RÉALITÉS INDUSTRIELLES

« Se défier du ton d'assurance qu'il est si facile de prendre et si dangereux d'écouter »
Charles Coquebert, Journal des mines n°1, Vendémiaire An III (septembre 1794)

La protection des données dans une économie globalisée

UNE SÉRIE DES
ANNALES DES MINES
FONDÉES EN 1794

AOÛT 2022

Publiées avec le soutien de l'Institut Mines Télécom

UNE SÉRIE DES
ANNALES DES MINES
FONDÉES EN 1794

RÉALITÉS INDUSTRIELLES

Série trimestrielle – Août 2022

Rédaction

Conseil général de l'Économie (CGEIET),
Ministère de l'Économie, des Finances
et de la Souveraineté industrielle et numérique
120, rue de Bercy - Télédoc 797
75572 PARIS CEDEX 12
Tél. : 01 53 18 52 68
http://www.annales.org

Grégoire Postel-Vinay
Rédacteur en chef

Gérard Comby
Secrétaire général

Alexia Kappelmann
Secrétaire générale adjointe

Magali Gimon
Assistante de rédaction / Maquettiste

Myriam Michaux
Webmestre et maquettiste

Publication

Photo de couverture
Libre de droits. Photo téléchargée sur le site de PIXABAY (https://pixabay.com/fr/photos/gdpr-les-donn%c3%a9es-big-data-s%c3%a9curise-3777666/).
Photo©John Collins

Iconographie
Gérard COMBY

Mise en page
Myriam MICHAUX

Impression
Dupliprint Mayenne

La protection des données dans une économie globalisée

Le dossier a été coordonné par le **Dr. Laure TABOUY, PhD**

Éditorial

Par Marie-Laure DENIS
Conseiller d'État, présidente de la CNIL

L'économie mondiale a connu depuis trois décennies un bouleversement majeur de son organisation. Si les années 1980 ont vu les débuts d'une ouverture des marchés, ce sont deux événements majeurs du début des années 1990 qui ont radicalement modifié les pratiques économiques et industrielles, dans la foulée immédiate de la chute du monde soviétique.

D'une part, en 1993, sont actées la pérennisation de l'Accord général sur les tarifs douaniers et le commerce (GATT) et la création, à partir de 1995, de l'Organisation mondiale du commerce, qui est chargée d'organiser un commerce international sans barrières majeures. Dans ce nouvel accord, la France s'est longuement battue (avec succès) pour préserver « l'exception culturelle et audiovisuelle ». Mais, concernant le numérique, le principe de la libre prestation de services informatiques entre membres de l'OMC s'impose.

D'autre part, la même année, le CERN (Centre européen pour la recherche nucléaire), qui accueille le chercheur britannique Tim Berners-Lee, inventeur du "World Wide Web", décide de verser le logiciel Web inventé par ce dernier dans le domaine public, permettant par cette décision une réutilisation libre de ses principales briques : adresses URL, protocole de transfert fondé sur les liens hypertextes (HTTP) et langage de mise en forme des contenus HTML.

La combinaison de ces deux initiatives, portées et soutenues activement tant par les États-Unis que par l'Union européenne, va permettre l'éclosion de nouveaux acteurs du numérique lors d'une première vague d'innovation à la fin des années 1990 (les "dotcoms"), suivie d'un reflux puis d'une dynamique constante à partir de la fin des années 2000. Cette période voit naître les entreprises qui aujourd'hui sont les plus capitalisées du monde : Amazon (fondée en 1994), Google (en 1998) et Facebook (en 2004), tandis que des entreprises de la première génération de l'informatique continuent de se développer (Apple et Microsoft, toutes deux créées en 1976). Au cours de la même période, certains acteurs historiques de l'informatique, comme IBM ou les opérateurs télécoms, voient leurs positions relatives diminuer dans le paysage de l'innovation technologique.

Plus récemment, cette vague d'innovation numérique touche l'ensemble des secteurs traditionnels en y imposant le modèle dit "winner takes all", permis à la fois par la technologie du Web (et, à partir de 2007, des applications mobiles) et par l'ouverture des marchés, notamment occidentaux. Dans le tourisme, l'industrie, la musique et même, *in fine*, dans l'audiovisuel pourtant protégé par des accords commerciaux, des acteurs globaux s'imposent *via* le numérique. Malgré la diversité des secteurs concernés, l'ingrédient commun de ces succès est souvent la force d'un modèle de services unique pour le monde entier, mais hautement personnalisé à l'échelle de chaque utilisateur.

En effet, alors que l'industrialisation des XIX^{e} et XX^{e} siècles conduisait à une uniformisation des produits, l'avènement du numérique résout en partie cet apparent paradoxe entre uniformité et personnalisation, grâce à l'exploitation des données générées par les utilisateurs eux-mêmes ; ces données offrant une information fine et automatique sur les goûts et les préférences de chacun. S'il est alors particulièrement aisé de personnaliser un service complètement dématérialisé, comme le sont un réseau social ou un service de *streaming* de musique, ce phénomène touche aussi largement d'autres secteurs d'activité comme l'automobile, la santé et même l'aéronautique ou la chimie.

Dans ce contexte foisonnant, il est probable que la décennie 2020 soit marquée par le retour de la régulation concernant le numérique, comme le montrent en Europe (mais aussi des initiatives équivalentes aux États-Unis) les législations DSA (Digital Services Act), DMA (Digital Markets Act), "Data Act", "Data Governance Act" et, bien évidemment, la première d'entre elles, le RGPD (Règlement général sur la protection des données).

En effet, que représente la protection des données dans l'ensemble de la dynamique d'innovation numérique des dernières décennies ?

Les fondements de la protection des données personnelles datent de la première vague de développement de l'informatique, dans les années 1970. À la suite d'un intense débat public dans la plupart des pays occidentaux, la France, comme ses homologues européens, se dote en 1978 d'un cadre de protection des données personnelles tiré du droit à la protection de la vie privée, droit fondamental réaffirmé dans la Convention européenne des droits de l'homme de 1950. Ce cadre repose sur des principes simples, qui se sont révélés particulièrement robustes par rapport aux évolutions de la technologie, comme l'a souligné le Conseil d'État dans son étude annuelle sur le numérique et les droits fondamentaux en 2015 : la finalité du traitement des données, sa proportionnalité, sa sécurité et l'existence de droits directs pour les personnes concernées, notamment le droit d'accès et le droit de suppression.

Ce cadre a toutefois d'abord été pensé en référence à des systèmes informatiques bien identifiés, situés dans des centres de données fixes et créés spécifiquement pour la mise en place du traitement de données projeté. La CNIL, dans sa mission originelle et toujours intense de contrôle des traitements de données personnelles mis en place par l'État, continue d'ailleurs d'appliquer cette grille de lecture à de nombreux systèmes qui lui sont soumis pour avis.

Néanmoins, il est clair que le développement de l'informatique distribuée, notamment *via* l'informatique en nuage, le recours à des services en ligne plutôt qu'à des serveurs physiques ou encore les usages massifs en mobilité, représente des défis majeurs en matière de déclinaison des principes de la protection des données au niveau de l'informatique moderne.

Parmi ces défis, trois dimensions illustrent particulièrement la tension entre les impératifs de protection des données et le fonctionnement de l'économie globalisée, dans laquelle nos sociétés évoluent.

Premièrement, il est très difficile de prendre la pleine mesure de l'état des technologies et de leur évolution. Ce problème concerne tout à la fois les pouvoirs publics, les régulateurs et la société civile dans son rôle de veille : l'effort annuel de R&D des GAFAM représente 117 milliards de dollars d'investissement[1] répartis sur des dizaines de lignes technologiques différentes, quand le budget annuel de l'Inria, le principal centre de recherche français sur le numérique, représente environ 250 millions d'euros. Cette asymétrie de moyens, qui n'est pas propre au numérique, rend toutefois difficile le dialogue technico-juridique avec les acteurs principaux du numérique : les produits et les services sont souvent conçus dans des services d'ingénierie eux-mêmes mondialisés, dont les principaux centres de conception sont situés hors d'Europe et sont dirigés par des personnes qui ne sont généralement pas familières des exigences juridiques européennes. Le RGPD a introduit le principe du "privacy by design" pour favoriser la prise en compte de la protection des données dès la conception des produits ou des services. Mais l'application de ce texte reste lié à la mise sur le marché européen du produit ou du service visé et s'applique donc imparfaitement à des travaux situés très en amont de cette mise sur le marché et le plus souvent réalisés hors d'Europe.

Deuxièmement, l'informatique évolue de plus en plus vers un modèle de fourniture de services à distance dans lequel les serveurs sont partagés, virtualisés et répartis dans plusieurs centres de données. Ce modèle est à la fois plus simple à opérer pour les clients comme pour les fournisseurs et plus efficace en termes de coût, mais il engendre d'importants flux de données entre les utilisateurs et les serveurs du fournisseur et de ses sous-traitants. Or, le RGPD, comme la Directive qui l'a précédé, fixe le principe d'une protection des données personnelles, y compris en cas de transfert des données hors de l'Union européenne. En l'absence de législation comparable dans de nombreux pays, notamment les États-Unis d'Amérique, la question des transferts vers ces pays s'est posée dès la fin des années 1990 : en la matière, l'adoption d'un premier cadre de transfert, le "Safe Harbor", a été finalement annulée par la Cour de justice de l'Union européenne, comme l'a été son successeur, le "Privacy Shield" (arrêts Schrems et Schrems 2). Si l'Europe a toujours été très favorable à l'ouverture des marchés et au développement du commerce international, il est important de relever qu'elle a également toujours revendiqué une prééminence de la protection des libertés fondamentales (dont la protection des données personnelles) sur les objectifs commerciaux ; une prééminence par ailleurs prévue par les accords de l'OMC.

Enfin, la troisième tension majeure, liée à la précédente, réside dans les conflits de législations applicables aux services numériques : dans un monde où tout service est conçu de manière uniforme pour toute la planète, il est particulièrement difficile pour les entreprises qui les opèrent de se conformer à de multiples législations, parfois contradictoires. Cette difficulté, au cœur de l'arrêt Schrems 2, s'exprime au sein de l'Union européenne au travers d'interprétations parfois divergentes entre les autorités nationales. Le RGPD a visé à harmoniser les pratiques, mais le « mécanisme de cohérence » instauré par ce texte pour faciliter le dialogue entre les autorités, connaît une mise en œuvre progressive et soulève encore des questions. En particulier, il reste difficile pour des autorités nationales d'accéder en pratique à des infrastructures de traitement de données qui ne sont pas situées sur leur territoire.

Les nouveaux textes législatifs de la Commission européenne (DSA, DMA, DGA, "Data Act", "AI Act") poursuivent également l'objectif de consolider un cadre réglementaire harmonisé en Europe et d'en faire un standard de niveau mondial à l'instar du RGPD. Mais les impacts sur les acteurs européens et non européens restent à déterminer.

En conclusion, les trois tensions développées ci-dessus pour le cas de la protection des données personnelles et observées depuis quelques dizaines d'années illustrent également les difficultés qui pourront se poser pour la protection des données à d'autres titres, comme le secret industriel ou le droit d'auteur (droit *sui generis*), même si, à ce niveau, d'autres problématiques entrent en jeu comme la capacité des acteurs industriels à mettre en commun certaines données dans un environnement de confiance. Pour autant, les décisions récentes de la CJUE comme les projets législatifs de la Commission européenne ouvrent la voie à une application stricte et directe des législations européennes à tous les services opérant sur le marché continental, ce qui implique des évolutions des pratiques des acteurs globalisés pour mieux prendre en compte les spécificités du cadre juridique européen.

[1] https://en.wikipedia.org/wiki/List_of_companies_by_research_and_development_spending

Introduction

Quel monde voulons-nous ? Jusqu'où pouvons-nous et devons-nous aller ?

Par le Dr Laure TABOUY, PhD
Neuroscientifique et éthicienne, équipe Éthique et épistémologie, CESP-INSERM U1018,
Espace éthique APHP, Université de Paris-Saclay

Confiance, fiabilité, responsabilité, éthique, transparence, rigueur, confidentialité, disponibilité, intégrité : voici quelques mots clefs importants pour parler de la protection des données et de ses enjeux.

Concevoir l'avenir de la protection des données dans une économie globalisée et dans un contexte international incertain de guerre et de crise sanitaire nous demande d'être plus responsables, plus vigilants, plus réfléchis au regard de nos données. L'utilisation que nous en faisons et la protection de ces dernières figurent parmi les préoccupations majeures en 2022.

Mais de quoi parlons-nous ? Qu'est-ce qu'une donnée, ou, plus précisément, une donnée sensible ou une donnée chiffrée ? Est-ce qu'une donnée identitaire soulève les mêmes enjeux qu'une donnée de recherche ou de santé, une donnée cérébrale, démographique, d'interaction ou de comportement, ou encore transactionnelle, financière ou industrielle ? De plus, définir les données de la recherche n'est pas un exercice aisé, dans la mesure où elles sont de natures extrêmement diverses.

L'usage et la réutilisation des données en vue d'une confirmation de résultats de la recherche ou de l'apport d'une aide à la décision, ou d'accroître la confiance dans l'innovation, mais aussi en vue de favoriser le partage des données et la définition de modèles économiques de valorisation de celles-ci sont des enjeux cruciaux en 2022 pour tous les acteurs du monde de la recherche et des entreprises. L'attractivité de la France en dépend.

Mais il ne faut pas oublier que le *business* des données va bon train : le *business* de la vente mais aussi du vol des données des utilisateurs d'Internet ou des patients a des incidences sur les actions de la vie quotidienne des citoyens, sans qu'ils ne s'en rendent compte. Il peut ruiner les fondations d'une démocratie en favorisant l'addiction et la manipulation[1]. Car les réseaux sociaux sont une drogue, en tant qu'ils déclenchent le relargage de la dopamine et favorise le renforcement du circuit de la récompense. Les outils technologiques, l'IA et les algorithmes sont aujourd'hui au cœur des innovations ; ils ont été étudiés et conçus pour permettre de contrôler l'attention des utilisateurs, tout en collectant leurs données.

Le respect de la vie privée et la protection des données sont des enjeux majeurs de notre société, où, dans la période actuelle, le piratage des données s'accentue, où des milliards de dollars sont investis dans but de monétiser ces données, où les risques d'usurpation d'identité et de manipulation des populations à grande échelle sont à prendre très au sérieux.

Les développements récents issus de la convergence de technologies émergentes se traduisent par une accélération de l'innovation se concrétisant par des applications dans les entreprises et le monde de la recherche que ce soit dans les domaines de la santé, de la finance, de l'aéronautique, de la cybersécurité, de la Défense, ou en matière de recrutement, d'accompagnement des projets, de consommation ou de déplacement, mais aussi dans le domaine de la recherche, toutes disciplines confondues en sciences humaines et sociales et sciences exactes, notamment en neurosciences et en génétique.

Au regard de la place que prend aujourd'hui la problématique de la protection des données dans notre société, le comité éditorial de la série *Réalités industrielles* des *Annales des Mines* a donc souhaité contribuer à la mise en œuvre des recommandations émanant des différentes autorités et des dispositions de la loi Informatique et libertés en dédiant le présent numéro à l'avenir de la protection des données dans les entreprises et dans la recherche dans un contexte international plus responsable.

