait un peu d'épidémiologie, et il existe quelques données structurées qui circulent en vase clos pour produire du soin, voire du remboursement. Pourtant, il y a d'autres données qui pourraient entrer dans le cadre d'une prise en charge et d'une production de soin, celles du smartphone (géolocalisation, objets connectés pour le sport, mesure du rythme cardiaque, etc.), celles venant de l'entreprise du patient, celles concernant le mode de transport pour se rendre au travail, etc. ; on sait mesurer les calories d'un plat à partir d'une photo, on peut connaître le nombre de pas effectués, et déduire de tout cela, l'état de santé. Ces éco système de communiquent pas, c'est interdit. Un assureur identifie les gens en arrêt de travail, le stress ou les problèmes de dos, et pourrait alerter son actuaire qu'il est dans un processus pouvant le conduire à tomber malade. Impossible pour l'assureur de contacter le médecin, impossible de vendre une assurance en fonction du risque. Pourtant le potentiel existe.

Un groupe de travail a fait des recommandations, non reprises, sur l'accès aux données, la valorisation des connaissances, la monétisation des données, la création d'experts technique, juridique et médical, sur les tiers de confiance, sur les producteurs et les utilisateurs, sur le consentement personnel, le don et la session de données comme pour le don du sang ou d'organes, la portabilité, l'éthique médicale, sur un modèle qui vise l'efficacité du médicament, notion d'intermédiaire de confiance ou d'opérateur de services sécurisés. Au niveau de l'Europe sur l'année, on serait capable d'économiser 250 milliards d'euro, fourchette basse.

4. La carte vitale numérique

Dr Adnan El Bakri (InnovSanté)

Promouvoir le passeport numérique universel de santé, tel est l'objectif du docteur El Bakri. Interne en urologie, diplômé de chirurgie urologique, chercheur-spécialiste en prévention de l'évolution du cancer du rein, il a obtenu, en 2015, un master 2 de santé.

A partir de son expérience professionnelle, il explique que les cancers sont traités selon des standards internationaux, avec une mise sous surveillance pendant 5 ans. Ensuite, en absence de métastases, d'évolution ou de récidive, le protocole indique l'arrêt du traitement. Or, on s'est aperçu que, pour le cancer du rein, les patients revenaient 5, 10 ou 15 plus tard avec des métastases, alors qu'ils étaient considérés comme guéris, et donc hors surveillance. D'où la question : « A partir de la néphrectomie de la tumeur partielle ou totale, est-on capable de dire au patient : dans 5 ans, vous avec un risque élevé de métastases, aussi, notre surveillance ne doit pas être stoppée ?». Face à cette limite du système, et au besoin de disposer d'outils de médecine prédictive, l'idée du recours au big data s'est imposée. D'abord il a fallu imaginer un modèle, jusqu'alors inexistant. A partir d'une cohorte de 100 patients, il a, avec son équipe, utilisé une technologie de micro-spectroscopie infrarouge qui envoie de la lumière sur la tumeur et qui génère énormément de donnée. L'analyse des tumeurs prélevées chirurgicalement sur 100 patients, a abouti à 4 millions de données, qui ont ensuite été contextualisées, classées, en s'appuyant sur des outils de machine learning et de data mining, et obtenir

un algorithme d'IA. Le big data c'est le volume, la vélocité, la variabilité (provenance de plusieurs sources) et la véracité, combiné à la contextualisation sur laquelle IBM a buté. Outre l'auto-apprentissage du système visé par le Dr Adnan El Bakri, il s'agit aussi de pourvoir l'enrichir, afin d'obtenir une automatisation du diagnostic et de pouvoir parler de la médecine des 5P : préventive (arguments scientifiques), prédictive, personnalisée (car chacun réagit différemment), participative (enjeu primordial pour faire aboutir un parcours médical, l'émergence des patients experts, des e-patient) et précise (couplée par exemple à la génétique). L'enjeu est aussi de faire apparaître les éléments de terrain épidémiologique région, qui n'existent pas actuellement. Pour le Dr El Bakri, il n'est pas utopique de parler de blockchain. Il constate qu'il y a des silos d'informations, même dans le big data. On rajoute des applications et des systèmes, des silos non interconnectés. Dans le domaine financier ou monétaire, la blockchain est un moyen pour sortir des silos d'information, et cela pourrait être la révolution des systèmes de soins et de recherches, en matière de traçabilité des médicaments, pour les essais cliniques (80 % des publications seraient non reproductibles et on parle d'une quantité considérable d'essais non publié jusqu'à 5 ans après approbation de l'AMD). La blockchain propose une data base collaborative, sans passer par une autorité centralisée, avec des accès distincts, un horodatage, les smart contracts pour la monétisation, le remboursement, la e-prescription (e-ordonnance). Des e-prescription avec la blockchain pourraient éviter d'anonymiser l'ordonnance et permettre de partager l'ordonnance sur l'éco système sans dévoiler l'identité du patient. Ce serait également de contrer les GAFA et BATX qui sont actuellement les seuls capables de faire de l'IA, ayant 10 ans de recul de données (ex. : l'affaire FaceBook qui a utilisé des données de 50 millions de citoyens américains pour influencer les élections américaines). La blockchain c'est la révolution de la confiance, et on parle d'un « blockchained healthcare system » encore théorique pour l'instant, mais néanmoins en test. On estime que 50 % des essais cliniques ne sont pas signalés, et qu'il existe des problèmes de sécurité, des lacunes de connaissance, etc.

Actuellement, on en est encore loin, il existe des archives, mais certains services hospitaliers n'ont toujours pas d'ordinateurs. On pourrait faire du Big data, mais de façon limitée, il y a peu de données cliniques hospitalières, ni de suivi et de parcours de soins d'un patient. Par ailleurs, chacun utilise des logiciels différents, d'où une gestion archaïque, avec un retard français dans la réalité de la pratique médicale quotidienne. Il faut à 10 minutes pour récupérer l'information. Aux urgences il faut pouvoir appeler la famille, la pharmacie, le médecin traitant pour récupérer l'info d'un patient souffrant. Watson, le programme d'intelligence artificiel d'IBM, s'est heurté à des limites éthiques et scientifiques. IBM a acheté des données de santé aux établissements de santé, les a compilés, et faute de contextualisation de ces données, a dû reculer. IBM lance maintenant le cloud Watson sur les entreprises et a abandonné le programme Watson Health., alors qu'ici la loi avance avec la RGPD et la création de l'INDS.

Le Dr Adnan El Bakri, face à ces situations, s'est convaincu que le flux unique de circulation de données de santé c'est le patient. Il est le vecteur qui se déplace entre tous les intervenants. Aussi pourquoi ne pas lui proposer la technologie, le connecter à tous les intervenants autour de lui. Le patient au milieu, et tout le monde se connecte sur le patient. Il a ainsi lancé l'idée du Passeport numérique universel de santé, qui serait délivré au patient qui le détient, et qui serait connecté à une plateforme cloud sécurisée selon la réglementation ; le patient se rendant à une consultation autoriserait son

intervenant à se connecter, en 30 secondes. Le patient gardant a maitrise sur ces données de santé, que l'intervenant (médecin, infirmière, pharmacien, etc.) connaisse ou pas le système, qu'il soit proche ou distant. Une plateforme cloud a ainsi été installée, et un POC est également initié. La carte Passeport, baptisée PAssCare a été présentée à Las Vegas, tandis qu'une blockchain privée hybride dénommée ChainForHealth permet de rechercher gratuitement l'empreinte de chaque personne. Comme pour les crypto-monnaies, chaque patient a en effet une empreinte/ Le patient est de fait anonyme, tout en permettant d'accéder à ses données, à ses données de santé, à condition que toutes soient sur le registre blockchain.

Outre les soins et leurs suivis, et la recherche de l'historique du patient, la start up a ainsi pu réaliser et commercialiser une étude sur le suivi de la connexion de 3 000 patients à leurs médecins, qui va être publiée bientôt. Afin d'éprouver sa faisabilité, le Pass Care est expérimenté auprès d'un groupement de 150 pharmacies en France quelque soient leurs systèmes de gestion informatiques. L'utilisation d'appli va au contraire les lier entre eux, et à leurs 500 000 patients, d'ici l'année prochaine. Et les données de ces derniers seront immédiatement partagées via la blockchain qui crée l'interopérabilité. Avec le Pass Care, le patient garde le pouvoir. Avec la blockchain, on passe de l'information centralisée du passé, à l'information partagée, pour entrer maintenant dans l'ère de de l'information distribuée. Dans 10 ou 15 ans, avec le passeport de santé universel, il sera possible de parler d'open data en santé, estime le Dr Adnan El Bakri. Il permettra également de favoriser l'accès aux soins aux pays en voie de développement qui n'ont pas d'organisation de leurs systèmes de soins. Le modèle économique d'InnovCare s'appuie sur le modèle de Jean Tirole, prix Nobel 2014, qui a décrit le modèle des e-plateforme. Le Dr Adnan El Bakri explique que l'ère des applications est terminée que les applications santé d'Apple ne marchent pas, et des dizaines de millions d'euro ont été vainement dépensées parce que ce modèle est centralisé. Elles créent des silos d'informations supplémentaires, entraînent des problématiques de stockage dans les smartphones ; une étude a montré que les appli e-santé sont désinstallées au bout de 15 jours. Le cloud permet, en revanche, permet l'interopérabilité des systèmes informatiques. Actuellement, des start up font du profilage biologique à partir de photo sur Instagram, les données sont partout, (Twitter, Facebook, Linkedin, Instagram, YouTube), certains font du e-report en récupérant les effets secondaires des médicaments sur Twitter, considérant qu'il y a beaucoup plus d'informations sur les effets secondaire dans les réseaux, que chez les médecins etc.

