CNIL.

air 2021

ENTRE PARTAGE ET PROTECTION : QUELLE ÉTHIQUE POUR L'OUVERTURE DES DONNÉES ?

air

AVENIRS INNOVATIONS RÉVOLUTIONS

Air est l'acronyme d'Avenirs, Innovations, Révolutions, trois mots-clés qui forment le nom que la CNIL a donné à la mission éthique qui lui a été confiée par la loi pour une République numérique de 2016. Ses objectifs : explorer les avenirs souhaitables, questionner les innovations qui façonnent notre temps et appréhender les révolutions en cours.

Programme de l'édition 2021

Le lundi 8 novembre 2021, de 14h à 18h

OUVERTURE

Marie-Laure Denis, *présidente de la CNIL*

METTRE LES DONNÉES AU SERVICE DE L'ACTION PUBLIQUE

Amélie de Montchalin, *ministre de la Transformation et de la Fonction publiques*

DIALOGUES CROISÉS : DONNÉES ET INTÉRÊT GÉNÉRAL

L'odyssée politique de l'ouverture des données en France

Axelle Lemaire, *directrice générale déléguée à la Stratégie, la Transformation et l'Innovation de la Croix-Rouge française*

Eric Bothorel, *député des Côtes-d'Armor*

Open data, une nouvelle morale pour le partage des données en Europe ?

Malte Beyer-Katzenberger, *DG Connect, Commission européenne*

Eric Salobir, *membre du Conseil national du numérique*

LA DONNÉE COMME COMMUN NUMÉRIQUE

Stéphane Gigandet, *fondateur d'Open food facts*

LE PARTAGE DES DONNÉES DE SANTÉ À L'ÉPREUVE DU COVID-19 ?

Nathalie Mesny, *présidente de l'association de patients Renaloo*

Julien Marchal, *co-directeur de l'innovation de l'Agence régionale de santé d'Île-de-France*

Martin Daniel, *co-fondateur de Covidliste*

POTENTIELS, RISQUES ET LIMITES DE L'EXPLOITATION DE LA DONNÉE

Anonymisation et pseudonymisation des données ?

Yves-Alexandre de Montjoye, *professeur associé au London Imperial College*

Intelligence artificielle, l'ouverture contre les biais

Anne Bouverot, *co-fondatrice et présidente de la Fondation Abeona*

Datajournalisme, une déontologie de la donnée

Pierre Romera, *CTO de l'International Consortium of Investigative Journalists*

L'ART DU PARTAGE

Caroline Goulard, *cofondatrice et présidente de Dataveyes*

CONCLUSION

Marie-Laure Denis, *présidente de la CNIL*

somm

1 *Comment mettre les données au service de l'intérêt général ?*

L'odyssée politique de l'ouverture des données en France 7

Open Data, une nouvelle morale pour le partage des données en Europe ? 8

3 *Potentiels, risques et limites de l'exploitation de la donnée*

Anonymisation et pseudonymisation des données ? 19

Intelligence artificielle, l'ouverture contre les biais 20

Datajournalisme, une déontologie de la donnée 20

naire

2 *Open data : savoir allier éthique et partage*

La donnée comme « commun numérique » 12

Le partage des données de santé à l'épreuve du COVID-19 13

4 *L'art du partage*

L'universalisation de la lecture et de la compréhension de la donnée 22


éxito

« La protection des données peut tout à fait s'articuler avec l'open data : tout est affaire d'équilibre. »

Marie-Laure Denis,
présidente de la CNIL

Pour sa deuxième édition de l'événement Avenirs, Innovations, Révolutions (air2021), la CNIL a choisi de s'intéresser à l'ouverture des données. Pourquoi ? Parce que la crise sanitaire a montré que leur mise à disposition en grand nombre n'était pas seulement un enjeu technologique, mais aussi un défi économique, scientifique, démocratique et politique. Économique, car la donnée est une source de valeur et un levier pour l'innovation. Scientifique, car elle constitue un vecteur de connaissance. Démocratique, car elle favorise la compréhension des enjeux de société par le grand public. Et politique, car elle permet de nourrir la confiance dans l'action publique.

Depuis quelques années, le cadre juridique autour de l'ouverture des données fait l'objet de réformes ambitieuses en France et en Europe. En 2015, les premiers textes publiés sur le sujet à l'échelle nationale répondent à des besoins sectoriels, comme l'illustrent la loi de transition énergétique pour la croissance verte ou encore la loi « Macron » sur la mobilité bancaire. En 2016, un nouveau cap est franchi avec la loi pour une République numérique, qui crée un régime global applicable aux données dites « d'intérêt général ». Deux ans plus tard, le rapport de Cédric Villani, député de l'Essonne, sur l'intelligence artificielle, souligne l'importance de la souveraineté numérique de la France. Fin 2020, le sujet revient sur la table avec un rapport sur la politique publique de la donnée, remis par Eric Bothorel, député des Côtes-d'Armor. En 2021, dans la lignée de tous ces travaux, la ministre de la Transformation et de la Fonction publiques, Amélie de Montchalin, présente sa feuille de route sur l'ouverture des données publiques, qui constitue la clé de voûte de la stratégie gouvernementale de modernisation de l'État.