[1] Voir le documentaire « Derrière nos écrans de fumée », https://www.thesocialdilemma.com/

Ce numéro des *Annales des Mines* s'articule autour de trois axes qui le structurent et ont pour intitulés :

1. L'utilisation des données : entreprises privées et recherche académique, ensemble pour un avenir plus responsable de la gestion des données.
2. L'économie des données : quelles utilisations des données au cœur des collaborations, des partenariats ou des plateformes collaboratives. Quelles données partager ? À qui en ouvrir l'accès ?
3. Les sciences humaines et les données : les enjeux de la mutation des sociétés à travers les données. Faire confiance au temps du numérique.

La problématique de la protection des données étant subordonnée à la singularité des spécificités de chaque entreprise et de la recherche, nous avons fait le choix, dans la première partie de ce numéro, de nous focaliser sur l'utilisation des données et sur les enjeux et les questions auxquels nous sommes confrontés en la matière.

L'obligation d'obtenir un consentement libre et éclairé des personnes avant une utilisation de leurs données ; la nécessité que les entreprises, les industriels, mais également les organismes de recherche soient rigoureux, fiables, responsables, transparents, intègres et honnêtes dans l'exercice de leurs activités ; le besoin urgent de réfléchir à la pertinence de la collecte des données et de leurs lieux de stockage, d'ouvrir le débat au sein de la société mais aussi des entreprises et des communautés de chercheurs, de légiférer, de mettre des bornes, des limites et de mettre en balance les bénéfices et risques en matière de partage des données : ce sont là des problématiques qui doivent être considérées de manière très sérieuse aujourd'hui. Ce sont des questions que prennent à bras le corps les délégués à la protection des données (DPO) présents dans les entreprises mais aussi certains cabinets d'avocats, qui sont les interlocuteurs privilégiés des entreprises et des organismes de recherche, et qui sont les garants du respect de la loi.

Que signifie une donnée nous appartenant en droit ? Que recouvrent les termes « souveraineté » et « confidentialité » des données ? Qui a le droit de réutiliser des données, notamment celles de clients, pour une autre utilisation que celle prévue à l'origine ?

Le respect de la confidentialité des données collectées dans le cadre des activités des entreprises ou de recherche, qu'elles soient civiles ou militaires, n'est plus une option. Des protections appropriées des données numériques comme non numériques présentes sur des espaces privés, de santé et relatives à notre identité individuelle, ainsi que des contenus sur Internet doivent être mises en œuvre pour nous aider dans notre compréhension des droits de l'homme qui s'y attachent.

Cette juste protection des données va permettre aux entreprises et aux organismes de recherche de continuer à être, au plan national et international, des acteurs influents majeurs de la valorisation et du transfert vers la société des innovations, dans toutes leurs dimensions. Il en va de la confiance que nos concitoyens accordent et vont continuer d'accorder aux entreprises, à l'État et aux chercheurs qui travaillent au sein des organismes de recherche académiques et privés.

Les enjeux de la protection des données qui sont au cœur de l'activité de nombre de filières industrielles, notamment de la *healthtech* et de la *medtech*, et de celle des laboratoires et autres organismes de recherche sont colossaux.

Quels sont les enjeux à relever par la France sur le plan de la protection des données pour qu'elle reste attractive aux yeux des entreprises et de la recherche ? Qu'est-ce qu'une donnée sensible en recherche, selon les disciplines ? Comment les organismes de recherche s'organisent-ils ? Comment ces différentes disciplines, notamment celles travaillant sur le microbiote et celles des neurosciences, s'organisent-elles à l'échelle européenne ?

Pour apporter des éléments de réponse à ces nombreuses questions, nous inviterons certains acteurs de la science des données à s'exprimer en tout début de la deuxième partie de ce numéro. Souvent associée à la collecte des données massives et à l'analyse de celles-ci, cette discipline interdisciplinaire utilise entre autres « des méthodes et des processus d'apprentissage automatique, des algorithmes pour extraire des connaissances et des idées de nombreuses données structurelles et non structurées. »

L'internationalisation des données est un véritable enjeu. Qu'est-ce que représentent ces données ? Qu'est-ce que l'anonymisation et le chiffrement des données ? Comment sont-elles classifiées et catégorisées ? Comment le *machine learning*, les algorithmes et la *blockchain* interviennent et permettent de représenter ces données ? Comment sont organisées ces données au sein d'une bibliothèque de données de recherche ?

Puis nous nous intéresserons plus précisément à l'utilisation de l'anonymisation et du chiffrement qui permettent d'apporter une réelle protection aux données stockées dans des *cloud*, sur des serveurs ou dans des entrepôts et dont le partage s'opère au sein de consortiums et de collaborations nationales et internationales, comme le Health Data Hub[(2)] ou le Paris-Santé Campus[(3)].

(2) https://www.health-data-hub.fr/

(3) https://parisantecampus.fr/

Assurer la protection des données issues de la recherche à l'ère du numérique est l'engagement pris par l'État pour permettre l'ouverture de la science[4] : « La France s'engage pour que les résultats de la recherche scientifique soient ouverts à tous, aux chercheurs, aux entreprises et aux citoyens, sans entrave, sans délai, sans paiement ». L'enjeu pour le monde de la recherche est inédit et gigantesque ; il demande d'instaurer un véritable dialogue entre tous les acteurs concernés : « La science ouverte est la diffusion sans entrave des résultats, des méthodes et des produits de la recherche scientifique. »

Les différents projets de recherche internationaux, comme le Human Brain Project[5] et l'International Brain Initiative[6], mettent l'accent sur le développement des neurotechnologies qui visent à mieux comprendre les fonctions cérébrales et à intervenir sur elles. Le Health Data Cloud[7] vise à fournir des services en lien avec les données cérébrales sensibles. Ce consortium comprend un ensemble d'infrastructures existantes partenaires du Human Brain Project/EBRAINS[8] et de fournisseurs de services de données de santé qui ont récemment rejoint ce projet.

Propriété intellectuelle, domaine et données publics, protection des données personnelles... Les données sont à la croisée d'une multitude de régimes juridiques, parmi lesquels il peut être difficile de se repérer.

Enfin, parler de protection des données nécessite de s'attarder sur les enjeux de celle-ci à l'heure des réseaux sociaux et d'Internet.

Quels sont les enjeux et les limites de cette protection ? Quelles sont les questions et les craintes qu'elle soulève pour les futures générations ? Qu'en est-il du rôle et des responsabilités sociales que doivent assumer les citoyens, les industriels, les entreprises, les politiques, les chercheurs et les organismes de recherche publics et privés, dans la génération de ces données et le devenir de celles-ci ?

Dans la dernière partie de ce numéro, le lecteur pourra naviguer au cœur de la problématique des données et de l'innovation responsable, au travers des contributions de multiples acteurs, dont votre serviteur, spécialistes du droit français et européen, comme l'OCDE et la CNIL, de la cybersécurité sous la plume de représentants de la gendarmerie nationale et de l'armée de terre et, enfin, de l'éthique et de l'épistémologie à travers la position exprimée par la HTF (Human Technology Foundation).

Les questions soulevées sont nombreuses et demandent de faire preuve d'une exigence, d'une pertinence et d'une rigueur sans commune mesure pour traiter ces sujets si brûlants et si politiques. Les enjeux liés aux données, à leur protection, à leur collecte, à leur réutilisation peuvent être de différents ordres.

À la seule évocation des mésusages potentiels des données, apporter des limites éthiques, juridiques, économiques et sociétales à leur utilisation s'avère nécessaire pour que s'instaure une confiance informée des patients, des consommateurs et, plus largement, des citoyens, pour que les investissements et les innovations s'orientent vers des biens et services à forte utilité économique, sociale et environnementale. L'évocation de tels risques permet aussi d'éveiller les consciences à la nécessité de protéger ce que les hommes sont en eux-mêmes, car derrière les données, ne l'oublions pas, il y a des êtres humains, avec leurs histoires singulières.

Les rôles joués par la CNIL et l'AMF, les recommandations de l'OCDE, les dispositions du RGPD ou du Data Act sont centraux à l'échelle française et européenne, en tant qu'ils régissent les aspects juridiques de la protection des données.

L'accélération de l'essor des technologies et des innovations rend indispensable l'engagement de réflexions sur les enjeux qu'elles soulèvent au regard des données, et de travailler à la conception interdisciplinaire de certifications, de normes, de dispositifs d'autorisation, de systèmes d'évaluation et de surveillance et de cadres de gouvernance adaptés aux valeurs sociologiques, économiques, éthiques et juridiques de la France et de l'Europe. Compte tenu du large spectre que recouvrent les questions soulevées par l'usage des données, la Commission européenne, au travers du Règlement général de protection des données (RGPD)[9], entré en application le 25 mai 2018, mais aussi l'OCDE[10] et [11], au travers de ses recommandations n°463, adoptée le 6 octobre 2021, et n°433, adoptée le 13 décembre 2016, et la France, par l'intermédiaire de la CNIL[12], et l'Autorité des marchés financiers[13] ont pris leurs responsabilités en adoptant les lois et recommandations qu'exige la protection des données.

---

(4) https://www.ouvrirlascience.fr/

(5) https://www.humanbrainproject.eu/en/

(6) https://www.internationalbraininitiative.org/

(7) https://www.healthdatacloud.eu/

(8) https://ebrains.eu/

(9) Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016. https://www.cnil.fr/fr/reglement-europeen-protection-donnees

(10) OECD, Recommendation of the Council on Health Data Governance, OECD/LEGAL/0433.

(11) OECD, Recommendation of the Council on Enhancing Access to and Sharing of Data, OECD/LEGAL/0463.

(12) https://www.cnil.fr/professionnel

(13) https://www.amf-france.org/

Comme le souligne Marie-Laure Denis dans l'éditorial qui introduit ce numéro : « Ce cadre, qu'est le RGPD, a toutefois d'abord été pensé en référence à des systèmes informatiques bien identifiés, situés dans des centres de données fixes et créés spécifiquement pour la mise en place du traitement de données projeté. »

Les réflexions éthiques et épistémologiques liées aux données, au *Big Data*, à leur protection et à leurs usages s'expriment en général par le biais de questionnements et d'intuitions. Le rapport « *DATA* altruisme, une initiative européenne »[(14)] aborde les enjeux des données sous l'angle de leur mise au service du bien commun.

Réfléchir aux enjeux éthiques nécessite de s'interroger, de s'ouvrir à la démarche de questionnement sur les valeurs et les finalités, les principes et les normes, les contextes et les conséquences de nos actions à des moments où ces dernières sont ambiguës, en tension, voire en conflit avec notre environnement. Une telle réflexion, essentielle à l'innovation responsable, permet alors de poser factuellement le débat.

C'est sur ces trois aspects que ce numéro se clôturera, en nous questionnant sur ce que ces données disent de l'homme, et sur ce que cela change dans notre compréhension de l'humain.

De même, la question de la protection des données sera abordée à la lumière de l'éthique de la recherche, de l'intégrité scientifique, de l'épistémologie et de la neuroéthique, pour en venir à nous questionner sur les enjeux que la science ouverte sous-tend.

Car l'idée que les sciences « s'ouvrent », renvoie à des pratiques de mise en culture desdites sciences, d'une présence de celles-ci dans un grand nombre d'espaces sociaux et sociétaux. Le monde de la recherche aussi bien privée qu'académique et celui des entreprises sont-ils capables de rester autonomes et indépendants, tout en travaillant conjointement pour le bien commun et en interdépendance avec la société et sont-ils prêts à ouvrir largement leurs données ?

(14) https://www.human-technology-foundation.org/fr-news/rapport-data-altruisme

Dans un monde à la fois connecté et en tension, quels enjeux et quelles approches pour assurer valorisation et protection des données ?

Par Jérôme ANDRES
Systèmes d'information et communication sécurisés du groupe Thales

La révolution numérique intervenue au cours des trente dernières années est à l'origine de très nombreuses mutations au plan économique et dans notre vie quotidienne. Internet qui en est le socle s'est bâti sur l'échange de données toujours plus importantes, mais surtout sur leur valorisation dans tous les champs : scientifique, sociétal, politique, artistique, ludique... L'information et la connaissance qu'elles rendent possibles constituent un actif qu'il s'agit aujourd'hui de protéger au regard de leur valeur intrinsèque, mais aussi des conséquences que peut induire, directement ou indirectement, leur utilisation. À l'image du numérique et comme toute autre innovation, l'émergence des données n'est pas neutre, représentant tour à tour un risque ou une opportunité, un remède ou un « poison ». La cybersécurité est une pratique et un champ industriel qui peuvent permettre de mesurer et de canaliser ces différentes alternatives, mais qu'il faut aussi appréhender sur le plan politique, au niveau national comme international, pour tenter d'en réguler les équilibres.

Nouvel Eldorado, nouvel or noir ?

Les métaphores aurifères sont fréquentes concernant l'opportunité que revêt l'usage des données depuis une trentaine d'années et l'émergence d'Internet. À l'évidence, elles ont fait la fortune de quelques nouvelles entreprises nord-américaines (les GAFAM[1]), mais aussi chinoises (BATX[2]), qui ont su en quelques années seulement acquérir des positions dominantes inattendues bouleversant nos quotidiens. Aujourd'hui, ce sont nos *smartphones* qui en sont l'exemple le plus éclairant : ils nous sont aujourd'hui indispensables, que nous nous déplacions en recourant à un service de navigation, que nous suivions nos performances sportives ou physiologiques et, depuis deux ans, les courbes d'infection au Covid-19, les taux de vaccination ou d'hospitalisation et les clefs numériques de notre passe sanitaire, ou bien que nous consultions les derniers résultats sportifs, les annonces immobilières ou les biens d'occasion en vente, ou encore pour maintenir le lien avec nos proches au travers de réseaux sociaux ou avec des inconnus dont les centres d'intérêt nous sont similaires. Les applications qui rendent ces services possibles reposent bien sûr sur les technologies de l'électronique et des télécommunications, mais plus particulièrement sur un échange toujours plus varié de données, plus rapide d'informations et plus vertigineux de connaissances entre chacun d'entre nous ou les organisations au sein desquelles nous exerçons.

Hier, nous naviguions à partir d'ordinateurs fixes contraints par des débits bien moins importants. Demain, l'on nous fait la promesse que la 5G permettra de nous connecter à tous nos objets environnants, démultipliant encore nos échanges, telle une rhapsodie folle dans un rythme infernal, à la source de multiples inquiétudes (vie privée, réchauffement climatique, perte de souveraineté, espionnage, manipulation…). Tandis que les thuriféraires de l'innovation nous renvoient, hier comme aujourd'hui, à une seule crainte, celle du changement.

(1) GAFAM : Google, Amazon, Facebook, Apple et Microsoft, qui sont les cinq mastodontes américains de la nouvelle économie et qui sont plus puissants que bien des États. Il est à noter que les trois premiers cités n'existaient pas il y a vingt-cinq ans de cela.

(2) BATX : Baidu, Alibaba, Tencent et Xiaomi : les équivalents des GAFAM pour le Web chinois, lesquels sont des acteurs dominants en Chine et prépondérants en Asie. Leur création remonte à peine à vingt-quatre ans pour les plus anciens.