La data est le nerf de la guerre, avec deux éléments clés, l'interopérabilité et la portabilité. Pour anticiper la révolution de l'open data, pour ne pas faire du diagnostic, la Carte Vitale 2.0 devrait finalement s'imposer. Innov Santé vient de nouer un partenariat avec Alcatel-Lucent qui équipe en télécoms un hôpital sur deux, en France, et qui possède déjà des outils de traitement de data et d'IA. Innov Santé apporte ses connaissances en Deep Learning. L'un des objectifs concerne les déserts médicaux, auxquelles seront proposés des consultations en visioconférence.

5.Aspect juridique

Marc-Antoine Ledieu (Bardehle-Pagenberg)

Tout ce qui tourne autour de l'open data, des contrats Saas côté prestataire face aux géants du net, le dark net et la cryptographie, au traitement et problématiques des données personnelles sont le quotidien de l'avocat Marc-Antoine Ledieu. Les litiges qui peuvent en découler font l'objet le plus souvent de conciliations que de jugements, en raison de la complexité grandissante des sujets qui deviennent hors de portée du pouvoir judiciaire actuel. Du coup, Marc-Antoine Ledieu ne plaide quasiment plus auprès des tribunaux, ce qui pourrait augurer de leur avenir !

Néanmoins, le vrai problème lorsque l'on parle de la donnée numérique, de l'open data, il s'agit de droits dérogatoires au droit des bases de données électroniques. Le fond du problème, pour un juriste, est que la data n'a pas de définition légale. On dispose seulement de connaissances techniques, juridiques, intuitives, de ce qu'est la data. Il n'y a pas de définition légale, ce qui est délicat pour écrire un contrat. Pour donner du sens dans ces matières techniques, Il faut partir de ces trois notions qui concernent la donnée de contenu, c'est-à-dire que l'on va dire, écrire, filmer, c'est du contenu. Ensuite, il peut y avoir du droit d'auteur ou pas de droit d'auteur, la donnée personnelle. La base de distinction pour comprendre le droit de la data et de l'open data, c'est la notion de contenu. Si l'on s'attache au contenu, celui-ci est protégé par le secret des correspondances, que ce soit de la conversation type GSM, 3G ou 4G, du contenu consulté par URL, de l'email, de la messagerie instantanée, il s'agit de data, protégées par le secret des correspondances, le secret des communications et télécommunications.

L'Union européenne parle de RGPD et de données personnelles, la suite qui arrive concerne l'e-privacy qui devait être voté fin 2017 mais qui va l'être en 2018, c'est une réglementation complète sur va porter sur le contenu des communications électroniques et des métadonnées des communications électroniques. Métadonnée signifie « donnée de/à propos de donnée »), c'est une donnée utilisée pour définir ou décrire une autre donnée papier ou électronique. Les métadonnées sont toutes les données numériques qui vont être générées par les systèmes qui permettent de faire les transferts de données de communication électronique. Ces systèmes transfèrent du contenu d'un terminal vers un autre terminal, vers un serveur, tout cela génère de la métadonnée. Cette métadonnée reste de la data, et se pose la question de qui la gère, qui peut l'utiliser, qui en est propriétaire, qui peut la commercialiser et dans quel cas.
Avec la réglementation e-privacy il y aura un cadre européen sur ce sujet, pour les 27 états membres, sans les Britanniques. Tout cela pour arriver à cette notion très particulière, c'est une donnée, à laquelle on va donner une qualification particulière, de la donnée personnelle, de la donnée personnelle dans laquelle il y a de la donnée de santé, et quantité de catégorie et de sous-catégories, etc. Il y a un truc intéressant, en cherchant vraiment, le droit de droit de propriété sur la data n'existe pas, nulle part, ni aux Etats-Unis, au Japon, ni en Russie ou en Chine. Toutes les notions d'appropriation, de business, ou que l'on peut commercialiser librement n'existent pas. Il n'existe qu'un cadre, pour toute l'Union Européenne, qui fonctionne assez bien, celui de la directive n°96/9 du 11 mars 1996 sur les bases de données numériques. La question est comment cette directive peut être utilisée, quel type de data sont encadrées par les bases de données numériques. Pour comprendre ce cadre, quelle que soit le type de data mises dans une base de données, Il faut savoir ce qu'est une base de données, il y a deux bases de données, dans la directive de 96, il y a la base de données contenant qui concerne le régime logiciel, la notion de contenant, la structuration de la donnée, le sac. La partie contenant c'est du logiciel, c'est le régime légal du logiciel, avec tout, sauf le mot logiciel. La base de données contenant, le système de gestion de bases de données que l'on appelle le SGBD (système de gestion de la base de données), de la protection logicielle avec deux choses à retenir ce régime de protection du logiciel de gestion du logiciel de pilotage ne va concerner que la partie structuration, mais pas le contenu. La deuxième partie dit clairement, le droit d'auteur ne couvre pas le contenu, la data, qui est dans le « sac ».

La deuxième partie de cette directive de 1996 qui est compliquée porte sur le contenu de la base de données. Il peut y avoir d'autres législations qui s'appliquent : droit d'auteur (exemple du propriétaire d'un stock de MP3 légalement chargés), du secret d'affaires (nouvelle réglementation en attente), des données personnelles, des documents secret défense/confidentiels. Tout cela va être en plus, donc on va empiler les couches de protection sur la partie data. L'autre partie c'est la notion de contenu de la base de données, comment on va protéger la base de données elle-même, qui va être protégé, qu'est-ce que l'on a comme droits pour arriver à l'open data qui va être une sorte de dérogation à tout cet ensemble. Le critère légal offre une protection du contenu d'une base de données lorsque que l'on peut prouver un investissement substantiel sur l'obtention, la vérification, la présentation de sa base de données. Cela est valable pour une personne privée ou public, tant qu'elle investit d'une manière ou d'une autre dans le contenu de la base de données, avec un mix de critères de Bruxelles et français, du point de vue qualitatif et quantitatif, avec des investissements humain, matériel ou financier qualitativement et quantitativement substantiels. Les tribunaux, en cas de conflit, au cas

par cas, vont dire, là il y a investissement substantiel, là non. Cela vaut pour la base de données Météo-France, pour l'annuaire de France Télécom, pour le programme des courses du PMU qui a donné beaucoup de contentieux au départ parce que l'on ne savait pas comment le protéger. Il y a des jurisprudences importantes, sur ce sujet. On protège donc le continu de la base de données, mais au profit de qui, et le qui, est-ce le producteur, et ce producteur est-il défini comme celui qui va prendre l'initiative et le risque. Donc personne publique ou personne privée, si je décide d'investir dans du logiciel, du matériel, de l'achat de data, dans la récupération de data éventuellement privée. Si on récupère de la data qui est en open data et qu'on la traite et que l'on a du personnel et que l'on a investi du matériel, cette data va générer un droit à mon profit qui va être ce droit du producteur de la base de données. Même si l'on a récupéré des données de manière gratuite. Donc l'open data va juste être une dérogation à tout cela qui va permettre de récupérer de la data. Cette initiative et prise de risques, en droit européen cela s'appelle le fabricant du contenu de la base de données ou, en droit français, le producteur de la base de données. Encore faut-il déterminer tout ceci contractuellement pour savoir comment utiliser et réutiliser cette data, il faut voir si le contenu est protégeable, s'il y a bien un producteur, et on arrive sur deux droits. Le droit des bases de données qui n'est pas très connu, mais il est revenu à la mode avec les licences de base de données cartographiques pour faire les premiers services de guidage avec son téléphone, et ce droit des bases de données, qui est un monopole d'exploitation, c'est un titre de de droits d'auteur, on est considéré comme le créateur, l'auteur, donc on a un monopole d'exploitation sur la base de données, sur le contenu de la base de données ensemble, c'est la même chose, sauf que l'on a deux droits qui sont attribués en droit latin, le droit sui generis (certains droits de propriété intellectuelle spécifiques peuvent être considérés comme *sui generis)*, le droit d'extraction et le droit de réutilisations.