Et l'Europe dans tout ça ? Depuis 2018, le règlement général sur la protection des données (RGPD) encadre la libre circulation des données personnelles au sein de l'Union européenne, en fixant un certain nombre de principes, pour garantir les droits fondamentaux des citoyens européens, leur droit à la vie privée et plus globalement la protection de leurs données. Deux textes renforceront bientôt l'approche européenne sur l'*open data* : le *Data Governance Act* (DGA) et le *Data Act*. Ils compléteront les règles du marché numérique et les outils disponibles en vue de faciliter l'accès aux données et leur réutilisation, notamment par le secteur public, tout en assurant le plein respect de nos valeurs.

La CNIL promeut un partage des données respectueux de la vie privée dès sa conception. La protection des données peut tout à fait s'articuler avec l'*open data* : tout est affaire d'équilibre. Nous devons créer un cadre précis, pour que chaque acteur sache ce qu'il peut faire et comment, sans ambiguïté. C'est dans cet esprit que la CNIL et la CADA[1] ont co-rédigé en 2019 le *Guide pratique de la publication en ligne des données publiques*, en partenariat avec les services d'Etalab[2]. C'est aussi ce qui a motivé la création d'un groupe de travail au sein de la CNIL, afin de clarifier la façon dont les textes doivent s'appliquer en matière de publication et de réutilisation des données, qu'elles soient publiques ou privées, publiées sur internet ou partagées. Ce travail sera piloté par Anne Debet, commissaire à la CNIL, avec l'aide de la CADA. Il sera également nourri des réflexions issues de notre colloque air2021, qui s'intéresse à deux impératifs aussi légitimes l'un que l'autre : entre partage et protection, quelle éthique pour l'ouverture des données ?

Pour revoir l'intégralité des échanges, rendez-vous sur cnil.fr #air2021

[1] Commission d'Accès aux Documents Administratifs.

[2] Etalab est un département de la direction interministérielle du numérique (DINUM), dont les missions et l'organisation sont fixées par le décret du 30 Octobre 2019. Il coordonne notamment la conception et la mise en œuvre de la stratégie de l'État dans le domaine de la donnée.

1

COMMENT METTRE LES DONNÉES AU SERVICE DE L'INTÉRÊT GÉNÉRAL ?

L'odyssée politique de l'ouverture des données en France

L'ouverture des données au nom de l'intérêt général soulève de nombreuses questions : qu'entend-on par « intérêt général » ? Comment déterminer celles qui en font partie ? Sur quels critères ? Éric Bothorel, député des Côtes-d'Armor, a remis en 2020 un rapport sur la politique publique de la donnée en France et revient sur cette notion centrale d'intérêt général.

Éric Bothorel

Député des Côtes-d'Armor depuis 2017, Éric Bothorel a remis en 2020 un rapport sur la politique publique de la donnée en France.

« En rendant public un grand nombre de données, on crée les conditions d'un débat très large qui empêche de se saisir d'éléments non publiés pour déformer la réalité. »

Quelles ont été les grandes étapes de l'ouverture des données en France ?

Derrière l'enjeu de libération de la donnée, il peut aussi y avoir un sujet polémique ou mal interprété. Mais nous avons tout à gagner à mettre à disposition la *data* ! Pourquoi ? Parce qu'en rendant public un grand nombre de données, on crée les conditions d'un débat très large, qui empêche de se saisir d'éléments non publiés pour déformer la réalité. Par ailleurs, nous l'avons vu avec la pandémie, il est important d'enrichir les données publiques avec les données privées. Superposer les données entre elles a permis aux puissances publiques de mieux comprendre l'impact du COVID-19 sur la population.

Comment définir les données d'intérêt général ?

Tout peut être qualifié d'intérêt général. Les enjeux sanitaires et environnementaux qui nous attendent vont mobiliser des jeux de données très différents : météo, automobile… On ne peut pas savoir à l'avance quels jeux de données seront indispensables.

Pour moi, permettre à la puissance publique de réquisitionner des données qu'elle qualifie « d'intérêt général », c'est prendre le risque de nier les réalités économiques de ceux qui produisent et collectent cette donnée. Il faut clarifier ce « droit à la réquisition » de la donnée, comme on l'a fait pour les biens et services.

Demain, nous aurons besoin de *hubs* sectoriels (agriculture, santé…) qui soient pensés avec de l'interopérabilité. On ne peut plus réfléchir en silos, ni de façon verticale, tout comme on ne peut plus ignorer ce qui se fait ailleurs. L'interopérabilité sera déterminante pour permettre de servir les besoins de la population.

Open Data, une nouvelle morale pour le partage des données en Europe ?

Dans les coulisses du Data Governance Act

Malte Beyer-Katzenberger

Malte Beyer-Katzenberger fait partie de la DG CONNECT au sein de la Commission européenne. Il a travaillé à l'élaboration du *Data Governance Act* (DGA).

« En cas de crise sanitaire ou de catastrophe naturelle, il faut établir un mécanisme qui permette d'obliger une entreprise privée à nous donner accès à certaines bases de données »

Comment est né le *Data Governance Act* (DGA) ?