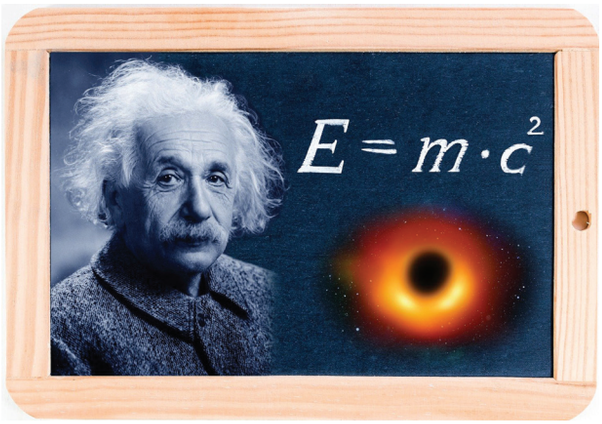


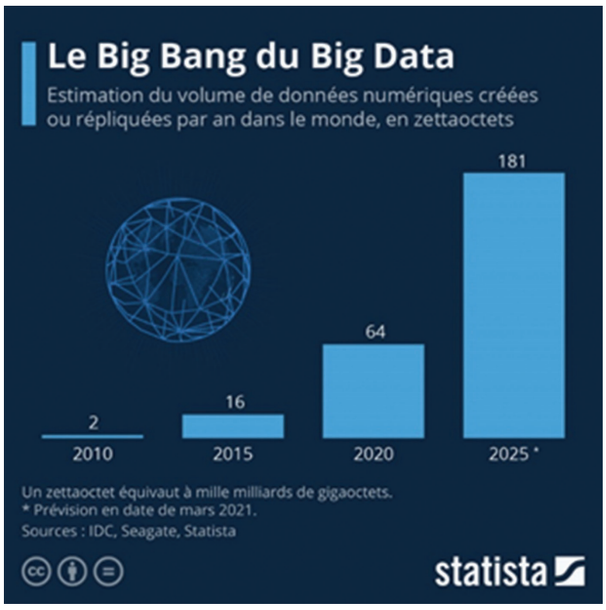
Photo © STUX

Photo 1 : La théorie de la relativité générale d'Albert Einstein, à la base de la découverte des trous noirs.
Source : Photo libre de droits téléchargée sur PIXABAY,

En adoptant une image plus financière, les données sont perçues comme un atout, ou encore un actif (le même mot « asset » recouvrant les deux sens en anglais) que l'on cherche à la fois à protéger et à valoriser. S'il s'agit de la manipuler tel un minerai précieux, il faut alors successivement la collecter, la transformer, la raffiner, la livrer et, enfin, la recycler en recourant à des outils et des processus adéquats et efficaces. À l'instar du pétrole et de la pétrochimie, l'émergence d'informations utiles passe par la combinaison de données, un alliage de celles-ci qui parfois est explosif socialement, politiquement ou économiquement parlant, rendant nécessaire la mise en place de cadres normatifs de différents ordres pour tenter de réguler les enjeux qui y sont associés.

Quelles sont les données à protéger ?

Toutes les données ne sont pas égales. Numérisées, ce sont toujours des suites de zéros et de symboles (bits) combinés en octets. Prises ensemble, elles peuvent représenter des volumes importants, comme une image ou une vidéo pour les plus massives que nous manipulons avec nos *smartphones*. On parle alors de résolutions pouvant aller de 330 x 480 pixels[3] pour les vidéos VHS, soit 0,16 Mpixels[4], jusqu'à 70 Mpixels pour l'IMAX qui a été utilisé par Christopher Nolan pour réaliser le film *Interstellar*, format actuel le plus fin en matière de caméra cinématographique. Cela est encore bien peu en comparaison des données accumulées pour aboutir à l'exploit scientifique rendu public le 12 mai 2022 : la reconstitution par l'EHT[5] du trou noir Sagittarius A* localisé au centre de la voie lactée. Cette prouesse a reposé sur la mobilisation de plusieurs



radiotélescopes répartis sur la surface du globe pour constituer un réseau d'antennes de la taille de la Terre, agissant comme une gigantesque lentille. Chaque campagne de recueil de données a permis d'amasser 7 pétabytes de données[6], sur cinq jours ; l'image divulguée en mai dernier a, elle, exigé la réalisation d'une multitude de calculs depuis 2019.

Ces chiffres vertigineux ne représentent qu'une partie des données numériques disponibles : IDC, Seagate et Statista estiment qu'en 2025, le volume des données créées ou répliquées sera égal à 181 zêta octets, soit mille milliards de Giga octets. Pour ne s'en tenir qu'à cette échéance…

Les données commerciales et industrielles

Les vidéos VHS des années 1970 à 1990 sont souvent les seuls souvenirs animés ou les seuls enregistrements des émissions télévisuelles de l'époque ; elles constituent un patrimoine familial, voire historique que l'on peut souhaiter préserver. S'agissant d'*Interstellar*, ce qui est ici protégé, c'est la propriété intellectuelle de C. Nolan et de ses ayants droit…, au point de ne pouvoir accéder sur Wikipédia qu'à une image d'un format 220 x 326, soit 0,07 Mpixels, donc une image bien moins précise que celle de notre bonne vieille cassette VHS. L'encyclopédie en ligne précise à ce propos que l'image de l'affiche précitée est présentée dans le cadre du *fair use* défini par la loi américaine du copyright : « tout usage ultérieur pouvant constituer une violation de ce droit », telle est l'infraction qui pourrait être imputée aux *Annales des Mines* en cas d'utilisation. Si la qualité en est suffisante et

(3) Un pixel, contraction de *picture element* en anglais, est l'unité de base mesurant la résolution d'une image numérique matricielle.

(4) Un méga pixel mesure une résolution d'image d'un million de pixel, soit, par exemple, une image matricielle de mille pixels de côté.

(5) Event Horizon Telescope est un consortium international de collaboration en matière d'observation des trous noirs grâce à l'utilisation combinée de plusieurs radiotélescopes. Dans les faits, il s'agit d'observer le flot de particules entourant le trou noir, qui ne délivre aucune information et dont aucun cliché ne peut être fait.

(6) Un pétabyte est une unité de mesure de volumes d'informations correspondant à un billion d'octets, soit la capacité de sauvegarde de 1 000 iPhone de dernière génération.

permet sans doute en outre de limiter les besoins de stockage afférents et par là même les coûts d'énergie et les émissions de gaz à effet de serre, la publication de cette image est source d'incertitudes juridiques : on ne sait quelle action pourrait engager les ayants droit, mais le risque pour l'éditeur des présentes *Annales des Mines* d'une exposition à des poursuites serait réel.

Le piratage massif de données dont a été victime en 2014 Sony Pictures Entertainment, la branche américaine de divertissement du géant japonais de l'électronique, a eu des retentissements et des conséquences bien plus cette fois-ci pour les ayants droit ou leurs intermédiaires. Un mois avant Noël, les employés de cette entreprise voient apparaître sur leurs écrans des images de squelettes et des messages de chantage exigeant la déprogrammation du film *L'interview qui tue !*, une comédie mettant en scène un projet d'assassinat du dictateur nord-coréen Kim Jong-Un. La Corée du Nord a toujours nié son implication dans ce piratage. Malgré l'appel au FBI et à Mandiant, une société de cybersécurité, cinq films inédits sont mis en ligne sur des sites pirates ainsi que le script du futur *James Bond*. Des courriers professionnels et confidentiels sont aussi divulgués.

Plus récemment, en 2020, les États et les acteurs mondiaux principaux de la pharmacologie engagés dans la recherche de vaccins contre le Covid-19 ont dû, semble-t-il, faire face, selon Microsoft, à des tentatives de piratage provenant de la Russie ou de la Corée du Nord, mais bien entendu sans reconnaissance explicite de celles-ci, comme c'est le cas pour la plupart des attaques étatiques supposées. À moins que ce n'aient été que de plus simples tentatives d'espionnage industriel, la valeur future attendue des informations de la recherche contre le Covd-19 représentant sans nul doute un énorme enjeu financier…

Les données personnelles

Les entreprises commerciales et industrielles ne sont malheureusement pas les seules à être victimes de vol de données. En marge des attaques qu'elle a subies en 2014, Sony Picture Entertainment a indiqué que des informations personnelles concernant ses employés ou leurs proches, comme leurs coordonnées bancaires, de cartes de crédit ou des mots de passe, avaient pu être dérobées.

Comme évoqué *supra*, le contexte du Covid-19 nous a aussi malheureusement rappelé que ces pratiques frauduleuses concernent aussi les particuliers, dont nombre d'entre eux ont vu leurs résultats de tests PCR divulgués. Les données de pas moins 1,4 million de personnes (identité, numéro de Sécurité sociale, coordonnées…) ont ainsi été dérobées en juin 2020. De manière intéressante, la « fuite » n'a pas émanée de l'application centrale SI-DEP[7] stockant les résultats des dépistages, mais d'une autre application[8] d'échange de fichiers, Dispose, qui joue un rôle de passerelle dans le cadre du traçage des contaminations. C'est cette dernière application qui, rendue vulnérable par une faille de sécurité inconnue à l'époque, a été à la source de la fuite.

Cela constitue un exemple fort intéressant de risques intrinsèques à la valorisation des données – ici le traçage des contaminations à partie des résultats positifs –, lorsque certaines mesures de précaution ne sont pas mises en œuvre – ici la sécurisation des échanges de données. Ce qu'illustre bien la reprise par le philosophe Bernard Stiegler, dans le contexte du numérique, du concept grec de *Pharmakon*[9], cher à Platon et Derrida, celui du double serpent du Caducée d'Hermès, qui est à la fois remède et poison. Fallait-il mettre en œuvre ce type de traçage pour tenter de casser le cycle des contaminations, au risque d'un éparpillement des données personnelles pouvant, par exemple, servir à des usurpations d'identité ou à du hameçonnage[10]? L'efficacité limitée du traçage en France pourrait incliner à regretter cette séquence malheureuse, mais pouvait-on en prévoir l'issue par avance ? À moins que l'on ne cherche un bouc-émissaire – la troisième face du *Pharmakon* antique… Le responsable du piratage et de la fuite des données s'est avéré être un opposant au passe sanitaire ; il s'est défendu en déclarant vouloir « démontrer la faiblesse et la faillibilité du système d'information de l'AP-HP »[11]. Mais il y a bien d'autres manières de rendre connues de telles vulnérabilités

D.R.

Photo 2 : Le philosophe Bernard Stiegler.

[7] Système d'informations de dépistage, qui est la « plateforme sécurisée où sont systématiquement enregistrés les résultats des laboratoires de tests Covid-19 ».

[8] Deux articles du journal *Le Monde*, datés des 15 (https://www.lemonde.fr/pixels/article/2021/09/15/covid-19-les-donnees-de-tests-de-1-4-million-de-personnes-derobees-aux-hopitaux-de-paris_6094806_4408996.html) et 21 septembre 2021 (https://www.lemonde.fr/pixels/article/2021/09/21/comment-les-donnees-de-1-4-million-de-franciliens-testes-pour-le-covid-19-se-sont-retrouvees-dans-la-nature_6095455_4408996.html).

[9] *Internet n'est pas neutre, Internet est un pharmakon. Ce qui nous arrive sur la Toile*, épisode du mardi 14 janvier 2014, par Xavier de La Porte, France Culture (https://www.radiofrance.fr/franceculture/podcasts/ce-qui-nous-arrive-sur-la-toile/internet-n-est-pas-neutre-internet-est-un-pharmakon-9357167).

[10] Technique de piratage, préalable à une usurpation d'identité numérique, consistant à se faire passer pour un tiers de confiance qui aurait un accès légitime aux données dérobées.

[11] Article du journal *Le Monde*, daté du 8 octobre 2021 (https://www.lemonde.fr/pixels/article/2021/10/08/vol-massif-de-donnees-de-sante-de-l-ap-hp-un-pirate-informatique-arrete_6097637_4408996.html).

informatiques : ainsi, les « Hackers éthiques » ont pris l'habitude d'en informer les éditeurs de logiciels incriminés ; autre exemple, celui de communautés comme les CERT (Computer emergency response team), qui traquent ces failles, facilitent l'élaboration des réponses à y apporter ou les conçoivent directement.

Le risque zéro n'existe pas, en tout cas, en ce qui concerne le numérique et le traitement des données, pas plus qu'en médecine ou en matière de transport. Tout est affaire d'anticipation, d'analyse et de compromis bien compris, mais il demeurera toujours une part d'incertitude.

Les acteurs de la cybersécurité s'appliquent à mettre en œuvre dans ce cadre une démarche dite d'analyse de risques, qui s'intéresse notamment :

• aux menaces pouvant entacher les processus supportés par les systèmes d'information ;
• aux risques proprement dits, à leur probabilité, à leur impact et à leur objet : la sécurité physique des personnes ou des biens, la perte ou le vol de données, leur modification, leur divulgation etc. ;
• enfin, aux moyens à mettre en œuvre pour réduire ces risques, diminuer la possibilité de leur survenue ou leurs conséquences, les délais et les coûts associés, à accepter ou à arbitrer.

Cela signifie qu'il existe toujours des risques dits résiduels, qu'il s'agit de documenter et, surtout, de surveiller ; autre domaine important de la cybersécurité.

Au niveau européen, la mise en place du RGPD (le Règlement général sur la protection des données), promulgué en 2016 et applicable depuis 2018, vise à définir le cadre d'exploitation des données à caractère personnel. C'est une référence juridique unique à l'échelle de l'Union européenne, qui, à elle seule, ne protège pas techniquement les usagers, mais donne un cadre à respecter par les entreprises et fait des émules à travers le monde, y compris en Californie, le berceau de la nouvelle économie d'Internet.

Photo © Gordon JOHNSON

Photo 3 : Le caducée, symbole de l'Ordre des médecins.
Source : Photo libre de droits téléchargée sur PIXABAY.

Données critiques et enjeux de souveraineté

L'impact d'une attaque sur les données peut-il toujours être circonscrit aux seules personnes ou entreprises qui en sont propriétaires ? Malheureusement, non. C'est dans ce cadre que l'on évoque la notion d' « opérateur d'importance vitale » (OIV), telle que définie dans le cadre du Code de la Défense en 2007. Si les entreprises sont invitées, ou plutôt incitées à protéger leurs données, si les particuliers sont sensibilisés à la nécessité d'être vigilants au regard de leurs traces numériques, certaines organisations, lorsqu'elles sont attaquées, peuvent être à la source, malgré elles, de conséquences bien plus larges.

La liste des opérateurs d'importance vitale n'est pas publique, mais ce sont pas moins de douze secteurs d'activité, relevant du public ou du privé, qui sont concernés depuis 2008. Depuis 2016, une série de neuf arrêtés ont été pris par le Premier ministre ; des arrêtés rédigés par l'ANSSI, l'Agence nationale de la sécurité des systèmes d'information[12], qui pour plusieurs sont au cœur des sujets abordés dans la présente revue : sous-secteurs de la recherche publique, de l'audiovisuel et de l'information, les communications électroniques et Internet, la santé, dont les données et leur protection sont considérées comme indispensables à la survie de la nation.

La recherche publique notamment, et toutes les connaissances qu'elle rend possibles, ont été jugées clefs par le législateur pour la pérennité de l'État et du corps constitué par les citoyens, mais aussi pour son rayonnement – et sans doute encore davantage pour l'innovation qu'elle permet en matière de santé, de matériaux, de composants, d'agriculture et d'environnement, et qui peut être à la source du développement de demain. Un laboratoire développant des vaccins est à cette aune sans doute plus « vital » qu'un éditeur de films ou de jeux.