Dans l'hypothèse où l'on est le producteur de la base de données, on peut justifier de ses investissements ou en tout cas dans le contrat, on va jurer dans le contrat, donner des garanties contractuelles que l'on est bien le producteur que cela soit vrai ou pas, parce qu'il faut ensuite aller devant les tribunaux et attendre le contentieux pour savoir si c'est vrai ou pas, mais en tout cas contractuellement on peut l'écrire.
Et si on se considère comme producteur du contenu de la base de données, alors on a la capacité d'interdire ou d'autoriser le transfert, permanent ou temporaire, de tout ou partie du contenu de la base. Le deuxième droit, c'est le droit de réutilisation, soit le droit d'extraction de la base de données et de ce que l'on peut en faire. Il y a un critère qui est très difficile à manipuler : la mise à la disposition du public. Ce n'est pas très clair, le public n'est pas forcément le public de manière générale, le particulier consommateur, cela peut être un public professionnel, cela peut être une capacité de mettre à disposition, d'un public BtoB, d'un public professionnel pour retraitement big data, d'un sous-traitant, au sens de la loi Informatique et Liberté. On a de la data dans une base de données, on autorise quiconque de piocher tout ou partie de son stock de données pour que ce dernier le réutilise à son profit. On est là au cœur de la mécanique de la protection du droit des bases de données. Avec un droit très particulier, un droit sui generis qui est un droit autonome qui confère une protection d'une durée de 15 ans après l'achèvement de la base de données, le droit d'auteur c'est 70 ans, là on est sur une durée de 15 ans après l'achèvement de la base de données, ce qui ne veut rien dire. On considère qu'à partir que la base de données est constituée, on part pour 15 ans C'est sujet à caution, parce que si l'on réinvestit de manière substantielle, quantitative et

qualitative, dans le contenu de la base, la protection est relancée pour 15 ans. C'est le premier droit où l'on a des durées de protection qui sont quasiment perpétuels, qui sont glissants. Si l'on met en place un process d'investissement pour mise à jour de sa base de données.

Une personne morale de droit public ou une institution de droit public va produire de la data dans le cadre de sa mission de service publique, on est là au cœur de l'activité, et cette capacité à réutiliser cette data pour autre chose. Là on vraiment au cœur de l'open data. On a un gisement de données, est-ce qu'il est possible pour ces institutions, ont-elles le droit, de les réutiliser ? On se heurte de suite à des principes et des exceptions.

Dans la loi Cada (la **loi** n° 78-753 du 17 juillet 1978, la Commission d'accès aux documents administratifs), « ne sont pas considérées comme des informations publiques…, les informations…, sauf quand cela est diffusé à titre purement public, élaborées ou détenus par les services publics, dans l'exercice d'une mission de service publique, à caractère industriel ou commercial. Si l'organisme public collecte de la data dans le cadre de sa mission industrielle ou commerciale, cela va revenir à lui et on va retomber, en fait, dans le droit commun de la base de données. On n'est donc plus en open data, parce qu'il y a la protection du droit d'auteur, ou alors si c'est bien un service public à caractère industriel et commercial, on repart dans le droit commun de la base de données, et cela ne sera plus du caractère public open data. On voit les difficultés. On a une loi de 78 (78-753) modifiée en 20O5 (2005-60), on a la loi PSI (2003-98) concernant la réutilisation des informations du secteur public, remodifiée en 2013 avec la directive n°2013/37) qui concerne l'extension de la réutilisation des documents aux bibliothèques. Les lois pour encadrer l'open data sont nombreuses, ordonnances, au point qu'il est difficile de s'y retrouve, et qui font que l'open data ne fonctionne pas jusqu'à maintenant, que cela ne risque de ne pas fonctionner.

La grande nouveauté c'est que l'on a la loi République Numérique d'octobre 2016, un code des relations entre le public et l'administration, qui impacte différentes lois. L'open data version française, c'est l'obligation de publier des documents administratifs (retour à la loi de 78). Ce qu'il faut en retenir, sont des documents administratifs ce qui est produit dans le cadre de la mission de service public : dossiers, rapports, études, comptes rendus, procès-verbaux, etc., et surtout statistiques, prévisions et codes sources (c'est du logiciel donc du droit d'auteur, donc pas de l'open data), des thèmes qui concernent la data. Le grand apport de cette loi, c'est l'obligation de communication spontanée de l'administration de publier un grand nombre de documents administratifs. C'est un renversement du système et de son esprit l'administration est obligé d'ouvrir des documents, puisque ce n'est plus au citoyen, à l'entreprise ou à un service de public, de demander l'autorisation pour accéder aux data. Il y a des exceptions pour les collectivités territoriales, de moins de 3 500 habitants, et une seconde exception pour les personnes morales de droit public dont le nombre de salariés est inférieur à un seuil, soit 50 personnes dans une entité, en équivalent temps plein. Dans la loi on voit base de données qui ne font pas l'objet d'une diffusion publique par ailleurs, soit c'est diffusé publiquement, auquel cas on y accède plus ou moins, ou cette notion de pas de diffusion publique, mais c'est bien une base de données mise à a jour de façon régulière que les administrations produisent ou reçoivent, et on a le même système sur la data. C'est un peu limité mais on a des : intérêts économiques, sociaux, sanitaires, environnementaux,

ce qui constitue déjà un stock de data important que l'administration est tenue de mettre à disposition, pour réutilisation, donc à d'autre fins que sa propre collecte.

Le drame de la législation à la française c'est qu'il y a les grands principes et quantité d'exceptions. On a là tout le problème qui bloque le système de l'open data, tel qu'il est conçu. Les exceptions ne sont communicables qu'à l'intéressé, donc on ne peut pas aller dans la base de données pour la réutiliser. Si l'on va faire son marché dans une base de données de l'administration, on se heurte à la protection de la vie privée, au secret en matière commerciale et industriel qui comprend le secret des procédés (droit des brevets), les informations économiques et financières, les stratégies commerciales ou industrielles. On peut mettre tout ce que l'on veut, là. Ce qui ressort de cette exception, si elle est de mauvaise foi, l'administration peut dire que systématiquement, il y a du secret de procédé, d'informations économiques et financières, de stratégies économiques et financières, industrielles et commerciales, etc. On peut faire dire ce que l'on veut à de telles notions. L'administration peut dire ce qu'elle veut et s'opposer à tout et n'importe quoi. C'est légitime, l'administration ne sait pas comment vont être utilisées, voire manipulées, ses data. On ne sait pas très bien comment on peut manipuler tout ceci. On sent que techniquement on est en train d'évoluer dans toutes les législations qui touchent au numérique. On commence à voir dans le droit à la portabilité, la RGPD, dans les formats, formats standards ouverts, régimes aisément utilisables ou réexploitables, on commence à avoir cette notion technique. Là, on n'est plus sur la base de données contenant, mais sur la capacité à entrer et récupérer la data. On explique comment l'on doit procéder, et c'est à la fois des bonnes et mauvaises nouvelles, mais c'est loin d'être gagné pour l'open data, en juridique.

6. OPEN DATA, BIG DATA et RGPD : Mission impossible ?

Nathalie Puigserver (P3B Avocats)

Parmi les spécialités du cabinet P3B Avocats, il y a le droit des nouvelles technologies et des données : droit transversal et désormais incontournable. Un domaine où se mêlent les problématiques d'identification, de ré-identification des personnes, où les contraintes légales sont fortes car il s'agit de protéger les droits et libertés fondamentales des personnes.

Les lois et réglementations applicables à la protection des données personnelles s'imposent à tous les acteurs de l'open data qui entrent dans un monde de principes et d'exceptions, un univers complexe où se mêlent les mesures techniques et légales, où il est souvent question d'équilibre, de « mise en balance », entre Privacy by Design et le Privacy by Using, entre anonymisation, pseudonymisation et chiffrement...

Il faut trouver des solutions pour intégrer, combiner et tenir compte des obligations qu'impose le RGPD, en conservant un esprit d'ouverture par rapport à la réutilisation des données. Les données publiques comme les données privées sont concernées, dès lors qu'il est possible d'identifier directement ou indirectement, ou de ré identifier une personne au moyen de la captation massive des données. La problématique des données personnelles est importante pour les objets connectés et par rapport à la réutilisation des données captées, stockées, traitées. Toutes les exigences technico-légales peuvent

« décourager » les initiatives innovantes mis bout-à-bout. Il n'est pas rare d'entendre : « ce n'est pas gérable », « ça freine notre modèle économique ». Oui, c'est une certitude : Protéger des droits fondamentaux entraîne des obligations technico-légales contraignantes.