Le *Data Governance Act* a été élaboré en 2020 en virtuel, par des équipes 100 % à distance. Il s'inscrit dans la nouvelle stratégie data de la Commission européenne, qui vise à ouvrir les données du secteur public, du secteur privé et les données personnelles. Pour permettre cette ouverture dans un climat de confiance, le DGA propose de mettre en place des structures intermédiaires, c'est-à-dire des acteurs comme Dawex, Cozy Cloud ou encore API-Agro, qui fluidifient les données sur une base volontaire, à des fins économiques. Mais il faut que quelqu'un organise cet écosystème. Le DGA impose donc une stricte séparation entre la fonction d'intermédiation et tout autre service aux données. L'objectif est de faire monter des acteurs très spécialisés, qui ne sont pas forcément issus de l'écosystème des *Big Tech*[3], pour donner la chance à de petites start-up européennes de s'imposer dans la cour des grands.

Le DGA encourage aussi les États membres à développer des « accès restreints », comme ce que fait l'Insee en proposant des services aux chercheurs qui leur permettent d'extraire des infos clés des données (sans jamais voir les données).

Où commence et où s'arrête l'altruisme des données ?

En tant qu'individu, je peux permettre à ma commune d'utiliser mes données Waze pour analyser mes déplacements dans la ville. C'est une démarche volontaire de ma part et un véritable altruisme des données.

Il faut également donner aux pouvoirs publics, quels qu'ils soient, en cas de crise sanitaire ou de catastrophe naturelle, un droit d'ordonner à une entreprise privée, un accès à certaines bases de données à des fins bien définies et de façon très encadrée. Pendant le confinement, la Commission européenne a par exemple demandé aux opérateurs télécoms de lui fournir des données agrégées des téléphones portables, pour voir si les gens se déplaçaient malgré les restrictions et quel effet cela avait sur la propagation du virus en fonction des mouvements enregistrés.

L'altruisme des données pose la question du don au sens large : pourquoi les gens le feraient-ils ? Peut-on concevoir une obligation d'altruisme ? Ou au contraire, doit-on y poser des limites, afin de laisser aux gens la liberté de ne pas être altruistes, voire la liberté d'être égoïstes ? Le DGA laisse une certaine souplesse aux États dans la mise en œuvre du partage des données à l'échelle nationale.

[3] Big Tech est un nom donné aux entreprises les plus importantes et dominantes de l'industrie des technologies de l'information aux États-Unis (Google (Alphabet), Apple, Facebook (Meta), Amazon et Microsoft).

L'altruisme des données est une question qui revient au fil des débats. Il s'agit d'un enjeu beaucoup plus large et philosophique qu'on ne le croit. Où commence cet altruisme ? Où s'arrête-t-il ? Malte Beyer-Katzenberger a participé à l'élaboration du *Data Governance Act* au sein de la Commission européenne et nous explique ce qu'est cet altruisme pour le législateur européen. Son intervention est suivie de l'analyse d'Éric Salobir, Président de la *Human Technology Foundation* et membre du Conseil national du numérique.

Quelles sont les limites de l'altruisme des données ?

Éric Salobir

Président de la Human Technology Foundation, Éric Salobir est aussi membre du Conseil national du numérique.

« Nous devons absolument créer les conditions de la confiance. L'opposé de l'altruisme, ce n'est pas tant l'égoïsme que la peur. »

Que pensez-vous de ces notions d'altruisme et d'égoïsme ?

Le danger est de les opposer l'une à l'autre. Si je partage mes données avec une plateforme en échange d'un service, ce que je fais n'est pas désintéressé. Ce n'est pas égoïste non plus. Ce partage se fait très largement de nos jours, dans un contexte de donnant-donnant. Alors comment se fait-il que des résistances se dressent dès qu'on propose à des particuliers de partager leurs données pour l'intérêt général ?

Les applications qui ont été développées pendant la crise sanitaire montrent bien l'extrême frilosité du grand public. Les gens se demandent ce qu'ils y gagnent : *« je paie déjà mes impôts, il faudrait en plus que je partage mes données ? »*. Si les données détenues par les entreprises et les particuliers représentent un réservoir immense qui pourrait être valorisé pour l'intérêt général, il y a néanmoins une peur que les données soient mésutilisées.

Les entreprises craignent qu'elles soient récupérées par la concurrence, tandis que les particuliers redoutent qu'elles soient utilisées à des fins de surveillance gouvernementale. Quand on parle de réquisition de données anonymisées, le grand public n'entend pas le mot « anonyme ». C'est pour cela que nous devons absolument créer les conditions de la confiance. L'opposé de l'altruisme, ce n'est pas tant l'égoïsme que la peur…

Cela rejoint le projet du DGA de création de tiers de confiance…

Tout à fait. Dans le cadre du RGPD, on accorde une immense importance au consentement. Pourtant, on constate qu'il n'est pas toujours éclairé, comme lorsqu'on partage ses données avec une plateforme pour accéder à un service ou qu'on accepte des conditions d'utilisation sans les avoir lues. Je pense qu'au-delà du consentement, c'est toute la gouvernance qu'il y a derrière qui est capitale. C'est pourquoi la création de tiers de confiance indépendants est essentielle. Ceux-ci doivent être indépendants des utilisateurs finaux de la donnée, mais aussi indépendants du régulateur. En outre, ils doivent pouvoir se doter de financements rendant réalisable la mise à disposition d'une grande quantité de données de qualité, sans remettre en cause cette indépendance. C'est aussi ce que prévoit le projet de texte européen, et c'est une avancée. Avec le soutien de la CNIL, notre fondation étudie actuellement les cas d'usage dans lesquels le DGA pourra ainsi promouvoir de façon réaliste l'altruisme des données.