L'audiovisuel traditionnel et, de plus en plus, les moyens numériques d'échange sont les médias de base de tout partage de données, mais aussi de formidables moyens d'influence. Le piratage de la chaîne TV5 Monde, le 8 avril 2015, a été un vrai coup de semonce : à la suite de la mise hors service en cascade d'équipements, la chaîne francophone doit interrompre toute diffusion ; en parallèle, ses comptes officiels Twitter et Facebook sont piratés et utilisés pour diffuser des messages en soutien à l'État islamique (EI). L'identité de proches de militaires français engagés dans la lutte contre l'EI est divulguée ; le président de la République François Hollande y est

[12] L'Agence nationale de la sécurité des systèmes d'information, service créé en France en 2009 à la suite de la création de la direction centrale de la Sécurité des systèmes d'information, « assure la mission d'autorité nationale en matière de sécurité des systèmes d'information. À ce titre, elle est chargée de proposer les règles à appliquer pour la protection des systèmes d'information de l'État et de vérifier l'application des mesures adoptées. Dans le domaine de la défense des systèmes d'information, elle assure un service de veille, de détection, d'alerte et de réaction aux attaques informatiques, notamment sur les réseaux de l'État » (https://www.ssi.gouv.fr/).

également mis en cause, quelques mois après la série d'attentats de janvier 2015. La piste de l'EI est progressivement abandonnée, en juin de la même année, les soupçons pointant alors vers le groupe de hackers AP28, proche du gouvernement russe ; l'attaque informatique a eu lieu quelques semaines seulement après l'annulation par la France d'un contrat de fourniture à la Russie de deux porte-hélicoptères Mistral, comme sanction à la suite de la crise ukrainienne d'alors, l'annexion de la Crimée par la Fédération de Russie.

En plein contexte de l'invasion récente de l'Ukraine, d'autres règlements européens ont été proposés, comme le DMA (Digital Markets Act), lequel, attendu pour 2023, vise à encadrer les pratiques des entreprises du numérique, ou encore le DSA (Digital Services Act) sur les termes duquel la Commission et le Conseil européens se sont mis d'accord, modernisant ainsi la direction dite « e-Commerce », afin de réguler les contenus numériques illégaux, la publicité et la désinformation.

Cette « crise » s'est en effet transformée depuis le début de cette année en une guerre – présentée comme une « opération spéciale » par les gouvernants russes – qui se déroule aux portes de l'Union européenne. Comme dans tous les conflits, la présentation des faits, les mots utilisés pour les nommer, le sens que l'on veut leur donner sont autant de moyens d'influence ou de contre-influence. Une des plus grandes surprises de ce conflit, et sans doute des succès enregistrés par l'Ukraine, concerne le domaine de « l'information », où l'on pouvait s'attendre à une bien plus grande domination de la part de la Fédération de Russie. Tous les champs de la communication ont été utilisés par les Ukrainiens et certains soutiens occidentaux :

• une protection Cyber plutôt très efficace des systèmes d'information ukrainiens, même si le piratage du satellite civil KA-SAT, dès le 24 février 2022, a eu des conséquences allant bien au-delà de l'Ukraine, pour les utilisateurs civils européens (un exemple typique, celui des opérateurs d'importance vitale) ;
• la divulgation par les Américains par avance des plans d'attaque russes, comme forme de « contre-propagande » très originale jouant sur la transparence de l'information ;
• l'utilisation de la constellation de satellites Starlink du milliardaire Elon Musk pour diriger les attaques dévastatrices de drones sur les colonnes de blindés russes ;
• la communication du président Zelensky devant les représentations nationales de très nombreux pays, la visite de Kiev devenant un lieu tendance pour les leaders internationaux ;
• la prise de contact avec les proches de soldats et de conscrits russes pour les informer du décès de leur fils ou de leur conjoint par le truchement de la reconnaissance faciale et l'analyse de données ouvertes afin de retrouver leur famille. Cette démarche accompagnée par la société Clearview AI a été jugée comme fort cavalière et bien peu respectueuse des droits humains et du droit de la guerre.

Ce sont autant d'exemples de valorisation de données sur tous les champs informationnels : télécommunications, médias, intelligence artificielle, transparence contre désinformation… Autant d'exemples de l'effet majeur de l'utilisation de la donnée et de l'information qu'elle permet de produire. Il faut donc tout à la fois pouvoir utiliser ces données, mais aussi les protéger, les tracer et en contenir l'exploitation.

Un autre parallèle pourrait être fait avec l'énergie nucléaire : à la différence des hydrocarbures, l'utilisation de cette énergie peut sembler être plus discrète, infinitésimale (au niveau de l'atome), moins directement polluante, mais surtout extraordinairement efficace, voire dangereuse si on libère sa « puissance » et qui peut nuire très longtemps, à l'image des données qui ne sont jamais réellement effacées. À moins qu'elles ne soient absorbées par un trou noir, mais peut-être pourront-elles s'en échapper par un trou de ver[13], le nouvel horizon de la recherche astrophysique.

[13] Concept théorique et encore hypothétique, suggéré par Albert Einstein en 1935, et dénommé ainsi en 1957 par Charles W. Misner et John A. Wheeler. S'il était avéré, il permettrait, en toute hypothèse, de rejoindre deux régions de l'espace-temps, le voyage temporel, et donc de libérer « ailleurs » l'information annihilée par le trou noir. L'hypothèse est utilisée par Christopher Nolan dans *Interstellar* comme dénouement de l'énigme des « fantômes » à l'origine de la chute de livres dans une bibliothèque. La future chercheuse, encore enfant, est sans le savoir en contact avec son propre père qui tente de lui faire passer des messages à travers ce trou de ver.

Cibler, à travers la donnée, les ruptures dans les parcours de santé : une chance pour les patients et pour les innovateurs

Par Marco FIORINI
Directeur du projet « Intelligence artificielle et cancer » du contrat stratégique de filière des industries et technologies de santé

Stéphanie KERVESTIN
Déléguée générale de l'Ariis (Alliance pour la recherche et l'innovation des industries de santé)

Et Virginie LASSERRE
Directrice des Affaires externes de Janssen France, co-lead de l'axe « IA et Parcours de soins » du contrat stratégique de filière des industries et technologies de santé

Dans cet article, nous mettons en avant l'idée que les filières des industries pharmaceutiques, de conception de dispositifs médicaux, de diagnostic et de santé numérique présentent des histoires et des technologies distinctes.
En regard, la production de données de santé par chacune de ces filières crée un terreau commun pour l'innovation. Ce terreau est particulièrement fertile lorsque son action s'oriente vers la description des parcours de santé des patients, lesquels recouvrent les phases de prévention, de diagnostic, de traitement et de suivi. Cette « perspective du patient » est celle que servent l'ensemble de ces filières et, au-delà du secteur privé, celle qui fait l'objet de politiques publiques.
Nous décrivons ici comment, au sein d'un contrat stratégique de filière des industries et technologies de santé, la capacité de ces industries à se doter avec l'État d'une sémantique commune autour de la description des parcours de santé permet d'envisager une méthode de quantification des ruptures intervenant au sein de ces parcours, lesquelles sont autant de pertes de chances pour les patients.
L'objet ici est d'en objectiver l'importance pour les prioriser au regard de technologies dont nous disposons en France.

Propos introductif

Le parcours de santé s'étend tout au long de notre vie. Ce parcours passe par une prise en charge globale et se poursuit ensuite, non seulement pour diagnostiquer et prévenir des pathologies, mais aussi pour accompagner, soigner et suivre tous les patients.

Il s'agit par là même d'un enjeu commun à toutes les filières qui composent la famille des industries de santé (diagnostics et dispositifs médicaux, objets connectés, industries du numérique, médicaments et services à la personne) : celui de faire progresser la qualité du service rendu aux patients.

Jusqu'à récemment, ces écosystèmes évoluaient, se structuraient, se projetaient s'appuyant sur des référentiels distincts hérités de leurs histoires respectives.

La généralisation des données de santé transforment cette dynamique : de plus en plus, les dispositifs médicaux, les industries du médicament et les outils numériques génèrent des données de santé qui constituent une matière commune et transverse aux industries du secteur.

Issues de services ou de produits distincts, ces données tour à tour donnent des informations complémentaires au regard de nos paramètres physiologiques, reflètent notre complexité biologique, restituent l'état de l'art des capteurs qui permettent la mesure ou la sociologie d'usage de chacune des solutions médicales à travers les applications de nos portables.

Mais au-delà d'être un produit commun aux différentes filières industrielles de la santé, ces données représentent également un potentiel considérable à exploiter de façon partagée.

C'est l'idée que nous aimerions expliciter dans cet article : regroupées, ces données permettent de disposer d'une vision intégrée, à date, des « parcours de santé » des patients[1].

Sur ce point, nous avons travaillé à développer la capacité des différents acteurs à mieux définir les parcours de santé. Mais au-delà de la définition et de la description de ces parcours, il nous est rapidement apparu important d'en cibler avant tout les ruptures : pertes de chances en termes de prévention efficace, de diagnostic conduisant à la prescription d'un traitement personnalisé, d'accès aux médicaments ou de suivi performant.

Il est également rapidement apparu qu'à travers l'analyse de la donnée, la perspective de définir un parcours de santé du patient pouvait déboucher sur une vision intégrée, libérant une opportunité de construire un référentiel partagé grâce à une mutualisation sémantique entre les différentes filières concernées, entre les secteurs public et privé, voire une méthode commune de description des parcours et des ruptures associées.

Au regard de ces défis, il nous semble important de nous doter dès à présent d'une doctrine commune : comment peut-on décrire de tels parcours ? Comment peut-on en cibler les ruptures ? Comment, grâce à des données de plus en plus nombreuses aujourd'hui, peut-on quantifier ces ruptures ? Comment structurer la fillière pour y répondre ?

Pour l'heure, la France peut s'appuyer sur son système national de données de santé (SNDS), une base médico-administrative unique au monde. Celle-ci présente deux forces : elle est le reflet des coûts des solutions de santé effectivement mises en œuvre aujourd'hui et elle est exhaustive. Cette base de référence peut donc nous servir d'ancrage pour construire une doctrine commune.

Demain, la France pourra également compter sur les données issues d'objets connectés remboursés par l'Assurance maladie ; celles-ci seront versées dans le domaine public. Ces données seront porteuses d'informations complémentaires à celles médico-administratives évoquées plus haut : mesures biologiques, physiologiques, de ressenti ou d'exposition. Il se dégagera en conséquence des opportunités en matière de constitution de *datasets* (jeux de données) inédits.

C'est à travers la diversité des typologies des données composant les *datasets* supports que vont se constituer ces opportunités qui vont se traduire par un accroissement de notre capacité à nous poser des questions inédites, de plus en plus fines sur les parcours de santé, mais aussi de notre capacité à cibler, préciser et quantifier de plus en plus précisément les ruptures évoquées ci-dessus.

Une initiative prend corps au sein du contrat stratégique de filière des industries et technologies de santé

Issu du Conseil national de l'industrie et signé par les représentants de l'État et ceux des acteurs privés du secteur, le contrat stratégique de filière des industries et technologies de santé a vocation à opérer un rapprochement en la matière entre le secteur public et le secteur privé, lequel est chargé de faire des propositions stratégiques pour la nation avec pour objectif de créer de la valeur en termes économiques et de création d'emplois et, en ce qui concerne plus particulièrement les industries de la santé, de la valeur pour les patients, les soignants et le système de santé.

C'est dans ce contexte qu'est né un projet dédié aux parcours de santé. Son objet est très simple : valoriser les données publiques au travers de méthodes de description des divers parcours de santé, mais aussi utiliser ces données pour faire apparaître les ruptures dans ces parcours aux yeux des patients et des professionnels de santé et, pour aller plus avant, objectiver et quantifier l'importance de ces ruptures en termes de nombre de patients concernés et de coût pour la société, pour enfin considérer l'état de l'écosystème d'innovation français au regard de ce qui se fait dans le monde dans le but d'apprécier notre capacité à prioriser et à nous employer à réduire telle ou telle rupture.

Cette méthode a l'avantage de pouvoir se focaliser sur l'innovation, sur la coordination des moyens de la prévention et des soins autour de ce qui nous concerne tous, à savoir le développement d'une médecine efficace et accessible pour les patients et d'outils modernes pour les praticiens.

Par ailleurs, ce cadre vise à offrir aux *start-ups* du numérique en santé et aux *biotechs* une capacité d'accéder plus rapidement au marché. En d'autres termes, nous entendons ici la possibilité pour elles de contextualiser et d'objectiver leurs propositions d'innovation dans un ensemble plus vaste, des propositions qui soient lisibles pour le secteur public. L'idée est de leur permettre d'apporter rapidement les preuves de l'utilité et de la viabilité du concept que chacune porte, au travers d'une expression claire et partagée des besoins, pour, dans un premier temps, leur donner les moyens de se déployer au niveau de la France, avant d'envisager de s'ouvrir à l'export.

Au-delà du principe, comment s'y prendre concrètement ? Le projet « Parcours de santé » repose sur le lancement de sous-projets pilotes pour apprendre à décrire les parcours de santé, à organiser des échanges où des praticiens s'expriment et réagissent au regard de leurs pratiques quotidiennes, et où les patients ouvrent des perspectives au regard de leur vécu.

[1] Nous utilisons l'expression « Parcours de santé » plutôt que celle de « Parcours de soins », car nous souhaitons ériger l'étape de prévention en un objectif sociétal à systématiser. Dans ces parcours de santé, nous distinguons ainsi la succession de différentes étapes : celles de la prévention, du diagnostic, du traitement et du suivi des patients.

Zoom sur le retour d'expérience de l'initiative « Impact »

Lancée en septembre 2021, l'initiative « Impact – Accélérateur d'innovation en santé mentale » a pour ambition de faire émerger des solutions utilisant des technologies et des données innovantes pour remédier aux ruptures identifiées dans le parcours de soins de santé mentale du jeune adulte et de l'adulte. Cette initiative s'articule autour d'un appel à projets collaboratifs et partenariaux ; cet appel a été ouvert aux *start-ups* de la *e*-santé, qui ont été invitées à proposer des solutions autour de thématiques prioritaires, telles que l'accès aux soins ou la prise en charge, le suivi et la continuité extrahospitalière des patients.

En janvier 2022, les huit organismes à l'origine de cette initiative (PariSanté Campus, la Fondation Université de Paris, l'Assistance publique – Hôpitaux de Paris, France Biotech, l'ARIIS, France Assureurs, Axa France, Janssen France, Otsuka France et Eisai) annonçaient que cinq lauréats avaient été retenus : Fedmind, Litdhospi, Tech2Heal, Tricky et ResilEyes Therapeutics. Pour accélérer leur déploiement, les *start-ups* lauréates bénéficient depuis janvier 2022 d'un programme d'accompagnement de neuf mois, reposant sur une forte implication des partenaires de l'initiative.

Au-delà de celui de la santé mentale, l'ambition est de mener d'autres pilotes qui pourraient reverser dans le domaine public des données issues de dispositifs médicaux et d'objets connectés, enrichissant ainsi en données médicales les bases médico-administratives qui constituent un socle commun de connaissances.

Ces données permettraient de participer à la construction de visions utiles à la direction générale de l'Offre de soins du ministère chargé de la Santé, pour mieux comprendre l'articulation entre la médecine de ville et la médecine hospitalière, pour construire des indicateurs dans le cadre du dialogue avec les praticiens et pour promouvoir la création d'outils qui mettent les patients en capacité d'être acteurs dans le déroulé de leur parcours de santé.

D'autres cas d'usage pourraient également contribuer à montrer comment utiliser le système national de données de santé comme support de comparaison des coûts des parcours de santé actuels, et à démontrer de quelle manière les innovations peuvent en réduire le coût tout en augmentant leur efficacité.