Au niveau des systèmes d'information, il faut être capable de « maîtriser » les flux de données : comment les données rentrent, comment elles sont utilisées et comment est-ce qu'elles partent. Il faut avoir une traçabilité permanente de ces flux de données, pour en assurer la confidentialité, ce qui vaut pour les objets connectés et le big data. Il faut intégrer toutes ces mesures techniques, structurelles permanentes et les faire contrôler par le responsable des traitements, l'exploitant et le concepteur des données. Il y a une multitude de mesures à mettre en œuvre, et les mesures techniques sont tout aussi importantes que les mesures légales. C'est la combinaison des deux, qui fait l'efficacité de la protection !

L'article 25 du RGPD a consacré la notion de Privacy by Design, pour intégrer la protection des données dès la conception des produits, des outils, des services numériques, des données personnelles, et également la protection des données par défaut. Deux alinéas de l'article 25 donnent des lignes à suivre mais selon des préceptes généraux. Ils concernent le responsable des traitements qui doit déterminer les moyens, et donner un certain nombre d'indications sur les finalités que doivent poursuivre ces mesures, comme les aspects sécuritaires, le fait d'en limiter l'accès, la quantité des données qui doit être connectées. On a, ainsi, des concepts posés par le RGPD, et qui doivent mis en œuvre pas le responsable des traitements. A ce stade on ne dit toujours pas comment.

Le RGPD comprend 99 articles et 170 considérants qui expliquent les articles. Le considérant 78 donne des exemples de mesures concernant le Privacy by Design, avec le principe d'anonymisation des données, de les rendre pseudonymes dès que c'est possible, ou permettre à la personne concernée de garder le contrôle sur ces données. Ce considérant explique le type de mesure que le responsable doit mettre en œuvre pour le Privacy by Desgin, et concerne les fabricants de produit, de services, les producteurs de logiciel. La conception d'une innovation, d'une plateforme de services en ligne, de marketing digital implique de mettre en place des procédés numériques pour effectuer de la collecte d'emails, des campagnes d'emailing, d'un outil CRM. Le champ est très vaste dès lors qu'il y a des données personnelles qui sont captées, stockées, traitées. Cela demande de développer des techniques, des outils et des algorithmes, qui permettent de faire que ces produits et services soient, avec une protection de la vie privée « all inclusive », c'est-à-dire par défaut, sans que l'utilisateur ait à faire ne manipulation pour protéger les données, par exemple : GPS désactivé par défaut.
Tout est à prévoir à l'avance, c'est compliqué pour le fabricant, pour les entreprises qui utilisent les données, il faut qu'elles respectent bien les engagements du RGPD. Cela implique d'aller auditer son éditeur de tel ou tel de logiciel, son prestataire cloud, chez qui sont stockées et hébergées les données. Cela responsabilise le concepteur qui doit s'assurer qu'il utilise un outil conforme, voire de changer de solution. Ce n'est pas anodin parce que les outils ne sont souvent pas d'origine européenne, mais américaine, lesquels ne sont pas nativement conçus pour être Privacy by Design.

L'application concrète du principe de Privacy by Design est encore assez floue. Il est demandé au responsable du traitement des données, au créateur d'objet connecté ou de base de données, de s'assurer du respect de principes, tels que celui de minimisation des données, ou de pseudonymisation des données. Comment faire pour que les mesures techniques intègrent ces concepts ? Il n'existe pas encore de solutions toutes faites, mais il y a des recherches en cours, des algorithmes en cours d'élaboration.
Si on prend le principe de minimisation, il est demandé de collecter uniquement le minium de données nécessaires à un traitement, de déterminer à avance ce besoin, alors que pour les objets connectés, il faut un maximum de données, et les stocker massivement dans le cloud ? Or, si on limite la captation des données cela ne va-t-il pas ruiner à néant l'intérêt même d'une innovation ? L'objet connecté sans des données cela n'a aucun intérêt.

Il faut trouver comment, sans brider ces aspects-là, respecter le principe de minimisation des données. C'est difficile d'intégrer des mesures de protection des données, alors que l'on ignore à l'avance tous les risques qui vont survenir. De même que limiter la captation des données serait contreproductif. Il faut essayer de coupler, selon les solutions qui émergent, entremêler les mesures techniques et légales. Il faut coupler le Privacy by Design avec des mesures de type Privacy by Using orientées sur l'utilisateur. Il faut responsabiliser l'utilisateur, ses données n'ont pas besoin d'être toutes stockées sur des serveurs externalisés, mais rester sur les appareils numériques avec lesquels elles sont collectées. L'utilisateur en garderait ainsi le contrôle et la gestion.

Le RGPD impose un haut niveau de protection des données privées, ce qui est rassurant, les usagers étant par ailleurs assez préoccupés de ce que l'on peut faire de leurs données. Les entreprises qui utilisent les données ne peuvent, également, plus se passer des dimensions Privacy friendly qui sera un avantage commercial, bénéficiant de la confiance des utilisateurs, et qui est un avantage compétitif sur le marché européen.

Deux initiatives intéressantes :

Mis en forme : Police :Couleur de police : Automatique

1/ A propos de Privacy by Design, la CNIL et Inria ont décerné le prix "protection de la vie privée" 2017 à une équipe de recherche européenne, à l'occasion de la 11e édition de la conférence internationale *Computers Privacy and Data Protection* (CPDP) pour leur article : « *Engineering privacy by design reloaded* ». https://www.inria.fr/actualite/actualites-inria/le-prix-cnil-inria-a-ete-remis.

2/Le club Urba-EA (www.urba-ea.org), partenaire du Cigref, qui aborde différents sujets liés à l'architecture d'entreprises (partage d'expériences des processus métiers, données et systèmes d'information) entend construire une plateforme « modèle » de conformité RGPD avec une architecture type. Elle serait capable de s'insérer dans n'importe quel système d'information existant, pour une mise en conformité de l'existant, en tenant compte des contraintes, en vérifiant la traçabilité des usagers qui se connectent.

Le problème le plus important de l'open data est la possibilité de ré-identifier les personnes. Plus on a de données plus on pourra ré-identifier, c'est le défi majeur des entités publiques et des entreprises : ouvrir les bases de données conformément au

RGPD. Le pacte conformité open data de la Cnil bientôt publié, pourrait aider à intégrer ces mesures.

On peut ré-identifier une personne malgré l'anonymisation. Il faut anonymiser différemment, moduler l'ouverture des données en fonction des risques, c'est un sujet en devenir, et il y a beaucoup de choses à inventer. En résumé, il ne faut pas voir le RGPD comme suscitant des charges en plus. Il faut au contraire le voir comme une opportunité. Ce n'est pas mission impossible, mais cela ne sera pas facile, on ne peut pas s'improviser dans le monde de la data et de la protection des données. Il est primordial de s'emparer très amont de ces questions, afin d'éviter les risques de perte d'argent et de temps, et de s'entourer d'experts.

7. La propriété intellectuelle dans le contexte de l'open data

Benjamin Jean (Inno3)

Juriste spécialisé en propriété intellectuelle et fondateur du cabinet Inno^3, Benjamin Jean est spécialisé en gestion de la propriété intellectuelle dans le cadre de modèles ouverts (Open Source, Open Data, Open Hardware, interopérabilité ou plus généralement Open Innovation et Open Access). Inno^3 est un facilitateur et catalyseur, au profit de la création, de l'innovation et des systèmes collaboratifs ouverts et partagés, convaincu de leurs vertus économiques et sociales.

Le sujet de l'open data n'est pas neutre. Cela fait une dizaine d'années que les collectivités publiques parlent d'open data, et une vingtaine d'années qu'un certain nombre de projets communautaires s'intéressent à la mutualisation et au partage des données. On connaît bien ce que l'on appelle l'open source, cette idée de collaboration dans le domaine du logiciel (Linux, par exemple), l'open data est également fondé sur la collaboration et le partage et qui a vu le jour avec Internet. Nombre d'acteurs et d'entreprises, et surtout d'individus, ont constitué entre eux des bases de données qui étaient le référentiel qu'ils utilisaient pour leurs usages personnels comme pour leurs usages professionnels. L'open data a ensuite connu une croissance et une visibilité de plus en plus importante avec un certain nombre de réformes législatives, au point de devenir un sujet central pour les précédents gouvernements.