OPEN DATA : SAVOIR ALLIER ÉTHIQUE ET PARTAGE

La donnée comme « commun numérique »

Stéphane Gigandet

Informaticien diplômé de l'école d'ingénieurs Centrale Nantes, Stéphane Gigandet est le co-fondateur d'Open Food Facts, une application citoyenne et collaborative qui permet de connaître en détail la composition des produits que nous achetons. Son immense base de données d'intérêt public guide les consommateurs vers des choix éclairés en matière d'alimentation, les entreprises dans l'amélioration de leurs produits et les scientifiques dans la connaissance de l'impact des produits sur notre santé.

Dates clés

2012 : *lancement d'Open Food Facts, association de loi 1901*

2016 : *83 000 produits sont référencés sur l'application*

2018 : *Open Food Facts reçoit le 1er prix du datathon de la Commission européenne et de l'Autorité européenne de sécurité des aliments*

2021 : *2 millions de produits sont référencés sur l'application*

« Il y a des données tellement importantes pour la société qu'elles doivent être rendues publiques et accessibles. C'est le cas des données sur notre alimentation, en raison de leur impact sur notre santé, la planète et la société. Elles sont essentielles pour appréhender des problèmes très actuels, comme l'obésité dans les pays développés ou encore l'empreinte carbone réelle des produits que nous consommons. Nous en avons besoin pour choisir nos aliments dans les rayons, les entreprises en ont besoin pour améliorer leurs produits, les scientifiques en ont besoin pour connaître l'impact de ce que l'on consomme sur notre santé…

C'est pourquoi nous nous sommes dit qu'il fallait l'équivalent d'un Wikipédia pour les produits alimentaires. Nous nous sommes inspirés de la célèbre plateforme, mais aussi d'Open Street Map, pour créer cette application où chaque personne peut apporter sa contribution. Au fil du temps, nous avons réussi à créer une transparence de fait. Au début, les entreprises étaient très réticentes. Puis certaines ont commencé à nous envoyer leurs données. Santé Publique France nous a même aidés à créer une plateforme pour faciliter ce partage d'informations. Toutes nos données sont ouvertes et nous recevons des centaines de milliers de contributions. Nos données sont factuelles, ce qui nous épargne le travail de modération que peuvent connaître, par exemple, les équipes de Wikipédia.

Open Food Facts est un bien commun et tout le monde a envie de participer. Nous voulons refléter au plus près l'impact environnemental réel des produits et influencer les entreprises vers des modes plus vertueux. Open Food Facts a aussi permis la création de centaines d'applications, dont la plus connue est Yuka, mais aussi des applications plus spécifiques pour les personnes diabétiques ou les personnes non-voyantes. Ce qui m'a rendu le plus fier ? Lorsque les chercheurs qui ont mis au point le Nutriscore se sont servis des données d'Open Food Facts pour valider ce système d'étiquetage en 5 niveaux. »

Le partage des données de santé à l'épreuve du COVID-19

C'est un sujet incontournable en 2021 : comment la crise sanitaire a-t-elle influencé l'ouverture des données de santé en France ? Quels défis ont dû relever les entrepreneurs, les pouvoirs publics et les associations de patients ? Martin Daniel, co-fondateur de Covidliste, témoigne aux côtés de Julien Marchal, co-directeur de l'innovation de l'Agence régionale de santé d'Île-de-France et de Nathalie Mesny, présidente de l'association de patients Renaloo.

Martin Daniel

En 2020, Martin Daniel a lancé Covidliste, une application qui a permis à plus d'1 million de personnes d'être alertées dès qu'un créneau de vaccination se libérait près de chez elles.

« Lorsque nous avons mis en ligne Covidliste, la pression du public, notamment sur les réseaux sociaux, a été très vive. Beaucoup de gens nous prêtaient des intentions qui n'étaient pas les nôtres, ils se demandaient ce que nous allions faire des données personnelles. Nous avons dès le début pris le parti de ne collecter aucune donnée de santé.

Si je dois tirer un enseignement de cette expérience, je dirais qu'il faudrait pouvoir simplifier le *privacy by design* : quand des développeurs conçoivent des applications et services, des éléments de protection des données par défaut devraient être disponibles. Par exemple, dans la Silicon Valley, l'incubateur YCombinator a simplifié les démarches de financement d'amorçage en concevant le SAFE (*Simplified Agreement for Future Equity*). Quel serait l'équivalent du SAFE pour la protection des données ?