L'unicité de ce projet réside dans la capacité de ses participants à se focaliser sur ce qui importe : canaliser l'énergie des innovateurs sur des avancées technologiques utiles aux praticiens et faire en sorte que les patients puissent bénéficier de parcours de santé de plus en plus modernes et de plus en plus efficaces, dans le but qu'ils deviennent, demain, acteurs de leur propre parcours.

Vers un droit de propriété des données personnelles

Par Alain BENSOUSSAN
Avocat à la Cour, Lexing Alain Bensoussan Avocats

Les défis liés à « l'émergence permanente de nouvelles technologies et l'omniprésence des traitements de données à caractère personnel dans tous les champs de la vie », encore récemment soulignés par la présidente de la CNIL, Madame Marie-Laure Denis[1], placent plus que jamais la *data* au centre de toutes les attentions et, avec elle, la question de la propriété des données personnelles, de leur patrimonialisation et du droit à terme pour tout un chacun de monnayer ses propres informations[2].

Il n'a pas fallu attendre l'explosion de l'hyperconnexion pour que les entreprises, quelle que soit leur taille, comprennent l'avantage compétitif qu'elles pouvaient retirer du nouvel or noir que constituent les données à caractère personnel.

Acheter, réserver, gérer, lire, jouer ou avoir des amis sont autant d'usages d'Internet qui multiplient les occasions de transmission et de collecte de données personnelles.

Or, même lorsqu'ils ne sont pas de nature commerciale, ces nouveaux usages ont donné naissance à un véritable marché des informations personnelles.

Un commerce très lucratif qui s'exerce au détriment des internautes qui ne sont pas réellement avisés de l'utilisation qui est faite de leurs données personnelles sur le Web.

Les débats récents sur le consentement à l'utilisation des *cookies* et autres traceurs, qui collectent des données au motif d'une optimisation de la navigation et de l'expérience client, le démontrent à l'envi[3].

Tous connectés

N'étant pas clairement informés, les internautes ne sont pas conscients des risques qu'ils encourent en termes notamment de protection de leur vie privée et de leur identité numérique, deux notions « à contenu variable » qui témoignent de la fusion maintes fois annoncée du « monde réel et du monde numérique », laquelle va s'accélérer de façon exponentielle avec l'arrivée du Métavers[4].

Combien de temps allons-nous encore accepter d'abandonner nos nom, prénom, adresse, date de naissance, nous transformant ainsi en consommateur-producteur de données au nom du profilage commercial devenu inhérent à l'IoT qui nous entoure ?

Comment s'assurer que les données que nous confions à des sites marchands ne seront pas cédées à des tiers ?

Une chose est certaine : ce phénomène débouchera immanquablement sur une monétisation de ces données, de sorte que l'évolution devra tôt ou tard conduire à passer d'un droit à la protection à un droit à la propriété des données personnelles.

Vide juridique

Qui est propriétaire des données personnelles ?

À ce jour, personne ne l'est, ni l'individu auquel se rapportent les données ni l'entreprise qui les traite : il n'existe en réalité pas de loi, en France comme ailleurs, qui consacre la notion de propriété des données.

Et pourtant, chacun d'entre nous peut déclarer : « C'est mon nom, mon prénom, ce sont donc mes données, mes informations », de sorte qu'il existe en l'espèce un indéniable rapport de possession, un rapport de propriété.

(1) CNIL, Rapport annuel 2021, La Documentation française, mai 2022, avant-propos.

(2) Voir BENSOUSSAN A. (2018), *Pour un droit de propriété et une monétisation des données personnelles*, le 28 février.

(3) BENSOUSSAN A. & AVIGNON C. (2022), *Cookies, traceurs et droit*, Lexing Editions, collection « Management juridique », juin.

(4) BENSOUSSAN A. (2022), « Le Métavers : quelles règles juridiques ? », *Chronique Au nom de la loi*, A News Sécurité, février.

Avant la loi « Lemaire » pour une République numérique (loi n°2016-1321 du 7 octobre 2016) et le Règlement général sur la protection des données personnelles (RGPD), il était clair que l'individu ne disposait pas, en l'état du droit applicable jusqu'alors, d'un droit de propriété sur les données à caractère personnel le concernant.

En effet, la protection des données personnelles, telle qu'elle était alors conçue par la loi du 6 janvier 1978 et par la directive 95/46/CE, n'était pas fondée sur une logique patrimoniale, mais sur une logique de droits attachés à l'individu.

Le Conseil d'État, de façon générale, se refusait à reconnaître un droit de propriété sur les données à caractère personnel, justifiant sa position par la crainte de « la fragilisation de toute la réglementation publique de l'utilisation des données personnelles », à laquelle la reconnaissance d'un tel droit conduirait. Il militait donc pour la consécration d'un droit de la personnalité, un droit attaché à l'individu, pour faire rempart à un droit de propriété sur les données.

La loi pour une République numérique (LRN) et le RGPD ont changé la donne

Désormais, est reconnu aux personnes dont les données sont traitées le pouvoir de disposer librement de celles-ci.

L'article 1er, alinéa 2, de la loi du 6 janvier 1978 modifiée prévoit aujourd'hui que « toute personne dispose du droit de décider et de contrôler les usages qui sont faits des données à caractère personnel la concernant, dans les conditions fixées par la présente loi. »

De même, l'article 20 du RGPD reconnaît désormais un « droit à la portabilité des données » au profit de la personne concernée.

Cette nouvelle prérogative de taille accordée aux individus concernés par un traitement portant sur leurs données leur accorde une apparente « appartenance-maîtrise » de celles-ci, qui se traduit en pratique par un amoindrissement de la protection dont bénéficient ces dernières, ce qui ne correspond pas à la notion de pleine et libre disposition telle que conçue dans le Code civil.

Ainsi, si certains droits sont reconnus à l'individu, ceux-ci sont finalement accordés sans qu'une position soit arrêtée concernant la propriété des données en question.

Les obstacles opérationnels ont ainsi pour origine ce « vide juridique » dans la législation.

Il paraît pourtant inéluctable de voir consacrer un jour un droit de propriété des données.

Nous défendons l'idée que chacun d'entre nous doit être le possesseur, l'archiviste, le procureur et le propriétaire de ses données personnelles.

Cela n'a rien à voir avec la monétarisation, par exemple, du corps humain. Il n'existe en effet aucun rapport entre l'approche biologique et l'approche numérique de cette question. Lorsque nous transmettons une information, nous n'en sommes pas pour autant dépossédée.

Monétarisation des données, NFT et micro-paiements

En la matière, l'individu devrait pouvoir monétiser ses propres données et devenir ainsi en quelque sorte le *trader* de leur exploitation, se transformant de fait en véritable maître de son identité informationnelle, tant biologique que numérique.

Cette monétarisation de la donnée passe nécessairement par la propriété de celle-ci et elle serait aujourd'hui très facile à mettre en place avec les NFC et les micro-paiements.

Une fois devenu propriétaire de ses données, chacun d'entre nous pourrait les contrôler et les monétiser, mais aussi les vendre, voire même les mettre sous licence. Et cela sans qu'il y ait forcément dépossession.

Il serait dès lors possible pour tout un chacun d'exercer un pouvoir politique, philosophique, éthique et financier sur son patrimoine informationnel, avec toujours comme règles essentielles à respecter : un droit de propriété sur les données toujours transmis à titre précaire et une propriété qui laisse intacts un droit de contrôle et un droit de repentir pour celui qui en est le détenteur, faisant que chacun d'entre nous conserve la pleine propriété de ses données pour lui-même et vis-à-vis d'autrui.

Le vide juridique actuel nous laisse toute latitude pour préparer cet avènement, et même d'en prendre l'initiative.

L'utilisation qui serait faite de nos données personnelles serait bien évidemment soumise aux différentes réglementations. Mais cet encadrement de leur utilisation permettrait aux entreprises d'envisager plus sereinement les échanges de ces données.

En réalité, ce « no man's land » est une opportunité pour nos entreprises, dans la mesure où il doit permettre de structurer ce marché de la monétisation des données.

Ce droit de la propriété est à écrire ; il revient aux entreprises d'en prendre l'initiative.

Quel droit pour les données dans une économie datacentrique ?

Par Bertrand WARUSFEL
Professeur à l'Université Paris 8, avocat au barreau de Paris (cabinet FWPA)

Dans une économie dont les innovations et la productivité reposent largement sur la production, l'échange et le traitement d'informations numérisées, les données acquièrent une valeur économique, sociale et politique croissante. Mais le corollaire de cette affirmation est que, comme toute valeur, la « data » fait l'objet d'une concurrence extrême et provoque à la fois des litiges et une demande de régulation européenne, voire internationale.
On voit donc émerger progressivement un droit des données dont nous voulons ici rendre compte synthétiquement. Ce cadre juridique reste assez hétérogène et très partiel. Mais les effets très importants (et sans doute assez perturbants) de l'exploitation et du traitement algorithmique des données dans les années à venir devraient servir d'accélérateur à la structuration de ce droit, à condition que des choix politiques clairs viennent préciser les valeurs essentielles qu'une économie numérisée doit respecter.

Différents motifs de protection des données

On ramène parfois le droit des données à la seule protection légale des données à caractère personnel, laquelle a été instaurée en France par la loi Informatique et libertés du 6 janvier 1978 et établie aujourd'hui dans toute l'Union européenne par le non moins célèbre RGPD (Règlement général de protection des données) de 2016[1]. Mais d'autres aspects ne doivent pas être négligés et peuvent reposer sur des fondements différents.

La protection des données personnelles

Le droit des données personnelles a été la première construction juridique autonome entièrement dédiée à régir les activités de traitement de l'information. Alors que dans d'autres domaines (comme les contrats, la propriété intellectuelle ou le commerce en ligne), le droit de l'informatique a préféré se construire à partir des dispositions de droit commun et de leur adaptation aux particularités du numérique, on a inventé là, de toutes pièces, un dispositif protégeant la vie privée des citoyens face aux moyens de collecte et de traitement de plus en plus puissants que la technologie numérique offre.

Pour ce faire, chaque citoyen s'est vu reconnaître un droit propre sur le traitement numérique de ses informations personnelles. Ce droit se traduit par autant d'obligations que les « responsables de traitement » (c'est-à-dire, les entités qui collectent et traitent ces données) doivent respecter, à commencer par en assurer la confidentialité et ne pas en faire une exploitation pour d'autres finalités que celle ayant justifié leur recueil ; le tout sous le contrôle d'organismes indépendants de contrôle dans chaque État européen (comme la CNIL, en France).

La protection des données professionnelles

Il existe d'autres types de données dont le traitement est juridiquement encadré. Pour s'en tenir au droit de l'Union européenne, on peut relever la directive de 1996[2] qui est venue préciser les prérogatives que peuvent revendiquer ceux qui investissent dans la constitution d'une base de données, de manière à ce que toute personne qui accède à une telle base (par exemple, à travers un service en ligne) ne puisse pas la reproduire à l'identique ou en extraire des parties substantielles. L'objectif avoué a été de préserver ainsi les entreprises européennes contre les possibilités de « parasitisme » de leur patrimoine immatériel.

Pour protéger également les entreprises européennes contre d'autres formes de détournement de leurs actifs immatériels, l'Union européenne a aussi établi en 2016 une protection juridique spéciale couvrant les différentes formes du « secret des affaires »[3]. Toute information d'une entreprise, de quelle que nature qu'elle soit (technique, commerciale, financière, stratégique…) et qui a été conservée confidentielle, peut en bénéficier pour autant que l'entreprise puisse prouver que cette information garde une valeur commerciale du fait de son caractère secret et qu'elle a fait l'objet

(1) Règlement général de protection des données 2016/279 du 27 avril 2016.

(2) Directive 96/9/CE du 11 mars 1996.

(3) Directive 2016/943 du 8 juin 2016.

« de mesures de protection raisonnables, compte tenu des circonstances, pour en conserver le caractère secret »[4]. Du fait de la dématérialisation en cours du fonctionnement des entreprises, la plus grande partie des secrets d'affaires sont accessibles sous la forme de données numériques, notamment tout ce qui va toucher aux algorithmes ou aux données d'apprentissage mises en œuvre par l'entreprise pour fournir son service à ses clients.

D'autres motifs particuliers de protection des données

Pour être complet, il faut rappeler que d'autres données peuvent également couvrir des informations protégées dans des domaines sectoriels particulièrement sensibles. C'est ainsi le cas des données portant sur des informations classifiées en matière de Défense (qui touchent à la sécurité nationale des États), des données de santé ou encore de toutes les données produites ou échangées avec leurs clients par des professionnels astreints au respect du secret professionnel (médecins et avocats, pour n'évoquer qu'eux). Enfin, lorsqu'un ensemble de données numériques exprime une innovation technique ou une création originale, lesdites données deviennent alors le support de droits de propriété intellectuelle, dont le titulaire peut faire un usage exclusif et dont il peut même, sous certaines limites, restreindre l'accès.

Mais à cette forte tendance à la protection des données sous toutes leurs formes et pour différents motifs, s'oppose une orientation assez orthogonale qui pousse, au contraire, à l'ouverture et au partage des données. Cette tension juridique et politique entre protection et ouverture n'est pas en soi nouvelle, elle anime notamment tout le droit de la propriété intellectuelle (lequel protège les droits des titulaires tout en organisant les conditions de la circulation des œuvres). Mais en matière de données numériques, elle n'est qu'un aspect d'un ensemble de contradictions qu'il faudra bien un jour trancher et arbitrer.

Des logiques contradictoires entre maîtrise et ouverture des données

Consciente de l'importance croissante de toutes ces données et de la tension que leur appropriation pourrait susciter, l'Union européenne a déjà ouvert le chantier de ce qu'elle a d'abord dénommée pudiquement « les données à caractère non personnel » et de leur partage au sein de l'espace européen. Elle veut aujourd'hui aller plus loin et établir une véritable liberté de circulation de toutes les données dans l'Union, poursuivant l'ouverture déjà effective sur un autre terrain, celui des données publiques ; au risque toutefois d'accroître indirectement les injonctions contradictoires auxquelles la politique des données en Europe doit faire face.

De l'ouverture des données publiques à la libre circulation de toutes les données numériques

C'est dans un objectif de stimulation de la croissance et de défiance à l'encontre des administrations publiques (soupçonnées d'abuser de leur pouvoir informationnel pour fausser la concurrence) que l'Union européenne a adopté en 2003 une directive sur « la réutilisation des informations du secteur public »[5]. Les gisements d'informations aux mains des administrations doivent être mis à disposition des entreprises privées pour leur permettre de développer de nouveaux produits ou services. Les « jeux de données » concernés doivent être mis en ligne sous une forme anonymisée sur des serveurs (comme le site data.gouv.fr français) afin que chacun puisse en extraire la valeur intrinsèque, y compris en y appliquant des traitements algorithmiques et en les croisant avec d'autres données.

Évidemment, cette politique de l'*open data* se retrouve souvent en conflit avec différentes autres limites juridiques, que ce soit la préservation de la vie privée des citoyens, la protection des secrets protégés par la loi (comme évoqué *supra*) ou certains aspects du droit de la propriété intellectuelle[6]. Un nouveau texte dénommé Data Gouvernance Act devrait d'ailleurs être prochainement adopté au niveau européen pour mieux organiser la compatibilité entre l'ouverture des données et le respect des secrets protégés ou des droits intellectuels des tiers.