Quelques aspects juridiques, pour faire un parallèle avec le domaine de l'open source. L'open source de manière pragmatique, c'est l'idée qu'un titulaire de droits, un contributeur, un auteur au terme de la loi, qui détient un monopole sur ce qu'il a créé, décide, plutôt que d'empêcher d'autres d'utiliser ce qu'il a créé, de les favoriser et les amener à collaborer entre eux. Dans le cas de l'open source, il émet une autorisation qui permet à n'importe qui d'utiliser le logiciel, de le modifier, de redistribuer le logiciel modifié. Ce même système se retrouve dans l'open data même si, ici, les droits ne sont pas des droits d'auteur, comme on peut le connaître dans le domaine du logiciel, mais des droits qui sont plus spécifiques comme le droit spécifique sur les bases de données. Dans ce contexte, l'open hardware est un mouvement également d'actualité, avec une logique de mutualisation autour de la fabrication de matériel, un sujet suivi en faveur de l'innovation dans les domaines de la mobilité et du spatial. Les acteurs, alors, se rassemblent pour construire, non pas des logiciels ou des bases de données, mais des lanceurs de fusées, des satellites, dans des logiques de brevets, de dessins et de modèles, où leur propriété intellectuelle est la base sur laquelle d'autres ensuite sont amenés à collaborer. On parle de collaboration et d'ouverture. L'open data est de plus en plus proche des préoccupations industrielles et économiques. A la base de la notion d'open data, la définition, la description de ce qui est le courant open data, il y a une définition, qui date un peu maintenant, mais qui a été rédigée par l'Open Knowledge Fundation (OKF). Elle définit comme open data ce qui, comme contenu ouvert, permet le libre accès aux données : liberté de redistribution, liberté de réutilisation, absence de restriction technique, attribution des auteurs et des contributeurs, intégrité des données, absence de discrimination entre les personnes et entre les groupes, absence de discrimination entre les domaines d'application, indépendance de la licence vis-à-vis des autres contrats (hhtp://opendatadefinition.org)

Tout le monde s'est mis d'accord à l'international, pour considérer qu'une base de données est en open data lorsqu'elle répond à ces différents critères. La licence est un moyen qui assure cette liberté d'usage, l'absence de discrimination quant aux différents utilisateurs ou aux champs d'application. Le cadre réglementaire, en France, qui est à la base de l'open data, est la loi de 78, la loi Informatique et liberté, qui traitait du contrôle à l'accès aux données. Les collectivités, l'administration au sens large étaient de plus en plus sollicitées pour avoir accès aux documents, qu'elles avaient en interne, notamment par des sociétés d'outre-Atlantique. Aussi, l'état a réagi, en estimant qu'il ne pouvait laisser libre court à des échanges bilatéraux entre, d'une part des acteurs économiques, et d'autre part, les collectivités, et administrations, une à une rencontrée. L'état a ainsi souhaiter donner un cadre permettant de pouvoir donner accès à tel ou tel type de données ou à tel type de documents administratifs. Il y a, donc, eu des ordonnances et différents lois (voir vidéo 6:02 partie 2), ainsi que des circulaires qui sont venus s'y rajouter. Ce cadre est venu construire le cadre de l'open data que l'on connaît aujourd'hui. La loi « Pour une république numérique », adoptée sous le gouvernement Hollande, est venue encore élargir le domaine de l'open data, en termes de périmètre d'application.

De plus en plus d'acteur sont concernés par ce régime de l'open data, et en premier lieu l'administration au sens large, avec des obligations, à la fois juridiques, techniques, afin de rendre accessibles certains documents, juridiques, en les associant à une licence, d'où des questions : quel type de licence et de quelles façons ? Sans oublier l'open data par défaut, par exemple, pour un certain type de données qui présentent un intérêt

économique. L'administration qui produit des données est désormais contrainte à mettre ses données à disposition, avec des qualités techniques et une qualité dans le flux de données qui sont ainsi fournies, contraintes qui n'existait pas auparavant, Avant le régime de l'open data indiquait de simplement remettre les données aux personnes le demandant. Désormais, les administrations doivent être proactives pour donner automatiquement leurs données, avec une qualité qui n'est pas des moindres. La loi « République numérique » intègre une jurisprudence de la Cada (Commission d'accès aux documents administratifs est une autorité administrative indépendante et consultative chargée de veiller à la liberté d'accès aux documents administratifs), qui considérait que les logiciels étaient des documents administratifs. Cette loi ordonne que tous les documents administratifs, qui sont produits et reçus dans le cadre d'une mission de services publique, d'une administration, sont par principe dans le régime de l'open data, et figurent également parmi ces documents le logiciel en fait partie. Ceci permet de demander à l'administration l'accès au logiciel qui aurait été produit et reçu dans le cadre du service public. Cette nouvelle perspective est importante sachant que la notion d'administration est excessivement large. Elle englobe l'état mais aussi toutes les collectivités territoriales, toutes les sociétés qui auraient une délégation de service public, soit un périmètre très large.

La DINSIC (Direction interministérielle du numérique et du système d'information et de communication de l'État -http://www.modernisation.gouv.fr/mots-cle/dinsic), travaille, en ce moment même, et un premier draft a été a validé, à définir le cadre dans lequel tous les agents de l'état peuvent et doivent contribuer au projet open source. Son objet est de fluidifier, automatiser un certain nombre d'accès à des documents qui sont des données, et aussi à des documents qui sont des logiciels.

A propos des licences, avant de parler de l'écosystème. Une licence d'open data est une licence qui répond à l'open définition de l'OKF. D'un point de vue juridique, la licence est un contrat ou une offre de contrat qu'un titulaire de droits de propriété intellectuelle met gracieusement à disposition d'un licencié qui ainsi acquiert la possibilité d'utiliser, de copier, de modifier, de distribuer librement la base de données en question. Une administration doit ouvrir ses données en les accompagnant d'une licence. Le site du gouvernement (data.gouv.fr//licence) propose la liste des licences qui peuvent être utilisées.

L'open data implique de réfléchir en termes de décentralisation. L'écosystème doit reposer sur un ensemble d'acteurs, en intégrant les pratiques issues des projets collaboratifs, telles que la transparence, l'inclusivité. Pour que des acteurs travaillent les uns avec les autres, toutes les ressources produites doivent être documentées, afin que quiconque puisse rejoindre l'effort commun. Pour un projet open data réunissant de 10 à 100 personnes, et devenant de plus en plus complexe au fil du temps, le risque, au bout de 5 ou 10 ans, existe que plus personne ne soit capable de rentrer dans le projet parce rien n'a été documenté. Si toute la connaissance repose sur quelques acteurs du projet, le projet est mort. Un projet open data collaboratif est dit collaboratif s'il réunit au moins une dizaine d'acteurs. Mais 10, c'est déjà insuffisant, il faut que d'autres puissent s'y greffer, rejoindre le mouvement, et l'y inclure et qu'il s'y inclue, afin d'entrer dans le « faire » très rapidement. Avoir des réflexions globales, se de donner des instruments, notamment juridiques, qui permettent à tous ces acteurs de faire ensemble, est une

chose, mais il faut expérimenter, se confronter au terrain, avoir des projets au fur et à mesure qui permettent de tester les données et les ressources.

Quelques exemples : la Fabrique des mobilités (projet porté par l'Ademe) réunit des acteurs des collectivités territoriales, des centres de recherche, des écoles, des sociétés, des PME, des individus, des acteurs de tous types pour produire de manière ouverte et collaborative toutes les ressources open data ou open hardware, qui leurs sont nécessaires pour leurs innovations communes ou spécifiques autour de la mobilité. Dans le domaine spatial, le projet Fédération réunit tous les acteurs ce secteur, sous l'égide du CNES. La Chine réfléchit et est active depuis longtemps dans le domaine de l'open data en réunissant tous les acteurs de différentes filières, un peu contraints et forcés certes mais tout de même. Leurs marchés sont énormes, et ils le font de manière concertée. Ces logiques systémiques leur offrant de travailler ensemble, de produire les ressources, permet à chacun de se spécialiser là où est la propre valeur ajoutée chacun.

Les données publiques sont désormais soumises au principe de l'open data sans ambiguïté. Tout dépend de la valorisation qui en résultera, ce qui inquiète tous les acteurs, en règle générale. Toute une série d'exceptions viennent, néanmoins, limiter la portée des craintes, selon le contexte dans lequel les données sont produites. Toutes les données ne doivent pas forcément être ouvertes, en raison des droits de propriété intellectuelle de tiers, du droit des données à caractère personnel, ou, encore, du secret industriel.

Un guide juridique pour faciliter l'ouverture des données de recherche (version 2), initié par un groupe travail animé par l'INRA, a été publié en avril dernier, https://inno3.fr/ressources/ouverture-des-donnees-de-recherche-guide-danalyse-du-cadre-juridique-en-france

8.L'open data, l'avenir du Big Data

Jean-Marc Lazard (OpenDataSoft)

Jean-Marc Lazard est le fondateur et président d'OpenDataSoft, plateforme française conçue pour l'ouverture et le partage de données. Dans 16 pays, plus de 130 villes (Paris, Bruxelles, Bristol, Potsdam …), administrations et entreprises (Saint-Gobain, Enedis, BPCE, Veolia, Suez …) l'utilisent pour la gouvernance et la valorisation de leur patrimoine de données.

La plateforme OpenDataSoft permet aux territoires et aux entreprises d'utiliser pleinement et de façon autonome le potentiel de leurs données pour en faire des leviers de développement économique et de pilotage de leur activité. Cette entreprise propose des softwares dédiés à l'ouverture des données dans 17 pays. Son objectif est la création de valeur à partir de la donnée. Ses clients : administrations, collectivités territoriales, régions, figurent parmi les plus dynamiques depuis bien avant la RGPS.