Autre élément important : quand nous avons conçu Covidliste, il nous a fallu mobiliser 80 bénévoles pour appeler les 4 400 centres de vaccination qui s'étaient inscrits, afin de vérifier qu'il s'agissait bien de centres de vaccinations et pas d'usurpateurs. Nos équipes ont utilisé le système d'authentification des professionnels de santé développé par l'Agence numérique en santé : la carte e-CPS. Le principe ? Les contacts sont référencés via un numéro, et non leurs données personnelles. Alors, pourquoi ne pas imaginer un système équivalent pour identifier les personnes éligibles au vaccin, sans mentionner leurs données de santé ? » ●

Nathalie Mesny

En 2004, la fille de Nathalie Mesny est diagnostiquée d'une insuffisance rénale terminale.
Depuis 2015, Nathalie est présidente de l'association de patients Renaloo, qui porte la voix des malades du rein.

« Début 2020, notre souci en tant qu'association a été de délivrer un maximum d'informations à notre communauté de patients sur les risques qu'ils couraient et leurs droits. Le 30 mars 2020, l'Agence de biomédecine publie un bulletin reprenant des données de contamination et de mortalité des dialysés et greffés. Nous avons demandé l'accès à ces données, de façon transparente, afin de pouvoir trier les données par tranches d'âges.

On sait désormais que les personnes atteintes d'insuffisance rénale sont beaucoup plus à risques que la population générale senior, tant en termes de formes graves que de mortalité. Il a donc fallu orchestrer les soins spécifiques dans les hôpitaux et informer les patients et nous nous sommes heurtés à un manque de transparence de la part des instances d'État sur leurs données.

Or, depuis sa création, Renaloo structure justement les données médicales et juridiques utiles pour permettre aux patients de devenir acteurs de leur santé. En 2016, nous avons développé la toute première plateforme participative administrée par l'association, qui complète les données publiques sur les centres de soins (qualité, confort…) et sert à mener des enquêtes auprès des patients dans le cadre d'études. Sa gouvernance inclut un comité d'intérêt des patients, qui vérifie le protocole des études qui nous sont confiées, leur utilité pour l'amélioration des soins, le consentement éclairé des patients ainsi que leur accès aux résultats finaux de l'étude. Chaque patient sait que ses données servent uniquement à l'étude à laquelle il a consenti. » ⬣

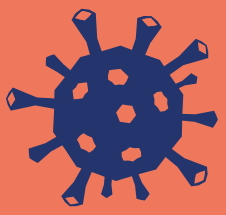
Julien Marchal

Julien Marchal a été nommé en novembre 2020 co-directeur de l'innovation de l'Agence régionale de santé d'Île-de-France, en pleine crise sanitaire.

« L'ARS travaille depuis longtemps avec les données de santé, en partenariat avec l'Assurance maladie, Santé Publique France, ou les remontées des établissements de santé. Pourtant, la première vague du COVID-19 nous a appris que toutes ces données ne suffisaient pas pour prendre une mesure exhaustive et suffisamment précise de la situation. Nous recevions des données hospitalières précises, qui nous permettaient de connaître le taux d'occupation au lit près. Mais il y avait un champ que nous avions du mal à étudier : l'épidémiologie, avec le nombre de contaminations, les clusters, la situation dans les ehpad…

Dans un premier temps, il a fallu se battre pour collecter ces données de façon artisanale, sous forme d'enquêtes, de questionnaires en ligne ou encore par téléphone. Le coup d'accélérateur a été donné au moment du déconfinement, avec la mise en place de nouveaux outils devenus indispensables à notre quotidien : SI-DEP pour les résultats de dépistages, Contact-Covid pour le *tracing*… Cela nous a permis de systématiser un certain nombre de choses et d'acquérir une connaissance très précise, inédite dans notre histoire, de la situation sanitaire en cours.

Notre défi maintenant ? L'anticipation. Nous avons développé des modèles prédictifs afin de gagner du temps sur l'épidémie. En couplant nos données à celles de la téléphonie mobile pour les déplacements, ou de Doctolib pour la fréquentation des cabinets médicaux, nous pourrions anticiper à moyen terme les reprises épidémiques ou les pics d'occupation dans les établissements de santé. » ⬣



Marin Dacos

Marin Dacos est coordinateur national pour la science ouverte au sein du ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation.

« Les données produites dans le cadre de la recherche ne relèvent pas du même régime juridique ni du même mécanisme de production que les données de santé : le volume peut être moindre, mais elles sont de grand intérêt car elles sont instrumentalement produites pour servir la connaissance.

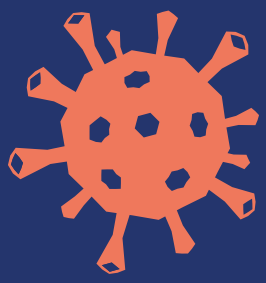
La pandémie mondiale du COVID-19 a été un accélérateur pour la recherche scientifique avec le développement d'un vaccin en un temps record. C'était également un test grandeur nature pour le développement des pratiques de science ouverte. Cependant, si l'ouverture des publications a en effet connu une accélération, et s'il y a bien eu des partages exemplaires de données, cela concerne surtout les données qui étaient déjà habituellement partagées, comme les séquences de SARS-CoV-2, très rapidement partagées par la Chine puis par le reste des pays. Malheureusement, le partage des autres données n'a pas réellement évolué. Le *Research on Research Institute* (RoRI) a ainsi produit une étude qui montre que le partage des données de recherche sur la pandémie n'a pas connu d'augmentation ou d'accélération par rapport à l'époque pré-pandémique. Il reste donc beaucoup de travail pour transformer la culture et les usages de partage des données en recherche en santé.
Les principaux enjeux d'avenir pour les données de recherche ? Arrêter de privilégier la publication et le partage des recherches qui ont fonctionné, car cela entraîne un biais éditorial qui pose problème sur le plan sanitaire et scientifique. Il faut aussi que l'on puisse publier les résultats négatifs. Le travail sur les données de la recherche doit être reconnu pour encourager les chercheurs à les partager. C'est ce à quoi appelle le Plan national pour la science ouverte porté par la ministre Frédérique Vidal. ●