Poursuivant plus largement un objectif d'ouverture de toutes les formes de données non personnelles, une directive peu connue de 2018 a défini quelques grandes lignes du cadre juridique applicable au « libre flux » des « données à caractère non personnel » au sein de l'Union européenne[7]. Mais le véritable véhicule législatif en la matière n'est encore qu'un projet, celui du "Data Act" qui vient de faire l'objet d'un accord politique à Bruxelles et dont le principe serait d'établir une nouvelle liberté de circulation en Europe, celle des données après celle des biens, des capitaux, des services et des personnes. Sont visés, en particulier, les flux de données techniques que vont produire les équipements connectés et l'IoT. Mais là aussi l'exercice ne sera pas sans complexité, tout d'abord en ce qui concerne la distinction entre données personnelles et données non personnelles.

[4] Article 2(1) de la directive du 8 juin 2916 précitée, repris en France dans l'article L.151-1 du Code de commerce.

[5] Directive 2003/98 du 17 novembre 2003 concernant la réutilisation des informations du secteur public (aujourd'hui modifiée par la directive 2013/37 du 26 juin 2013).

[6] Voir, notamment, WARUSFEL B. (2020), « Numérisation de l'action publique et *open data* : une révolution face à ses limites », *Propriétés intellectuelles*, n°75, avril ; et WARUSFEL B. (2018), « Enjeux et limites de l'ouverture des données en matière de sécurité et de Défense », *Revue française d'administration publique*, 2018/3, n°167, pp. 551-564.

[7] Directive 2018/1807 du 14 novembre 2018 établissant un cadre applicable au libre flux des données à caractère non personnel dans l'Union européenne.

D'autres contradictions à dépasser

Au carrefour des différentes normes juridiques qui s'appliquent de manière hétérogène, la donnée numérique est l'objet de logiques difficiles à concilier, voire contradictoires. Nous illustrerons ici quelques-unes des plus caractéristiques.

La première concerne la collecte des données personnelles. Comme nous l'avons déjà relevé plus haut, les réglementations françaises et européennes en matière de protection des données ont été établies pour que la personne concernée puisse échanger son consentement à la communication de ses données personnelles contre l'engagement du responsable du traitement d'en respecter la finalité et la confidentialité, et ainsi sa vie privée. Mais depuis la généralisation de l'usage d'Internet et le développement massif des communications numériques mobiles, ce n'est plus du tout ainsi que s'effectue la collecte : de ponctuelle et volontaire, elle est devenue permanente et occulte. Nul ne sait en effet la nature et le volume des données qui sont récupérées lors de chaque transaction avec un service en ligne, ni même le plus souvent qui sont le ou les prestataires de services qui en assurent la captation, le stockage, puis l'exploitation. Dès lors, derrière le paravent largement symbolique de l'autorisation préalable des *cookies* lors de la connexion à un site Web, se cache une réalité beaucoup plus cynique : celui qui souhaite bénéficier des avantages technologiques liés aux services en ligne (ou qui y est contraint) perd tout à la fois son droit à l'information et au respect de la finalité initiale de la collecte ainsi que son droit à s'opposer à (ou du moins à limiter) l'exploitation de ses propres données.

La deuxième difficulté, qui est plutôt d'ordre géo-économique, tient au fait que les États-Unis, qui sont le principal centre du cyberespace contemporain, ont jusqu'ici toujours refusé d'adopter un mécanisme de protection en ligne de la vie privée équivalent au droit européen des données personnelles. Et, là encore, le décalage entre la théorie et la pratique est flagrant : malgré la réitération par la Cour de justice de Luxembourg de sa condamnation des conditions de transfert des données personnelles vers les États-Unis[8], la collecte massive d'informations réalisée par les grandes plateformes sur les internautes européens n'a pas cessé et entraîne toujours un flux constant de données européennes à destination des États-Unis en dépit de leur refus de se doter d'un dispositif de protection équivalent au RGPD.

Conséquence indirecte de ce désaccord transatlantique profond (qui concerne aussi d'autres aspects du droit européen, comme celui des bases de données), une certaine partie de la doctrine anglosaxonne cherche à promouvoir un modèle d'appropriation privative des données personnelles (lesquelles deviendraient la propriété de chaque usager). Ce n'est pourtant pas une approche adaptée à un objet numérique qui doit à la fois demeurer en permanence sous la maîtrise de son producteur originel (puisqu'il s'agit de données qui lui sont personnellement attachées) et pouvoir être malgré tout partagé et transmis à de multiples interlocuteurs. Le véritable droit de propriété ne peut porter, au contraire, que sur une valeur dont le titulaire initial peut se défaire complètement et dont la possession peut être exclusive. La reconnaissance d'une telle propriété sur les données personnelles n'aurait pour effet que de marchandiser totalement ces données et de saper les bases théoriques du modèle européen, lequel repose sur la reconnaissance d'un simple droit de la personnalité qui ne comporte en lui-même aucune dimension patrimoniale directe.

Plus paradoxal, on peut évoquer le dilemme autour de l'usage des différentes techniques cryptographiques. Schématiquement, deux types de fonctions peuvent être mis en œuvre à partir de ces technologies de sécurité : l'une vise à l'authentification et à l'intégrité des données (notamment par des mécanismes de type signature électronique ou *blockchain*), tandis que l'autre permet d'en assurer la confidentialité par le chiffrement. Sur le principe, ces deux approches paraissent complémentaires, mais, en réalité, on s'aperçoit que deux logiques antagonistes s'opposent en arrière-plan : celle de la protection privative des données par les entreprises et les citoyens et celle des politiques sécuritaires publiques.

D'un côté, l'authentification numérique renforce certes la confiance dans les transactions électroniques et la prévention des cyberattaques, mais elle aboutit aussi logiquement à accroître la traçabilité des échanges et des personnes, ce qui peut être problématique pour le respect de la vie privée de celles-ci. De l'autre, et inversement, le recours au chiffrement par les fournisseurs de services numériques et les usagers protège efficacement leurs secrets, qu'ils soient privés ou professionnels, mais peut aussi réduire les capacités légales d'interception ou d'investigation des services des États (qu'il s'agisse de la police judiciaire ou des services de renseignement).

Une nouvelle manifestation de ces contradictions entre protection des données privées et prérogatives de la puissance publique a éclaté au grand jour avec deux arrêts de la Cour de justice de l'Union européenne d'octobre 2020[9] et d'avril 2022[10]. Faisant prévaloir le droit des données personnelles et la garantie des droits fondamentaux, les juges européens ont réduit la possibilité pour les services de renseignement d'accéder rétroactivement aux données de connexion conservées par les opérateurs de communication électronique au seul cas où la « sécurité nationale » est en jeu (ce qui couvre notamment la lutte contre le terrorisme ou l'espionnage). En revanche, la Cour a purement et simplement interdit d'utiliser ces mêmes données pour la seule lutte contre le crime organisé, estimant qu'il s'agissait d'une menace de moindre niveau qui ne justifiait pas une telle atteinte aux données personnelles des citoyens.

[8] Arrêts CJUE, 6 octobre 2015, aff. n°C-362/14 (dit « Schrems 1 ») et CJUE, 16 juillet 2020, aff. n°C311/18 (« Schrems 2 »).

[9] CJUE, arrêt La Quadrature du Net, 6 octobre 2020, aff. n°C-511/18 et autres.

[10] CJUE, arrêt Dwyer, 5 avril 2022, aff. n°C-140/20.

Conclusion : une cohérence à construire

L'adoption à venir du futur *Data Act* européen pourrait permettre de franchir symboliquement un pas significatif en donnant pour la première fois une définition – somme toute large et assez neutre – de la donnée (et non plus de certaines sous-catégories de celle-ci), qui recouvrirait « toute représentation numérique d'actes, de faits ou d'informations et toute compilation de ces actes, faits ou informations, y compris sous forme d'enregistrement sonore, visuel ou audiovisuel »[(11)].

Mais une simple définition ne suffit pas, encore faudra-t-il choisir et hiérarchiser les valeurs de principe qui devraient constituer les fondements du droit des données. Schématiquement, on peut en identifier quatre qui s'avèrent essentielles : le droit de contrôler l'usage de ses données ; puis celui de pouvoir en retirer de la valeur et d'en négocier contractuellement l'accès ou l'exploitation ; ensuite, le droit de les protéger techniquement et juridiquement contre ce qui peut affecter leur intégrité ou leur confidentialité ; et, enfin, la nécessité de faire prévaloir les droits de la personne sur tout traitement qui porterait atteinte à l'individu et à ses libertés fondamentales.

En effectuant de tels choix et en les articulant dans un cadre juridique cohérent, la décision politique pourrait permettre au modèle européen de régulation d'une économie datacentrique de devenir une base de négociation et d'harmonisation au plan international.

[(11)] Article 2(1) du projet de règlement « on harmonised rules on fair access to and use of data » (Data Act), 23 février 2002.

Pilotage par les données et l'IA en santé : faire vivre l'écosystème de la « Garantie humaine » MedTech et HealthTech !

Par David GRUSON
Directeur du programme Santé Luminess et cofondateur d'Ethik-IA

PariSanté Campus est le nouvel écosystème majeur de la santé numérique en France et au plan international. Il inscrit dans ses orientations prioritaires le développement d'une filière Healthtech et Medtech forte, centrée sur le pilotage par les données et l'intelligence artificielle (IA) en santé. L'attractivité de la France et de l'Europe passera non seulement par le déploiement de technologies et de méthodologies innovantes sur ces thématiques, mais également par un engagement conséquent dans une dynamique forte de régulation éthique positive centrée sur le nouveau principe de « Garantie humaine » de l'IA.

La diffusion rapide du pilotage par les données et de l'intelligence artificielle (IA) en santé représente un levier essentiel de modernisation pour nos systèmes de santé. Pour autant, cet esprit d'ouverture doit s'accompagner d'une régulation des enjeux éthiques. C'est le sens que revêt le principe d'une garantie humaine de l'intelligence artificielle en santé qui fait l'objet d'une reconnaissance de plus en plus forte aux niveaux national (dans le cadre de la loi de bioéthique) et international.

Les pouvoirs publics encouragent très fortement cette impulsion donnée au développement du *data management* en santé :

- constitution du Health Data Hub et lancement d'appels à projets stratégiques sur le développement de l'IA médicale ;
- engagement d'un « Grand Défi » IA en santé[1] dans le cadre du programme d'investissements d'avenir ;
- déploiement d'actions pilotes de BpiFrance ;
- identification d'un chantier prioritaire sur la définition de référentiels de régulation éthique des dispositifs médicaux avec IA dans le cadre de la Feuille de route de transformation numérique du système de santé mise en œuvre depuis 2018 par Dominique Pon et Laura Létourneau.

Les perspectives susceptibles d'être ouvertes par la vague actuelle de diffusion de l'IA et de la robotisation en santé sont très significatives.

Le déploiement de programmes de digitalisation des données médicales constitue une étape très importante sur la voie de la valorisation des données de santé. C'est le sens du programme NUMIA de Luminess qui intègre le déploiement d'un dispositif de pré-qualification des données de santé dans une logique de remontée progressive de la chaîne de valeur du *data management* en santé.

Concrètement, Luminess, l'opérateur du programme de numérisation des dossiers médicaux, élaborera, après cette numérisation, un segment de pré-qualification des données de santé, en articulation étroite avec le dossier patient informatisé (DPI) des CHU concernés.

Luminess apporte un appui concret à chacun des CHU dans cette dynamique de valorisation en mobilisant sa technologie d'IA-OCR et d'IA-reconnaissance de langage naturel dans trois premiers champs de spécialités identifiés comme prioritaires par chaque CHU. Les réflexions préalables exprimées par le CHU de Guadeloupe – au sein duquel cette solution a été prototypée – ont permis d'identifier des domaines de valeur (en termes de recherche, d'amélioration de la qualité de la prise en charge des patients et de développement technologique) dans les champs de l'ORL, de l'urologie, de la prise en charge de la drépanocytose[2] et de la neurologie. Ces domaines prioritaires pourront bien sûr être complétés par d'autres au fil de la mise en œuvre du projet médical et du programme de numérisation des dossiers médicaux.

[1] https://www.intelligence-artificielle.gouv.fr/fr/secteurs-prioritaires/diagnostic-medical/grand-defi-sante

[2] Une maladie génétique héréditaire touchant les globules rouges, https://www.inserm.fr/dossier/drepanocytose/

D'un point de vue méthodologique, la stratégie IA s'appuie sur :

• un accompagnement stratégique et technique du CHU dans l'identification de ses priorités de développement de l'IA en santé et des capacités techniques de mise en œuvre afférentes ;
• un accompagnement du CHU en matière d'activation du volet « Régulation positive éthique ».

Ce programme a été intégré comme initiative pilote au sein de PariSanté Campus dans le cadre de la sélection des activités Santé de Luminess s'inscrivant dans ce nouvel écosystème majeur de la santé numérique. L'engagement de Luminess au sein de PariSanté Campus vise, en particulier, à :

• apporter un appui aux établissements de santé dans le repérage stratégique de leurs champs spécifiques de valeur en matière d'intelligence artificielle. Cet accompagnement permet ainsi d'établir, pour les différents pôles d'activité d'un CHU, une cartographie dynamique des axes de développement prioritaires dans le domaine de l'intelligence artificielle ;
• aider les CHU à constituer leur *task-force* IA : rassemblant les acteurs clés de l'établissement de santé en la matière, elle a pour mission, sur mandat du directoire, de piloter la démarche IA et de faciliter la conduite opérationnelle des projets ;
• venir en soutien des établissements de santé dans la construction d'un modèle économique soutenable reposant sur le développement du pilotage par les données et de l'IA en santé, en portant une attention spécifique aux modes de financement innovants (appels à projets du Health Data Hub et du Grant projet IA médicale, article 51 de la LFSS, financements de parcours dans le cadre de la prise en charge des maladies chroniques métaboliques…) ;
• mettre en œuvre une analyse des capacités techniques des établissements de santé pour les appuyer dans la mise en place de cette politique de développement de l'IA. Cette démarche porte sur l'évaluation des capacités de digitalisation des données médicales et d'hébergement, ainsi que de celles de pré-qualification et de valorisation de ces données. Elle permet d'explorer les possibilités de mise en œuvre de ces capacités par l'établissement de santé lui-même et d'identifier les champs d'articulation possibles à l'échelle du groupement hospitalier de territoire (GHT), au niveau régional et au niveau national en lien avec les acteurs clés (Health Data Hub, Délégation au numérique en santé, CNAM…).

Cet écosystème MedTech et Healtech a vocation à se déployer au sein de PariSanté Campus dans un cadre éthique et juridique certes exigeant, mais constituant également un fort soutien à l'innovation. Pour pouvoir se positionner comme des acteurs porteurs de la nouvelle approche de régulation positive en phase avec la nouvelle loi Bioéthique française[3], les parties prenantes de cet écosystème devront s'inscrire résolument dans la dynamique de la « Garantie humaine » de l'IA.

Le principe de « Garantie humaine » de l'IA et du numérique en santé, introduit et porté par Ethik-IA depuis 2017, a été reconnu dans les avis 129 et 130 du CCNE et dans l'article 11 du projet de loi Bioéthique[4], lequel est devenu l'article 17 de la loi adoptée, qui est entrée en vigueur en août 2021. Le principe de « Garantie humaine » de l'IA, qui découle de cet article 17, impose à tous les utilisateurs et les concepteurs de solutions d'IA en santé :

• la mise en place d'une information préalable du patient sur le recours à l'IA dans sa prise en charge ;
• le déploiement d'une supervision humaine de la solution d'IA « en vie réelle », dans le respect de conditions de traçabilité mises en œuvre sous le contrôle et la supervision de la CNIL et de la Haute Autorité de santé (HAS).