Un acteur de la ville ne dispose que d'un morceau des données. Dans une ville beaucoup de données concernant la voierie sont gérées par Veolia, SNVF, (http://www.zdnet.fr/actualites/smart-cities-et-si-les-bus-devenaient-plus-qu-un-moyen-de-transport-39868060.htm), mais aussi par des sociétés privées telles que Waze. Aussi comment partager et exploiter les données et améliorer la vie des citoyens dans la ville. OpenDataSoft fournit des outils numériques pour gérer un projet d'ouverture des données. Dans une ville de 20 000 habitants, le responsable a du mal à

comprendre la question des données. Outre que les coûts ont été divisés par 100, il faut expliquer à quoi leur servent les données, et qu'avec, on peut mieux gérer, faciliter l'accès à la donnée, de la nécessité de l'open data, de comment desiloter les data. La Commission européenne s'interroge pour faire circuler les données. Les lois donnent un cadre, avant c'était la jungle. Les plus grands laboratoires pharmaceutiques dont de l'open data, et ont plus de 10 ans de données collectées, et abandonne ces données aux meilleurs start up qu'ils recrutent pour les gérer et les exploiter. OpenDataSoft entend relever le challenge de faire consommer les données comme on consomme de l'électricité. Cela sera long avant d'interopérer et standardiser les réseaux de données, avec les questions de comment monter une gouvernance, donner la main aux métiers (statisticiens, ingénieurs, informaticiens, scientistes, etc.). La loi a des vertus pédagogiques de contraignante, et c'est encourageant. Il est possible de se consacrer maintenant sur les deux bouts de la chaine, pour accéder à plus de données, pour chasser la bonne donnée soit en open data soit auprès d'acteurs moins connus, tels que les offices de tourisme d'une ville. Les grands acteurs privés offrent des données sur étagère, la SNCF par exemple, mais on peut aussi regarder chez O'Net (nettoyage hôtelier) pour avoir des informations sur les nuitées. Il y a besoin croissant d'actuaires qui utilise des techniques mathématiques principalement de la théorie des probabilités et de la statistique, en complément des data scientistes. Les grands groupes veulent des données potentiellement ouvertes. Alsthom a eu besoin de 6 mois pour partager des données avec la SNCF. Maintenant, avec l'open data by design, on va commencer à créer de la valeur. Un tramway est un device qui s'intègre dans la ville, qui peut prendre des données et en donner. Avec les open data, on parle de transparence. C'est seulement depuis un an que Pas beaucoup en France ouvert sur du vrai l'open data. Seulement depuis un an, les principaux utilisateurs de l'open data l'utilisent en interne pour casser les silos de leurs entreprises, et sont ainsi dans une logique d'efficacité. La data de la valeur si on la croise avec d'autres, c'est un matériau qui faisait défaut avant et dont on peut s'emparer maintenant. Le jour où Lille a mis à disposition, sur son portail, la disponibilité des temps réel des bornes de vélos, il n'a fallu qu'une semaine pour que soit créée un widget sur Garmin (Le widget est disponible sur le store Garmin Connect IQ). Ce n'est pas le métier d'un agent de la ville d'imaginer, d'utiliser les données de la ville et de développer ce type d'application-là. Le sujet open data qui fait prendre conscience aux entreprises d'un potentiel négligé. Schneider Electric s'est doté d'un Internal data-hub open data interne, afin que ses filiales partagent les données. Un hub big data, possède des prises à données prêtes-à-l'emploi. Ce n'est pas compliqué mais si on n'est pas dans cette logique d'industrialisation et de démocratisation de l'accès à la donnée, ce que l'on pourra en faire ensuite sera assez marginal. Les assureurs sont sensibles à cela, et ils ont investi massivement quand avec une approche big data. Quand ils gagnent 0,01 % dans leur score d'octroi (Un modèle de score d'octroi se définit comme étant une fonction mesurant le risque de défaillance d'une contrepartie), ce sont des millions d'euro qui sont gagnés ou économisés, les concurrents font pareils Il s'agit d'une mise à niveau, mais qui va encore évoluer quand les données seront mises à disposition de manière extrêmement simple. Les data scientistes qui eux ne prennent pas les meilleures décisions au quotidien, ne comprennent pas vraiment ce qui se passe et qui aujourd'hui n'ont pas de légitimé, se demandent s'ils vont encore servir à quelque chose.

Au final, ouvrir ses données c'est faire partie du jeu, si on n'est pas présent avec ses contenus on n'existe pas. Il a fallu deux ans pour convaincre Total de mettre ses stations-

service en ligne, pensant avoir le pouvoir en les gardant. Un jour une start up pourrait strapper les données que les stations-service n'ont pas. Le fantasme de la vente des données, les vendre chers ou pas, la question est infime, en comparaison de l'amélioration de la qualité de service en BtoB ou BtoC. Le boulanger Paul, en France et en Espagne, il y a 6 ans, a compris que les données ne sont pas seulement leur patrimoine, mais des informations pour savoir qui achète, où, en ligne avec le stock. Le directeur des services de la Métropole lilloise utilise Waze, pour, en temps réel, concernant la voirie, et en plus c'est gratuit. Ces données deviennent des outils de connaissances pour analyser ce que vivent les gens, prendre des décisions politique, mesurer les effets des décisions, réaffirmer un choix politique ou l'infléchir, en temps réel. Améliorer la qualité de service avec l'IA, sans data ne permet pas grand-chose dans le quotidien. Quand OpenDataSoft a commencé à travailler avec TER-SNCF (notamment sur les panneaux d'affichage), les données étaient assez orientées pour l'exploitation. Mais il faut aussi les partager avec l'extérieur, avec des outils grands publics, ou ceux que possèdent les régions. OpenDataSoft joue un rôle de proxy, et propose une surcouche pour stocker des données et surtout les redistribuer. Faire du stockage n'est pas son métier, mais doit savoir en faire. Elle propose de croiser les données. Elle propose aussi des outils appliqués au mix énergétique. Avec les énergéticiens, on parle beaucoup de cas d'usage. Les secteurs sont variés mais les problématiques similaires, mais il faut rentrer dans les métiers, tout en restant agnostique. D'un pays à l'autre, il existe des différences culturelles. L'énergie est peu challengé en France. Il peut être intéressant de croiser les données d'un pays l'autre, d'où l'intérêt de partager les données.

9. Open science : les questions d'éthique de l'open data

Jean-Gabriel Ganascia (CNRS)

Jean-Gabriel Ganascia est chercheur en IA, spécialiste d'apprentissage symbolique. Il a produit deux thèses, une thèse de docteur-ingénieur sur les systèmes à base de connaissance puis une thèse d'État sur l'apprentissage machine. Il est professeur à l'Université Pierre-et-Marie-Curie (devenue faculté des sciences de Sorbonne université depuis 1er janvier 2018). Depuis un an et demi, il est président du Comité d'éthique du CNRS en raison de son intérêt pour les questions de données, et en tant qu'informaticien.

L'accès ouvert aux données et aux publications scientifiques semble, au premier abord, partir d'une irrécusable générosité. Toutefois, un examen approfondi montre que, selon les disciplines et les contextes, il prend des formes extrêmement variées. De plus, le partage résultant de cette ouverture n'est pas toujours équitable ce qui pose parfois des problèmes d'ordre éthique. Ce sont toutes ces questions qui feront l'objet d'une présentation et d'un débat.

Tous les 7 ans, il y a une révision des lois de bioéthique sur la vie et la santé (dernière en 2011). Pour cette année nous avons été impliqués, en tant que Comité d'éthique du CNRS, dans des réflexions ; les questions actuelles portent beaucoup sur l'informatique,

les données et l'intelligence artificielle. Il y a d'autres questions comme tout ce qui a trait à la fin de vie qui sont assez bien balisées, mais tout n'est pas réglé, et les questions de GPA et de PMA laissent perplexes le corps médical. Le Conseil de l'ordre des médecins a fait un rapport sur les questions éthiques suscitées par les données et les algorithmes. Sans entrer dans le détail du RGPD, une question centrale sur le partage des données. Cela mérite réflexion, car le partage dans le monde numérique n'a rien à voir avec le partage dans le monde matériel.