Covidliste, la success story

En mars 2021, la vaccination est organisée par critères d'âges et de santé. Rapidement se pose le problème de l'offre et de la demande, avec d'un côté des personnes en train de faire le pied de grue devant les centres de vaccination pour obtenir un créneau et de l'autre, des centres avec des doses à écouler qui ne trouvaient pas preneurs…

Le 30 mars 2021, Martin Daniel et ses associés lancent Covidliste, la plus grande liste d'attente pour obtenir un vaccin en France. 4 400 centres de vaccination s'emparent de l'outil pour notifier par SMS les volontaires des doses disponibles. Pour s'inscrire sur Covidliste, chaque utilisateur doit d'abord fournir son nom, prénom, date de naissance, adresse géographique, numéro de téléphone et email. *« Ces données personnelles étaient indispensables pour faire fonctionner la plateforme »*, explique Martin Daniel.

Devant le succès fulgurant de l'outil, ses créateurs se mettent à craindre deux choses : *« les hackers russes et la CNIL »*. Afin de garantir la protection des données, la plateforme est alors hébergée sur un serveur ultra sécurisé en France, avec un code en open source. Les données sont chiffrées avant leur insertion en base, les noms et prénoms sont finalement supprimés, tout comme l'adresse exacte qui devient *« une géolocalisation à 100 m »*. L'association Hostolab est créée et devient la personne morale qui porte le projet Covidliste et le cabinet d'avocats McDermott Will & Emery rédige *pro bono* tous les documents juridiques (mentions légales, politique de traitement des données, procédés de suppression des données, etc.). La machine est lancée : entre mars et octobre, 1 124 913 personnes reçoivent une alerte pour un créneau de vaccin disponible. ●



POTENTIELS, RISQUES ET LIMITES DE L'EXPLOITATION DE LA DONNÉE

Le formidable potentiel de l'*open data* ne doit pas nous faire oublier les risques inhérents à l'ouverture non contrôlée de la donnée. Comment garantir l'anonymat ? Est-ce que l'intelligence artificielle peut reproduire certaines inégalités ? Comment révéler des données sensibles sans les citer ? Yves-Alexandre de Montjoye, professeur à l'Imperial College de Londres, Anne Bouverot, co-fondatrice et présidente de la Fondation Abeona et Pierre Romera, CTO de l'*International Consortium of Investigative Journalists* apportent leur éclairage.

Yves-Alexandre de Montjoye

Professeur à l'Imperial College de Londres

Anonymisation et pseudonymisation des données ?

« Lorsqu'on souhaite utiliser des données tout en protégeant la vie privée des personnes, on passe par l'anonymisation. Le principe est simple : il s'agit de casser le lien entre la personne et ses données. Le RGPD ne s'applique plus à des données qui sont anonymes, où quand la personne n'est plus identifiable.

Traditionnellement, l'anonymisation repose sur deux étapes : d'abord, on met en place la pseudonymisation (on retire tous les identifiants directs tels que noms ou numéros de téléphone), puis on utilise des techniques de désidentification (mélange ou suppression de certaines données pour empêcher que la personne ne puisse être réidentifiée). Mais aujourd'hui, ces méthodes historiques (le concept d'anonymité a été inventé dans les années 90) se heurtent à de nouvelles difficultés. Par exemple, la pseudonymisation est de plus en plus compliquée à obtenir. Il suffit de quatre données issues du téléphone mobile (liées à la localisation ou à l'horaire) pour identifier de manière unique 95 % de la population ! Par ailleurs, on constate que le risque de réidentification dans une base de données reste significatif, quels que soient le volume des données ou les tentatives de bruit pour les masquer. Enfin, il a été montré qu'il était possible de construire des modèles statistiques pour confirmer qu'on a bien trouvé la personne qu'on recherche dans une base de données.

Alors, comment garantir l'anonymisation aujourd'hui ? En opérant un changement de paradigme… Il ne faut plus voir l'anonymisation comme un état fixe d'un jeu de données qui aurait été anonymisé une fois. Il faut investir dans des techniques nouvelles telles que les *Query-Based Systems* ou l'ajout de bruit *Differentially Private* et les challenger en permanence tout en s'en tenant au cadre fixé par le législateur. »

Intelligence artificielle, l'ouverture contre les biais

Anne Bouverot

Co-fondatrice et présidente
de la Fondation Abeona

« La source de tout ce que l'on fait en intelligence artificielle, ce sont les données. Quand les données de départ sont biaisées ou incomplètes, le risque est de les reproduire ou de les amplifier. L'exemple le plus connu est celui d'Amazon en 2014, dont l'algorithme de recrutement discriminait les femmes. Alors, comment prévenir ces risques ?