Le concept de « Garantie humaine » peut paraître abstrait, mais il est, en réalité, très opérationnel. Le débat sur le recours au numérique pour faire face à la pandémie de Covid-19 – avec, en particulier, la question du *data tracking* – montre toute la nécessité d'une mise en application immédiate de ce principe. L'idée est d'appliquer les principes de régulation du numérique et de l'intelligence artificielle en amont et en aval de l'algorithme lui-même en établissant des points de supervision humaine, non pas à chaque étape, sinon l'innovation serait bloquée, mais sur des points critiques identifiés dans le cadre d'un dialogue partagé entre les professionnels, les patients et les concepteurs d'innovation.

La supervision peut s'exercer à travers le déploiement de « collèges de garantie humaine ».

Il est à relever que le principe de garantie humaine a reçu, au cours des années 2020 et 2021, des concrétisations dans trois cadres très significatifs autres que celui de la révision de la loi Bioéthique :

• tout d'abord, la « Garantie humaine » a été intégrée dans la grille d'auto-évaluation des dispositifs médicaux recourant à l'IA, une évaluation intervenant préalablement à leur admission au cadre de remboursement publié en octobre 2020 par la HAS ;
• ensuite, le principe de « Garantie humaine » est inséré dans les recommandations de l'OMS sur l'éthique et la gouvernance de l'intelligence artificielle en santé du 28 juin 2021 ;
• enfin, le principe a été repris dans le Livre blanc sur l'IA publié par la Commission européenne, le 19 février 2020.

Dans la nouvelle loi de Bioéthique, et comme indiqué précédemment, les dispositions concernées figurent à l'article 17 de cette loi. Les mots « Garantie humaine » ne figurent pas dans le texte lui-même, mais sont très directement repris dans l'exposé des motifs du projet de

[3] Loi du 2 août 2021, dont l'article 17 détermine le nouveau cadre juridique applicable à l'intelligence artificielle en santé.

[4] Le principe de « Garantie humaine » figure en tant que tel dans l'exposé des motifs et l'étude d'impact du projet de loi. L'article 11 a été adopté par l'Assemblée nationale et le Sénat dans des termes rédactionnels différents, mais en reprenant les mêmes principes : information préalable du patient sur le recours au numérique, s'inscrivant dans le cadre du recueil – obligatoire – de son consentement et de la mise en œuvre d'une supervision humaine du numérique et de l'IA en santé.

loi[5] et dans l'étude d'impact de celui-ci[6]. Ce positionnement correspond à la distinction que l'on peut faire entre un principe éthique général et ses concrétisations opérationnelles en droit positif, qui se retrouvent, quant à elles, directement mentionnées dans le corps des dispositions législatives.

S'agissant des dispositions législatives elles-mêmes, nous retrouvons la double dimension de la « Garantie humaine » recommandée par l'avis 129 du CCNE[7] :

• concernant le patient, un nouveau devoir d'information sur le recours à l'IA dans la prise en charge est reconnu aux fins de permettre, dans toute la mesure du possible, un consentement libre et éclairé au protocole de soins incorporant un traitement algorithmique : « Le professionnel de santé qui décide d'utiliser, pour un acte de prévention, de diagnostic ou de soin, un dispositif médical comportant un traitement de données algorithmique, dont l'apprentissage a été réalisé à partir de données massives, s'assure que la personne concernée en a été informée et qu'elle est, le cas échéant, avertie de l'interprétation qui en résulte » ;
• concernant la supervision humaine dans la conception algorithmique du traitement et dans son application en vie réelle, les termes utilisés par le législateur sont moins précisément définis. Néanmoins, si le principe général n'est pas formulé en tant que tel, l'article 17 impose une série d'obligations nouvelles traduisant effectivement la mise en œuvre d'éléments de supervision humaine : d'une part, « les professionnels de santé concernés sont informés du recours à ce traitement de données. Les données du patient utilisées dans ce traitement et les résultats qui en sont issus leur sont accessibles » ; d'autre part, « Les concepteurs d'un traitement algorithmique […] s'assurent de l'explicabilité de son fonctionnement pour les utilisateurs. »

Le principe de « Garantie humaine » de l'IA (*Human Oversight*) est également introduit dans l'article 14 du projet de règlement sur l'intelligence artificielle de la Commission européenne, qui, diffusé le 21 avril 2021, est en phase de finalisation dans le cadre de la présidence française de l'Union européenne (PFUE).

Cet article 14, qui institue ce principe de « Garantie humaine », se situe dans le droit fil des démarches pilotes et des recommandations précitées.

Le paragraphe 1 de cet article énonce ainsi que les solutions d'intelligence artificielle doivent être conçues et développées de façon à pouvoir être supervisées par des humains.

Le paragraphe suivant précise que la supervision humaine permettra de prévenir ou de minimiser les risques pour la santé, la sécurité ou les droits fondamentaux pouvant émerger d'un système d'IA susceptible de présenter un niveau de risque élevé. Par cet énoncé, le projet de règlement consacre ainsi la nécessité d'une garantie humaine pour un déploiement éthique de l'IA.

Le paragraphe 3 donne, quant à lui, des indications sur la mise en application de la supervision humaine de l'IA. En effet, la garantie humaine doit être identifiée et construite par le fournisseur avant la mise sur le marché ou la mise en service du traitement algorithmique, et/ou elle doit être identifiée par le fournisseur et pouvoir être mise en œuvre par l'utilisateur, et ce toujours en amont de la mise sur le marché ou de la mise en service du traitement. Cette « Garantie humaine » doit pouvoir faire l'objet d'un suivi en vie réelle de l'intelligence artificielle.

Les mesures prévues à ce paragraphe 3 fixent un certain nombre d'objectifs d'information s'articulant autour de cette « Garantie humaine » : comprendre entièrement les capacités et les limites du système d'IA et être capable de surveiller l'opération de façon à ce que les risques d'anomalies, de dysfonctionnements et de performance inattendus puissent être détectés ; être conscient des risques liés aux IA d'aide à la décision ; être capable d'interpréter correctement le résultat de l'IA à haut risque et, si nécessaire, ne pas tenir compte de ce résultat ou le remplacer ; et, enfin, pouvoir interrompre le recours à l'IA à tout moment.

Finalement, on retrouve dans cet article 14 du projet de règlement européen les deux axes essentiels de l'article 17 et des méthodologies construites par Ethik-IA depuis 2017 dans le champ de la santé :

• l'information des utilisateurs d'une solution d'IA ;
• la supervision humaine de l'IA dans la phase de conception de cette dernière et, dans une logique d'amélioration continue de la qualité, dans son utilisation en vie réelle.

---

[5] L'article 17 « vise à sécuriser la bonne information du patient lorsqu'un traitement algorithmique de données massives ("intelligence artificielle") est utilisé à l'occasion d'un acte de soin. Il décline également la garantie d'une intervention humaine. »

[6] https://www.legifrance.gouv.fr/dossierlegislatif/JORFDOLE000038811571/

[7] https://www.ccne-ethique.fr/node/174?taxo=0

People in the Sun
Illuminez vos données

Par Charles HUOT
Société People in the Sun

Depuis l'avènement des grands moteurs de recherche qui ont démocratisé l'utilisation de l'informatique dans les foyers, l'idée de la possibilité de créer un jumeau numérique complet de la planète et de tous ses constituants s'est fait jour, commençant par la géographie du globe, puis zoomant au niveau d'un pays, d'une ville, d'une maison, d'une pièce et, pourquoi pas, d'un individu. Mais que nous faut-il comme données (*data*) pour modéliser cet individu et quelles sont celles dont nous disposons déjà ? Que savons-nous de ses habitudes, de ses besoins, de sa santé, de ses relations, de son comportement ? Pourrions-nous mettre en place des capteurs qui nous fourniraient ces données ? Disposons-nous des technologies nécessaires à la réalisation de cet objectif ? Un tel projet, s'il était réalisable pourrait-il être rentable ?
Et nous pourrions poursuivre ce questionnement à volonté. Mais à l'heure où les ressources se font rares, où la planète court un danger écologique sans précédent, nous nous sommes posé la question de l'utilité pour l'homme d'un tel projet. Où devons-nous poser des barrières ? Comment mettre en place une stratégie positive sur la création et la valorisation des *data* ? Nous aborderons de façon très synthétique dans l'article cette question sous les angles technologique, économique, juridique et éthique.

C'est aujourd'hui un truisme que de dire que les données (les *Data*) sont l'un des moteurs essentiels de l'économie. Le digital est présent à tous les échelons de la société et l'une de son expression physique est l'augmentation croissante (+ de 50 % par an) de la production de données et des capacités de stockage de celles-ci. Les données naissent et croissent à l'envi ; jamais elles ne meurent ; et plus on les consomme, plus elles s'enrichissent. À l'inverse de nos ressources naturelles, les données sont une ressource inépuisable qui irrigue les réseaux de communication du monde entier sous l'eau, sur terre, dans les airs et dans l'espace. Au verbe « irriguer », nous devrions peut-être substituer celui d'inonder ou bien de submerger tant le volume des données vient engorger nos systèmes d'information. Nous pouvons même parler de boulimie digitale.

Chez People in the Sun, nous accompagnons les organisations dans leurs réflexions sur l'usage de la donnée. Les interrogations sont multiples et couvrent des champs technologiques, économiques et juridiques, mais aussi éthiques. Chacun de ses champs aborde un aspect particulier de la donnée.

Nous nous proposons dans cet article de balayer ces différentes facettes de la problématique générale de la donnée et de nous interroger collectivement sur les risques et les opportunités que le traitement algorithmique massif des données génère au sein d'une société démocratique, et ce au regard de la liberté de nos concitoyens.

Champ technologique

Nous débuterons par le champ technologique. Il nous intéresse naturellement, car il est le premier maillon de la chaîne de vie de la donnée. La donnée se crée grâce au numérique. Elle est fille du digital. Elle porte en elle toute la variété du monde qu'elle représente, qu'elle mesure, qu'elle écoute, qu'elle filme ou qu'elle raconte. Ainsi, en fonction de la forme que cette donnée va prendre, nous disposons aujourd'hui d'une panoplie très large d'outils d'interprétation, de croisement et d'analyse sémantique de celle-ci pour en comprendre le sens et en faciliter l'interprétation. Formidables outils d'aide à la décision, la simulation numérique, l'intelligence artificielle et l'analyse massive des données se développent toujours davantage pour générer et interpréter toujours plus de données. C'est la force de la technologie informatique, mais c'est également sa limite. Ainsi, nous pouvons diviser les aspects organisationnels du champ technologique en trois grands axes, à savoir : la donnée elle-même, les outils nécessaires à sa gestion, sa protection et son analyse, et, enfin, le mode d'usage et de diffusion de ses résultats pour répondre à un besoin précis de l'organisation. Il est fondamental d'identifier ses trois piliers lors de nos missions d'accompagnement à la valorisation des données de l'entreprise. Nous devons nous poser des questions aussi simples que celles-ci : que souhaitons-nous savoir/étudier/analyser ou comprendre pour rendre plus efficace la marche de notre organisation ? De quelles données disposons-nous pour cela et sous quelle forme se

présentent-elles ? Quels composants informatiques devrons nous mettre en œuvre ? Quels matériels, quels logiciels ou quels services devons-nous mobiliser ? » Il s'agit d'un aspect extrêmement important dans les missions que nous réalisons pour nos clients (Huot, 2019 ; Henke, Bughin, Chui *et al*., 2016).

Champ économique

Sur le plan économique, l'une des bornes au développement parfois anarchique des données est le besoin réel de l'organisation en termes d'information. Au-delà du vieil adage, « Mettons cette donnée de côté, elle pourrait nous servir un jour », il est temps de nous poser la question « À quoi pourrait-elle vraiment nous servir ? » Ou pourrait-elle servir à d'autres ? Revêt-elle un véritable potentiel économique ?

Faisant le constat que le marché de la donnée n'existe pas encore en France, GFII[1], en relation avec Syntec Numérique, la FNPS[2] et Cap Digital, a décidé de favoriser le rapprochement des producteurs-diffuseurs de données publiques ou privées de leurs futurs clients réutilisateurs de ces données. Le nouvel événement organisé en ce sens et baptisé Data&business Day (GFII, 2020) se tient sur un rythme annuel et s'adresse à chaque fois à un secteur économique différent. Pour sa première édition, le choix s'est porté sur les producteurs-diffuseurs de données professionnelles dans le domaine de la « Ville intelligente et durable » afin qu'ils puissent présenter leurs catalogues de données déjà disponibles à titre gratuit ou payant. La seconde journée, qui s'est tenue en décembre 2021, à l'initiative de Cap Digital, du GFII et d'OpenDataFrance, a abordé la question des données liées aux enjeux de la RSE (données relatives aux facteurs environnementaux, sociétaux et de gouvernance).

Dans un article récent (Margerie, 2022) qui aborde la question de la consommation en eau et en énergie et des émissions de CO_2 liées au développement du numérique, l'auteur cite Luc Julia (Julia, 2019), directeur scientifique du groupe Renault, qui nous invite à nous interroger sur la question de la complexité algorithmique et revient encore une fois sur les volumes gigantesques de données créées en permanence par les organisations.

La société française Dawex[3], leader technologique dans le domaine de l'économie de l'échange de données, lançait, dès 2017, sa Global Data Marketplace, une place de marché pour échanger, acquérir, distribuer et commercialiser des données dans un environnement sécurisé, de confiance et traçable. Trois ans plus tard, selon une étude mondiale menée par le MIT Technology Review (Insights, 2020) auprès de 1 000 dirigeants et experts en intelligence artificielle (IA), « 66 % des entreprises sont disposées à partager leurs données en externe pour contribuer à la création de nouvelles solutions, de nouveaux produits ou de nouvelles chaînes de valeur grâce à l'intelligence artificielle. »

La voie vers la valorisation des données des entreprises et donc vers une meilleure organisation et gestion de cette ressource que sont les données est aujourd'hui ouverte. Et c'est dès à présent la tâche allouée dans les organisations aux Chief Data Officer (CDO)[4]. En effet, en créant ce poste au milieu des années 2010 pour répondre à l'augmentation exponentielle des données à leur disposition, certaines entreprises ont ainsi décidé de construire un pont entre les départements informatiques et les dirigeants. C'est donc le rôle du CDO.

Nous entrons ainsi dans un nouveau modèle économique, où le Return on Data (ROD), l'équivalent du calcul de retour sur investissement (ROI), devient un indicateur de référence.

Ainsi, comme pour toutes les autres ressources, les entreprises, les organisations ou les États cherchent à s'emparer de ces données et à les exploiter afin d'en tirer le meilleur profit. Et l'objet du champ juridique est justement de fixer un certain nombre de règles pour encadrer cette exploitation.