Saint Martin est le saint patron du partage. Soldat de l'armée romaine en garnison à Arras, il a vu un pauvre hère qui n'avait pas de manteau et qui avait très froid. Par commisération pour sa souffrance, il lui donna la moitié de son manteau, et garda l'autre. Les autres soldats se moquèrent de lui parce son manteau était devenu trop court et qu'il avait un peu froid. Dans le monde matériel que nous avons connu le partage nous coûtait. Grande différence avec le monde numérique, car là, si l'on partage quelque chose, rien n'est perdu. On duplique ; cela ne coûte rien et on conserve l'intégralité des données que l'on a partagées. Ainsi, telle ville veut conserver la souveraineté sur ses données, et conserver ses données mais en même temps elle peut très bien les dupliquer et en faire profiter d'autres parce qu'elle ne perd rien. Dans ce contexte pourquoi ne pas partager ? C'est gagnant-gagnant, tout le monde est bénéficiaire. Mais ceci demande à être discuté. Si tout le monde est bénéficiaire cela peut poser un certain nombre de problèmes insoupçonnés jusque-là.
On conçoit que des juristes qui n'ont pas forcément une compréhension complète des enjeux technologiques aient pu s'enthousiasmer, imaginer de nouvelles solutions de partage, avec à la clé tout un tas de licence, etc., et ainsi promouvoir l'open data, etc. L'axiome était, et est encoure souvent, que tout le monde va gagner à partager les données.

Avant le numérique, l'accès à la connaissance était limité. Un livre a un coût et il faut se déplacer ou le déplacer. Dupliquer un livre dans une bibliothèque était d'un coût loin d'être négligeable. Aujourd'hui nous dormons tous dans une bibliothèque. A n'importe quelle heure du jour ou de la nuit, on a tous librement accès à pratiquement tous les classiques. Mais il y a deux questions qui se posent : toutes ces données dont on sait qu'elles ont une potentialité forte, vont permettre de construire des connaissances qui vont être utiles. On dit souvent que les données sont le pétrole de l'avenir. Mais jusque à quel point cela est vrai et qu'est-ce que cela signifie ? Les données sont à la fois extrêmement utiles, elles peuvent rapporter beaucoup, mais cela dépend de ce que l'on en fait. Une donnée seule n'a aucun intérêt.

Il y a un an une conférence à New York a montré que ce sont les GAFA qui vont exploiter les données. Ce que Google veut, c'est exploiter les demandes de requête sur son moteur de recherches, sachant qu'ils ont 20 000 milliards de requêtes par jour. Il y a aussi les photos, 9 milliards de photo sont partagées par jour sur les réseaux sociaux. Les GAFA entendent tirer avantage de ces données. C'est le pétrole de l'avenir pour autant que l'on donne un sens aux données. Pour être capable de leur donner sens, il faut accéder à ces immenses masses de données. Si l'on possède qu'une petite partie des données, on ne sera pas capable d'en tirer parti convenablement. Différentes solutions se dessinent alors : les très grands acteurs qui ont toutes les données, vont en tirer parti, les autres non. Autrement dit, le partage n'est pas équitable. Aux acteurs publics ou privés qui ont une toute petite partie des données, les grands acteurs vont demander d'être généreux

et de donner leurs données en les mettant à disposition de tout le monde. Ils vont donc les mettre à disposition. Ils n'en profiteront pas vraiment, tandis que les grands acteurs vont les utiliser et en tirer totalement parti. C'est, répétons-le, un partage devenu inéquitable. Chacun dit que le partage c'est formidable, en réalité ce partage devient vite dissymétrique. D'un côté, les grands acteurs sont capables de tout utiliser et de tirer bénéfice du partage et, d'un autre côté, les autres sont asservis par un partage qui va les engager, et dont ils ne tireront aucun fruit. Jusqu'à présent on avait peu de certitude sur ce sujet-là. D'un côté, les pouvoirs publics et les organismes de recherche étaient ambivalents. Certains disaient il faut être moderne et généreux, s'ouvrir au monde, promouvoir le partage (Charte de Budapest). D'autres s'alarmaient et disaient que cela pouvait devenir dangereux.

On peut se demander quelles sont les limites du partage dans le champ scientifique ? Dans le domaine de la biologie et dans certaines communautés de l'astrophysique le partage a été vite admis. Chez les astrophysiciens, les enjeux économiques sont faibles. Aussi, les gens se sont mis à tout partager, n'ayant pas d'enjeux industriels majeurs. Ils ont fait des choses extraordinaires ; ils ont accumulé beaucoup de données, comme les astronomes dont les radiotélescopes fournissent des masses de données colossales. Le catalogue des livres et imprimés de la Bibliothèque nationale contient 14 millions de volumes, ce qui pèse 14 téraoctets, ce n'est pas énorme. Avec des disques de 2 téraoctets, il suffit de 7 disques pour stocker la BNF. Le web, en 2015, pesait 7 Zeta octet, soit un demi-milliard de BNF. Twitter avec ses 7 téraoctets par jour, c'est une ½ BNF. Dans les années 90, les communautés ont commencé à avoir des grandes bases de données, qu'elles voulaient partager sur le web et mettre en accès libre. Et il y a eu tout un tas de déclaration (Bermuda Principles 1996, déclaration de Berlin en 2003 et 2005, CNRS 2014). Un grand nombre de communautés scientifiques sont très ouvertes à l'open data, et puis d'autres le sont moins pour des raisons particulières. Dans le domaine médical, les médecins hospitaliers n'avaient pas trop d'opposition à l'ouverture des données, encore que ces données ne leur appartenaient pas vraiment. En médecine se posent à la fois des problèmes patrimoniaux et des problèmes de protection de l'intimité des individus. Les industries pharmaceutiques ont été et sont toujours extrêmement réticentes à ouvrir leurs données pour des raisons de concurrence très rudes et, aussi parce qu'elles ont peur que l'on montre, à leur insu, que leurs médicaments seraient inférieurs en qualité à d'autres. Ils sont très réticents, et en même temps il y a une pression du grand public pour que l'on ouvre les données et en particulier les données d'autorisation de mise sur le marché des médicaments. Ce sont des questions qui sont très compliquées. Les objectifs de l'open data sont doubles, mettre la science à disposition de tous et permette à tous de valider les données et les résultats déclarés. Ainsi, la science est financée sur fonds publics, aussi il n'y a pas de raison que ses données ne soient pas mises à la disposition à l'ensemble de la société, même avec de petites restrictions. Face aux infractions scientifiques, aux questions de fraude, en mettant les données de la recherche publiques à disposition de tous, des acteurs externes peuvent valider les hypothèses qui ont été faites à partir des données. L'open data pourrait aussi empêcher la fraude. L'idée, est donc d'avoir des articles, des publications et données qui soient à disposition de tous. La nature des données scientifique est donc très variée, avec au moins deux types de limites à leur diffusion : les données à caractère personnel et les données à caractère sensible. En médecine, pour des raisons scientifiques, l'accès souhaité à toutes les données permettrait à tout le monde de les utiliser et de les interpréter. Il y a des données que l'on ne pourra jamais

anonymiser. Anonymiser, c'est rompre le lien entre la donnée et la personne. Anonymiser étymologiquement c'est enlever le nom, et si on enlève le nom, il n'y a plus de lien avec la personne. Apparemment, cela paraît simple. Mais, c'est trompeur. Si l'on se contente d'enlever le nom, le lien entre la donnée et la personne peut être reconstitué par recoupements. Prenons deux bases de données. L'une est une liste de votants qui comprend le nom de chacun, son adresse, sa date de naissance, son sexe, éventuellement son affiliation à un parti, etc., tout ceci n'est pas sensible, sauf s'il s'agit d'un militant d'un parti. Si on croise cette liste avec une base de données médicales du votant ici nommé patient, n'y figurent aucun nom, pour le protéger, mais ses ordonnances, ses visites chez le médecin, les médicaments qui lui ont été prescrits, sa date de naissance, son adresse ou au moins le code postal pour savoir dans quel environnement le patient vit, son sexe, etc. En croisant ces deux bases de données, il est possible d'identifier, avec la date de naissance, le code postal et le sexe, le nom de chaque individu de façon non ambiguë, mais aussi de récupérer chaque base de données.

Face à ces problématiques, des laboratoires de recherche, en particulier à l'Inria et à l'X, travaillent sur les questions d'anonymisation. Pour anonymiser les données, on leur rajoute du bruit. Toutefois, il est nécessaire de savoir à quoi servent les masses de données ; si on sait que l'on utilise les données dans tel but, on peut mettre du bruit sur les éléments qui ne servent pas à ce but, cela ne change rien. Mais le principe des masses de données, ou de ce que l'on appelle les *big data*, consiste à amasser tout un tas de données sans savoir à quoi elles vont servir. On s'aperçoit alors qu'il est très difficile de les rendre anonymes. Cela pose des questions sur le RGPD. La Cnil, par exemple, définit deux principes : le principe de finalité des données et le principe de proportionnalité. Le principe de finalité consiste à définir ce à quoi vont servir les données amassées pour constituer un fichier. Or, c'est orthogonal avec l'idée du big data.