Déjà, avant toute mise en service d'un système, il faut le tester. Pas seulement pour voir s'il marche, mais aussi pour vérifier s'il apporte un traitement juste et équitable. Autrement dit, il faut le tester sur des données réelles et notamment les variables sensibles, là où il y a un risque de discrimination. Cela doit se faire de manière encadrée, anonyme et via un tiers de confiance. En France, nous avons constitué un groupe de travail avec l'Institut Montaigne et la Fondation Abeona pour formuler des recommandations sur l'équité en matière d'intelligence artificielle.

Aujourd'hui, l'intelligence artificielle et le *big data* sont des domaines qui font encore peur : les conditions de création de la confiance reposent donc sur l'explication. Comme le disait Marie Curie, « dans la vie il n'y a rien à craindre, il y a tout à comprendre ». C'est pourquoi avec OpenClassrooms et l'Institut Montaigne, nous avons créé un MOOC (*Massive open online course*, un cours en ligne accessible à tous) qui explique ce qu'est l'intelligence artificielle, ses atouts et ses risques. Cette initiative citoyenne indépendante a porté ses fruits avec déjà plus de 150 000 personnes qui ont suivi ce cours. Par ailleurs, de la même manière que l'on audite les résultats financiers d'une entreprise, il faudrait faire auditer les systèmes d'intelligence artificielle qui présentent des risques par des tiers indépendants. »

Datajournalisme, une déontologie de la donnée

Pierre Romera

CTO de l'International Consortium
of Investigative Journalists

« *L'International Consortium of Investigative Journalists* (ICIJ) est connu pour ses missions d'investigation, qui amènent nos équipes à publier des informations sensibles : *Pandora Papers*, *Panama Papers*, *Paradise Papers*... Dans le cadre de notre mission, nous sommes amenés à collecter des données ouvertes, qui nous permettent de vérifier les documents internes et données sensibles que nous recevons. C'est un processus long. Pour les *Pandora Papers*, nous avons croisé près de 12 millions de documents internes relatifs à 14 fournisseurs de sociétés *offshore*. Ces données sensibles avaient été produites dans le plus grand secret et n'étaient pas destinées à être publiées. L'ICIJ mène une vraie réflexion éthique sur l'impact de ces données et l'intérêt qu'elles peuvent avoir pour le plus grand nombre. Nous protégeons aussi nos sources : l'ICIJ ne divulgue jamais ses sources. C'est une question de déontologie : notre métier est de servir le plus grand nombre et pas simplement de mettre des données à disposition du public. On travaille donc longuement pour trouver comment extraire ces données sans les publier : cela nous a pris deux ans pour les *Pandora Papers*. On met en avant des schémas récurrents pour révéler un problème systémique sans dévoiler les documents d'origine. Face au volume de données que nous recevons, nous ne pouvons pas tout traiter d'un coup. L'ICIJ noue ainsi des partenariats pour réutiliser les données même après publication de l'enquête. »

4

L'ART DU PARTAGE

L'universalisation de la lecture et de la compréhension de la donnée

Caroline Goulard

Co-fondatrice et présidente de Dataveyes, entreprise qui démocratise les données grâce à des outils et interfaces permettant à des personnes non expertes de se les approprier.

La vue, sens premier pour une compréhension immédiate

Coral Cities par Craig Taylor chez ITO World

L'idée de Coral Cities semble simple : cartographier la mobilité de 40 villes à l'aide de formes de coraux. Le résultat permet d'apprécier, en un coup d'œil, la distance qu'il est possible de parcourir en voiture en 30 minutes depuis chaque centre-ville. Mais également, de comparer chaque ville et son développement grâce au motif.

Le toucher, la physicalisation des données

Trends in water use par Adrien Segal

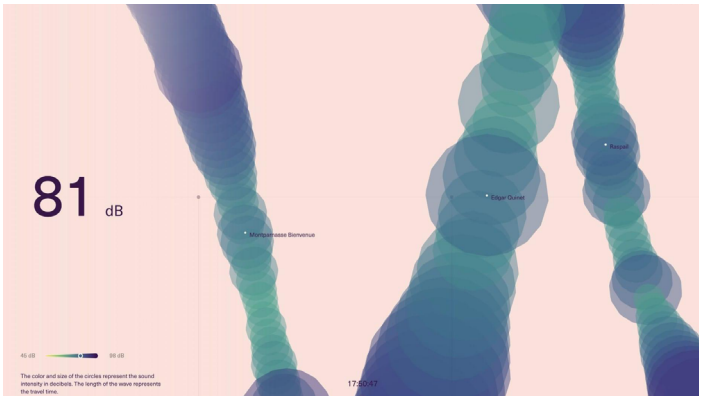
Adrien Segal utilise notre capacité à appréhender des objets physiques, à les toucher, à les retourner, à les manipuler, pour montrer la consommation d'eau aux États-Unis ces 50 dernières années. Chaque affluent divisant le canyon représente une utilisation catégorielle différente (irrigation et énergie thermoélectrique), la largeur des parois du canyon est définie par la quantité d'eau utilisée et la hauteur représente le temps, de 1950 (en haut) à 2000 (à la base).