Bibliographie

Les enjeux éthiques du partage de données scientifiques, rapport Comets 2015 (site http://www.cnrs.fr/comets/)
Une science ouverte dans une république numérique (Guide stratégique d'application, CNRS, DIST)
Déclaration internationale sur le libre accès de de Budapest, le 14 février 2002, connue sous le nom de sigle BOAI (https://wwwbudapestopenaccessinitiative.org)
Déclaration de Berlin sur le libre accès à la connaissance dans les sciences, 2003 puis 2005 (https://openaccess.mpg.de/68042/BerlinDeclaration_wsis_fr.pdf

10. L'open data dans le domaine public

Reynald Chapuis (Pôle Emploi)

Pôle Emploi qui a une triple mission : inscrire, indemniser, accompagner les actifs et les recruteurs est tenaillé tout comme les acteurs marchands par les Gafa qui veulent s'engager dans le domaine de l'emploi (diffusion d'offres, offres aux recruteurs, simulateurs, etc.. Cela l'amène donc à adopter une politique numérique particulièrement volontariste et tout azimut pour offrir toujours plus de services aux citoyens mais aussi vers les développeurs grâce aux API et aux licences qui permettent d'exploiter ses données publics ou certains services tout en veillant à la protection des données personnelles et commerciales.

Ainsi Pôle Emploi propose l'emploi-store.fr, une place de marché large et performante qui rassemble 320 services digitaux (dont 52 de Pôle Emploi) et un million de visites par mois et l'emploi-store-dev.fr qui met à disposition sous licences ses données et services auprès d'une communauté de près de 4000 développeurs. Mais Pôle emploi accompagne aussi dans des programmes internes (intrapreneuriat) ou externe (challenges) des start-ups en leur proposant des dispositifs d'incubation, lance des démonstrateurs avec des jeunes pousses, et ainsi favorise la création de son écosystème dans le domaine de la RH. Reynald Chapuis explique comment s'effectue l'entrée dans l'ère numérique de Pôle Emploi dont la stratégie est centrée sur les usages et donc les services. Un organisme qui a beaucoup de données, un champ de contraintes très fort et une mission de service public qui est au cœur de la préoccupation des gens. Un organisme qui doit aussi être

capable d'exposer ses données de manière réglementaire. L'open data et l'open innovation sont liées, si l'on veut développer des usages et générer ses propres données de manière à répondre au mieux aux besoins de ses publics.

La question de la donnée au sein d'un organisme public est souvent la première des questions. Les acteurs publics sont des tiers de confiance qui collectent massivement de la donnée de qualité (conformité), qui est mise à jour régulièrement, qui est structurée pour échanger entre acteurs publics. La première étape toutefois est avant tout une mise au clair des représentations de chacun sur les afin de déterminer celles qui ont de la valeur et celles qui sont davantage des commodités. Pour le personnel de Pôle Emploi, la donnée « offre d'emploi » a longtemps été considérée comme étant à forte valeur ajoutée et donc à protéger ; là résidait historiquement la valeur. Assez vite, il est apparu que l'offre d'emploi pouvait être diffusée n'importe où, depuis les petites annonces du bon coin jusqu'à Google ou encore via des agrégateurs. Dès lors, le choix « offensif » de partager des données « offre d'emploi » en API sous la protection d'une licence permet à la fois de coopérer et de peser sur le marché tout en garantissant un cadre juridique de protection contre du crawling sauvage (aspiration des offres sans mises à jour des données) ou l'historisation des offres qui présente un risque en matière de secret commercial car susceptible de faire de l'intelligence économique (lancement de nouveaux produits, etc.). En ce qui concerne les données personnelles, la protection en lien avec le RGPD est essentielle pour remplir ses missions et conserver la relation de confiance nécessaire aux échanges entre ses conseillers et ses publics.

Par exemple, le parcours professionnel, le CV comporte des données déterminantes mais extrêmement précieuses (coordonnées, mail, compétences, entreprises, postes occupés, loisirs etc.). Il est à noter par ailleurs que l'anonymisation d'un CV est impossible compte tenu de la singularité des parcours professionnel. Alors comment concilier partage de la donnée dans l'intérêt du retour à l'emploi et protection ? Il a été décidé de lancer une API de portabilité des données (Pole emploi Connect) permettant, sur la base du consentement éclairé de l'utilisateur, de partager tout autant son statut de demandeur d'emploi, que ses compétences, son montant d'indemnisation afin que d'autres acteurs publics ou privés puissent prolonger ou approfondir l'expérience de l'utilisateur au-delà de l'accompagnement par Pôle Emploi. Par exemple, liquider une assurance « perte d'emploi » dans le cas d'un crédit immobilier, accéder à des prestations attachées à son statut de demandeur d'emploi, partager ses compétences ou son cv pour candidater directement sur une offre ailleurs que sur le site de pôle-emploi, etc.
Pôle emploi possède également de très grande quantité de données agrégées autour du marché du travail (enquête Besoins de main d'œuvre, statistiques sur le marché du travail, référentiel des métiers et des compétences etc.) qu'elle expose sur l'emploi-store-dev.fr également.

Pôle Emploi s'est très vite intéressé à créer de nouveaux services à partir de ses propres données, en développant un programme d'intrapreneuriat…en proposant à ses conseillers par le biais d'un challenge de répondre à leurs problèmes quotidiens. Après une sélection sur le profil et la personnalité du conseiller, celui-ci est coaché et accompagné d'autres compétences (DATAscientist, frontdev etc.) pour prototyper dans un temps court un service répondant aux enjeux concernés et s'alimentant de ses données. Un exemple ? Le marché caché du travail ! Comment accepter de voir régulièrement des déclarations préalables à l'embauche (DPAE) et ne pas ou rarement

pouvoir proposer des offres aux demandeurs d'emploi. « Labonneboite » est née de ce constat. C'est un service (https://labonneboite.pole-emploi.fr) qui permet en saisissant une recherche géographique et son métier de découvrir quelles sont les entreprises susceptibles de vous recruter dans les 6 mois. Comment ? En modélisant un algorithme exploitant les pratiques de recrutement des entreprises grâce à la DPAE… « Labonneboite » est accessible en mode portable et web mais surtout en API et présent désormais autant sur www.pole-emploi.fr que sur l'emploi-store.fr ou les outils de travail des conseillers. Nous développons dès lors un véritable savoir-faire dans le smart data (la data dédiée à répondre à des cas d'usage) plutôt qu'au Big data.
Créer en permanence de nouveaux services grâce à la donnée est un défi permanent mais nécessaire à l'attraction des publics…car un service public sans public n'existe pas…et comment conserver le lien avec nos publics à l'heure des réseaux sociaux ?
Pôle emploi a lancé son réseau social interne en 2013. Son nom ? POLE. C'est un succès ! Il rassemble désormais 80% de ses salariés et près de 5000 collaborateurs chaque jour autour de 800 communautés. Désormais c'est vers l'extérieur que l'organisme se tourne pour créer et animer son propre réseau social d'accompagnement de ses publics et faciliter la mise en relation avec ses partenaires, les acteurs publics, ceux de l'insertion ou encore les entreprises. Ce réseau appelé « Sphère emploi » est en cours d'expérimentation et répond autant au sentiment d'isolement des demandeurs d'emploi qu'au besoin de mobiliser des compétences pour régler des freins périphériques à l'emploi : trouver un logement, accéder à des modes de transport, etc. Ce réseau social, géré par Pôle emploi, y compris sur les données, garantit la protection des informations des utilisateurs, face à une approche souvent prédatrice sur des réseaux marchands comme Linkedin ou Facebook.

Comme toutes les administrations publiques, Pôle Emploi est contraint d'ouvrir ses données qui sont engendrées par sa mission publique. Il existe toutefois une certaine forme de dissymétrie, plus ou moins organisée par les lois et règlements, entre cette obligation et les données des acteurs marchands qui conservent une liberté d'usage restreinte bien entendu par le RGPD. Il serait pourtant bon de s'interroger sur l'accès à certaines données pour la bonne réalisation de l'action publique au bénéfice des citoyens : par exemple, savoir dans le cas de la mission de contrôle qu'une personne recherche activement un emploi sur linkedin, permettrait à la fois au demandeur d'emploi de rendre compte facilement de son action et à Pôle emploi d'être plus efficace dans son accompagnement des personnes les plus fragiles.
La notion de « données d'intérêt général » introduite par la loi Lemaire (La loi Lemaire, adoptée le 28 septembre 2016, renforce la protection des données personnelles et anticipait le règlement européen 2016/679) n'est pas suffisamment définie à cette heure pour s'imposer aux acteurs marchands et participer ainsi à un « commun ».

Pôle Emploi dispose également d'un programme d'incubation (la Fabrique) qui aide les entreprises et startup à se développer dans le domaine RH. Ce faisant, et en se développant, ces entreprises et startup sont susceptibles de créer des emplois et donc offrir des opportunités pour placer des demandeurs sur le marché du travail. Faire de l'innovation digitale permet donc de répondre indirectement à nos missions.

Pour rappel chaque année, Pôle Emploi, place près de 3 millions de personnes sur le marché du travail.