Les données sont un mode d'encodage de l'information adapté aux ordinateurs, mais difficile d'accès pour nous, humains. Nous n'avons pas les mêmes mécanismes cognitifs et l'information ne nous parvient pas de la même façon. Mais comment doit-on, alors, traduire cette donnée pour l'adapter ? À travers des interfaces qui communiquent avec nos sens, s'amusent des particularités humaines et transforment une masse d'information en un message compréhensible par tous…
Tour d'horizon des idées les plus créatives.

Quel est le goût des données ?

Data cuisine par Moritz Stefaner et Susanne Jaschko[4]

Data Cuisine est un atelier qui explore la nourriture comme support d'information. Son objectif est de représenter des données ouvertes locales à travers la couleur, la forme, la texture, l'odeur ou le goût de la cuisine. Par exemple, des chocolats représentent le taux de mortalité d'un pays ou d'une région et leur texture les causes.



Sonification : traduire les mélodies en données

Commute par Dataveyes

Le but : éveiller les consciences sur la pollution sonore subie. Après l'enregistrement de la densité sonore dans toutes les lignes du métro parisien, l'application décompose le bruit enregistré en fréquences et en intensités, puis les transforme en motifs visuels et sonores esthétiques. L'application propose ainsi de comparer les lignes et « d'écouter » les moments où le son est le plus fort.

[4] http://data-cuisine.net

épilogue

« Le cadre de la réutilisation des données demande à être clarifié et précisé au regard des exigences et des possibilités offertes par le RGPD »

Anne Debet

Professeur de droit privé à l'Université Paris Descartes et membre de l'Institut Droit et Santé (IDS). Elle représente la Présidente de la CNIL au sein de la CADA et du Conseil supérieur des archives

L'évènement AIR 2021 a montré, grâce aux nombreuses interventions, l'intérêt majeur de l'ouverture et de la réutilisation des données (transparence de la vie publique, enjeu d'amélioration des politiques publiques, meilleure information des citoyens...). Cette ouverture et les réutilisations subséquentes présentent toutefois des risques en particulier pour la protection des données à caractère personnel d'autant que, comme cela a été évoqué, l'anonymisation réelle de ces données est difficile.

Or, dans ce domaine, le cadre juridique, en particulier celui de la réutilisation des données, demande à être clarifié et précisé au regard des exigences et des possibilités offertes par le RGPD. Cette clarification est d'autant plus indispensable aujourd'hui qu'une ouverture plus large des données est prônée tant au niveau national qu'au niveau européen. En France, quatre ans après la création du service public de la donnée par la loi du 7 octobre 2016 pour une République numérique, le rapport de la mission Bothorel, rendu le 11 décembre 2020 pose le constat général de la nécessité d'une ouverture plus large des données publiques et fait des propositions concrètes pour permettre cette ouverture. Au niveau européen, la Directive 2019/1024 du 20 juin 2019 concernant les données ouvertes et la réutilisation des informations du secteur public donne une nouvelle impulsion au mouvement de l'open data, mouvement que les projets européens relatifs au *Data Governance Act* et au *Data Act* devraient encore accélérer.

L'accélération ne peut toutefois se faire que dans un cadre respectueux de la protection des données et la CNIL comme la CADA sont très souvent saisies des interrogations des différents acteurs sur l'équilibre à respecter entre ouverture et protection. Les deux autorités ont déjà commencé à travailler sur ce sujet dans le cadre du Guide commun sur l'Open Data, publié en 2019, mais un approfondissement de ces travaux est aujourd'hui indispensable. C'est la raison pour laquelle la présidente de la CNIL a souhaité la constitution d'un groupe de travail sur le sujet et m'a confié la direction de ces travaux.

Pour avoir une approche plus transversale des questions, il a été décidé de ne pas se limiter aux questions de réutilisation en lien avec l'open data des données publiques mais d'envisager plus largement les réutilisations de toutes les données librement accessibles en ligne. L'objectif de ce groupe de travail sur la réutilisation des données, dont les travaux devraient commencer en janvier 2022 est, en partant de cas concrets, d'aboutir à la rédaction d'un guide et de fiches pratiques à l'intention tant des diffuseurs que des réutilisateurs des données, guides et fiches pratiques qui seront soumis à consultation publique. Il est nécessaire pour cela de tenir compte des spécificités des différents contextes dans lesquels les diffusions des données ou les réutilisations sont susceptibles d'intervenir (dispositions légales applicables, objectifs poursuivis, sources et nature des données en cause, conditions de collecte, etc.). Le champ ouvert est donc à la fois large et passionnant et les travaux auront pour objectif de trouver, dans le cadre légal aujourd'hui applicable, le bon équilibre entre la nécessité du partage et les enjeux de la protection.

Commission nationale de l'informatique
et des libertés
3 place de Fontenoy
TSA 80715
75334 PARIS CEDEX 07
Tél. 01 53 73 22 22

www.cnil.fr

CNIL.