Champ juridique

Sur le volet juridique concernant les données, les organisations ont, là aussi, besoin d'un accompagnement. Depuis la loi française du 6 janvier 1978[5] relative à l'informatique, aux fichiers et aux libertés, plus connue sous le nom de loi Informatique et libertés et dont l'objet est de réglementer la liberté de traitement des données personnelles, et le décret du 29 mai 2019[6] pris pour l'application de la loi précitée et visant à la mise en conformité du droit national avec le Règlement général sur la protection des données (RGPD)[7], le traitement des données informatiques fait l'objet d'une constante attention des États démocratiques dans le but d'assurer la protection de la vie privée de leurs citoyens. La dimension juridique se rattachant aux données devient alors essentielle pour les organisations. Le risque de se trouver en infraction par rapport à la loi devient une menace suffisamment forte pour bloquer le développement d'un certain nombre de projets. Dans ce domaine, une nouvelle fonction voit le jour au sein des organisations, il s'agit du délégué à la protection des données (DPO)[8]. Les missions du DPO, décrites à l'article 39[9] du RGPD, consistent à « informer et conseiller son organisation sur les obligations lui incombant en vertu des réglementations sur

(1) Le Groupement français des industries de l'information.

(2) La Fédération nationale de la presse d'information spécialisée.

(3) www.dawex.com

(4) Chief Data Officer (CDO) : rôle, compétences, formations et salaire, https://www.lebigdata.fr/emplois-big-data/chief-data-officer

(5) La loi n°78-17 modifiée du 6 janvier 1978.

(6) Le décret n°2019-536 du 29 mai 2019.

(7) https://www.cnil.fr/fr/reglement-europeen-protection-donnees

(8) https://fr.wikipedia.org/wiki/Délégué_à_la_protection_des_données

(9) https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre4#Article39

la protection des données personnelles, à contrôler le bon respect par son organisation des réglementations sur la protection des données personnelles, à dispenser des conseils en ce qui concerne la réalisation d'analyses d'impact relative à la protection des données, à coopérer avec les autorités de contrôle, comme la CNIL en France, ainsi qu'à faire office de point de contact entre son organisation et les autorités de contrôle sur des questions relatives aux traitements de données personnelles mis en œuvre par son organisation. »

Sans opposer la liberté individuelle à la nécessité collective, un certain nombre de traitements informatiques existent afin d'anonymiser (Bera M., 2017) des données individuelles. Cette fonction, qui peut être combinée avec une mesure du risque de réidentification (Bera M. (ed.), 2020), permet dans des secteurs sensibles, comme celui de la santé, de mener des études épidémiologiques nécessaires à l'évaluation de la santé de la population, ou celui de l'énergie, de mieux calibrer la grille de distribution électrique.

Les contrôles de conformité par rapport au RGPD réalisés par les autorités un peu partout en Europe touchent aussi bien les hôpitaux que les compagnies aériennes, et les amendes infligées peuvent se chiffrer en plusieurs centaines de millions d'euros, comme cela a été le cas pour British Airways avec une amende de 200 millions d'euros (Guzman-Scola, 2019). Le rôle du DPO est déterminant pour éviter de telles situations.

Champ éthique

Mais au-delà des champs de la technologie, de l'économie et du juridique, il convient à présent de nous intéresser à la question éthique. Chez Cap Digital, nous avons développé une vision responsable de l'innovation numérique[10] en nous appuyant sur un travail collectif porté par 150 de nos adhérents. Notre démarche repose sur cinq principes :

- devenir le pôle de compétitivité référent et reconnu en matière d'innovation responsable ;
- agir en cohérence avec l'objectif assigné au pôle Compétitivité des entreprises ;
- associer innovation technologique responsable, progrès social et compétitivité ;
- accompagner les membres du pôle précité plutôt qu'avoir une position décalée de type militantiste sur le sujet ;
- co-construire avec les membres du pôle une vision responsable de l'innovation numérique.

Doit-on systématiquement opposer éthique et performance au niveau d'une organisation ? Nos adhérents pensent que non, et à une très large majorité. Dans un monde où les ressources se font rares, il convient d'adapter notre consommation en conséquence ; et le numérique n'échappe pas à la règle. L'intelligence artificielle, la simulation numérique, le traitement massif des données…, sont autant d'activités qui relèvent pour une entreprise de sa responsabilité sociale et environnementale.

À la question « Jusqu'où pouvons-nous aller dans l'utilisation des données pour analyser les comportements et les contextes de nos actions, pour répondre aux attentes de nos clients, des entreprises dans un contexte mondialisé ? », la réponse technique est : « Il n'y a pas de limites ». Les capacités des outils actuels d'analyse de l'information combinées à une disponibilité massive des données permettent toutes les dérives. Il revient donc à la société et à ses citoyens de débattre et définir les limites qu'il convient de fixer pour éviter de tels risques. Aujourd'hui, ce sujet est de notoriété publique, et n'est donc plus réservé au seul monde de l'informatique. Il prend jour après jour de plus en plus d'ampleur dans la société et les consciences s'éveillent à ces enjeux dans bien des pays.

En s'accordant sur le texte final du Digital Market Act (DMA), le 24 mars 2022, la Commission, le Parlement et le Conseil européens ont entendu rééquilibrer les marchés numériques face à l'hégémonie des *Big Tech*. Avec l'adoption de ce DMA, l'Europe sonne-t-elle la fin du « Far West numérique » ?[11].

L'économie du numérique et de la *data* est encore jeune. Elle doit poursuivre son développement dans le respect de l'individu et de l'environnement.

Bibliographie

BERA M. (2017), « *Big Data* et anonymisation », *Le CNAM Mag'* 8, septembre.

BERA M. (ed.) (2020), *AFNOR SPEC Z90-030 : la méthode QaR de mesure du risque extrême de réidentification d'une base de données dans le cadre de l'évaluation de son assurabilité,* AFNOR.

GFII (2020), synthèse de la première édition du Data & Business day sur les villes intelligentes et durables, Data & Business day, Paris.

GUZMAN-SCOLA N. (2019), *Top 10 des amendes du RGPD,* 22 novembre, récupéré sur Blog CYBERSECURITY, https://blog.httpcs.com/top-10-des-amendes-du-rgpd/

HENKE N., BUGHIN J., CHUI M. *et al.* (ed.) (2016), *The Age of Analytics: competing in a Data-Driven,* McKinsey Global Institute.

HUOT C. (2019), « La normalisation et le *Big Data* », *Enjeux numériques,* n°5, mars, pp. 56-60.

INSIGHTS (2020), "The global AI agenda: Promise, reality, and a future of data sharing", *MIT Technology Review Insights*.

JULIA L. (2019), *L'Intelligence artificielle n'existe pas,* First Edition.

MARGERIE V. (de) (2022), « *Good data* plutôt que *Big Data* », *Les Échos*.

[10] Cap Digital, PUBLICATION – Notre vision de l'Innovation numérique responsable https://www.capdigital.com/publication-notre-vision-de-linnovation-numerique-responsable/

[11] https://www.aege.fr/agenda/nocturne-du-big-data-517

Valorisation de la recherche en santé humaine et protection des données à l'ère du numérique

Par Frédérique LESAULNIER
Docteure en droit, déléguée à la protection des données de l'Institut du cerveau

La recherche en santé humaine connaît une révolution numérique du fait des masses considérables de données disponibles, qui sont collectées dans des environnements multiples, et de la possibilité d'en extraire des connaissances et des corrélations grâce aux technologies qui augmentent les capacités de stockage et de traitement. À l'heure des grands défis de la Science ouverte, le RGPD conduit à un recentrage de l'organisation de la protection des données sur les organismes qui les traitent. Le respect de la réglementation doit être intégré dans une démarche globale de gouvernance des données et requiert l'implication des personnes à l'origine des données. Nous présenterons ici quelques pistes de réflexion et d'action développées par les acteurs de la recherche pour une utilisation éthique et responsable des données personnelles de santé.

Qu'il s'agisse de données scientifiques issues du soin et du système de santé, de données médico-administratives recueillies initialement à des fins gestionnaires ou encore de celles issues de l'utilisation d'objets connectés et d'Internet, les données personnelles de santé sont un enjeu de premier plan pour la recherche en santé humaine. Leur exploitation permet chaque jour de nouvelles avancées dans la compréhension des maladies, des facteurs de risque, mais aussi une amélioration de la prise en charge et de la prévention.

Les chercheurs soulignent la nécessité de favoriser l'utilisation optimale de ressources précieuses produites dans l'intérêt général, qui ont nécessité une mise en qualité indispensable à leur exploitation pertinente. Ils relèvent qu'en l'absence de partage ou d'utilisation de ces données et/ou des échantillons biologiques, de nouvelles collectes devront être effectuées, engendrant de nouveaux risques pour les personnes, mais aussi des délais et des surcoûts. Ils insistent enfin sur l'importance de garantir le respect de la confiance accordée par les personnes concernées, en tirant de leur effort de participation tout le parti qu'elles pourraient souhaiter. C'est particulièrement vrai dans le cas de maladies rares, où la recherche scientifique fait naître de grands espoirs d'identifier les causes de ces maladies, ce qui répond à une attente des patients, notamment pour prévenir leur transmission, identifier les facteurs de sévérité et, surtout, trouver des traitements[1].

[1] Comme souligné dans l'avis 129 du Comité consultatif national d'éthique, « l'insuffisance du recours au numérique dans la prise en charge des patients, pour la recherche ou pour soutenir le développement du pilotage par les données, induit, sur une large échelle, des situations non éthiques au sein de notre système de santé. »

Cela suppose que ces données sensibles soient exploitées avec la plus grande rigueur, l'expertise et l'esprit critique nécessaires, le tout dans le respect du cadre éthique, légal et réglementaire. Et c'est encore plus vrai pour des informations de nature à renseigner sur l'état de santé futur des personnes, et de toute la lignée de leurs apparentées génétiques.

Les données personnelles de santé, parce qu'elles relèvent de l'intimité de la vie privée des personnes, sont des données qui doivent faire l'objet d'une protection particulière. À ce titre, le droit leur reconnaît un statut particulier et impose le respect de règles ayant pour objet de garantir leur confidentialité. Elles sont ainsi soumises à un principe d'interdiction de traitement sauf pour un certain nombre d'exceptions prévues par la loi et assorties de garanties (de fond et de procédure) au respect desquelles la Commission nationale de l'informatique et des libertés (CNIL) veille.

Les garanties de fond sont les suivantes : une finalité de traitement déterminée, explicite et légitime ; des données adéquates, pertinentes et proportionnées au regard de l'objectif poursuivi (principe de minimisation des données) ; une durée de conservation déterminée à l'avance et dont la pertinence est appréciée au regard de la finalité poursuivie (droit à l'oubli) ; le respect des droits des personnes qui passe en premier lieu par le principe de loyauté et de transparence à leur égard ; et, enfin, la mise en place de mesures de sécurité de nature à garantir la confidentialité des données, qui est une composante majeure de la conformité s'agissant du traitement de données de santé.

Les spécificités du *Big Data* et de l'intelligence artificielle sont souvent présentées comme susceptibles de questionner certaines notions clés ou principes

cardinaux de la protection des données[2]. Ainsi, dans le contexte du *Big Data* et compte tenu des moyens techniques disponibles, l'anonymat devient relatif et des données réputées neutres peuvent devenir sensibles par recoupement. On est donc tenté de considérer la donnée comme personnelle à défaut de preuve contraire[3] et de raisonner en termes de genre plutôt que d'espèce. Avec la médecine personnalisée, la ligne de partage entre le soin et la recherche tend à perdre de son évidence : ainsi, l'objectif des instituts hospitalo-universitaires est de favoriser le développement d'une recherche translationnelle qui bénéficie directement au patient ; les législations « Informatique et libertés » raisonnent, quant à elles, par type de finalités. Par ailleurs, le contrôle de la pertinence des données est présenté comme difficilement compatible avec l'approche exploratoire inhérente au *Big Data.* De même, la limitation de la durée de conservation des données est difficilement compatible avec la constitution de vastes bases de données ayant vocation à être ouvertes à la communauté scientifique. Enfin, les formalités administratives préalables et l'information individuelle dont doivent bénéficier les personnes avant chaque étude, consacrées dans l'optique de recherches ponctuelles, sont difficilement compatibles avec les extractions réitérées et standardisées de lots de données très indirectement identifiantes. Il est donc nécessaire que les principes puissent être interprétés avec pragmatisme et souplesse.

Plusieurs dispositions du Règlement général relatif à la protection des données (RGPD) témoignent d'une prise en considération des enjeux de la recherche scientifique. Ce Règlement est devenu le texte de référence depuis le 25 mai 2018 dans l'ensemble de l'Union européenne et même au-delà, dès lors que l'on cible des résidents européens. Il maintient une dérogation au principe d'interdiction de traiter certaines catégories particulières de données, instaurée au bénéfice de la recherche scientifique, moyennant un certain nombre de garanties mentionnées à l'article 89 (art. 9-2-j). Il pose, notamment, le principe d'une présomption de compatibilité de la finalité de la recherche scientifique avec une finalité initiale différente et prévoit la possibilité de la conservation des données au-delà de la réalisation de la finalité initiale à des fins de recherche scientifique.

À l'heure des grands défis de la science ouverte, le RGPD modifie l'approche de la protection des données personnelles en renversant la charge de la preuve. Il opère un recentrage de l'organisation de la protection des données sur les organismes qui les traitent et/ou les mettent à disposition. C'est à eux qu'il appartient de prendre les mesures techniques et organisationnelles garantissant le respect des grands principes de protection des données et ils doivent être en mesure de démontrer que tel est bien le cas (*accountability*).

Le respect de cette réglementation n'est pas uniquement une question de conformité à la législation en vigueur, même si cette dimension est essentielle. Les bonnes pratiques induites dans la collecte, le traitement, le stockage et la diffusion des données peuvent concourir à l'amélioration de la science elle-même et favoriser la confiance des personnes qui concourent à son progrès. C'est pourquoi les organismes de recherche développent des écosystèmes facilitant l'accès et l'exploitation des données dans le respect de la vie privée et de la confidentialité des données personnelles et des libertés individuelles, avec toute la rigueur, le bon sens et l'éthique indispensables à la confiance dans le numérique.

Nous interrogerons ici les valeurs à l'origine de cette protection et présenterons quelques pistes de réflexion et d'action développées par les acteurs de la recherche pour une utilisation éthique et responsable des données personnelles de santé.

La protection de la valeur éminente de l'être humain

Il existe un cadre juridique riche et ancien qui définit les conditions d'accès et d'utilisation des données personnelles de santé et qui traduit le caractère sensible de celles-ci.

Il repose en Europe sur le principe selon lequel la protection des données à caractère personnel est un droit fondamental inscrit dans la loi (Charte des droits fondamentaux de l'Union européenne, art. 8 §1).

L'application de ce cadre juridique est subordonnée à l'existence de données susceptibles de permettre d'identifier la personne, que cette identification soit directe ou indirecte par référence à un identifiant ou tout élément qui soit propre à cette personne et qui, seul ou croisé avec d'autres (un faisceau de données), permet son identification.

L'information personnelle est une information dont le trait est de porter une empreinte individuelle et de la conserver. C'est dans le rattachement par un lien d'attribution à une personne individualisée que réside la notion de donnée à caractère personnel. Chacune porte l'empreinte d'un individu singulier et demeure une émanation de cette personne qu'elle singularise et caractérise. Toutes, elles forment une masse, un corpus dont la personne apparaît à la fois comme le sujet et l'objet, de sorte que sans englober toute la personne, le corpus des informations personnelles contribue à la plénitude de son *animus*.

Les données personnelles sont des attributs de la personne humaine[4]. Révélatrice est à cet égard la référence à l'identité humaine dans la formulation

[2] VULLIET-TAVERNIER S. (2014), « *Big Data* et protection des données personnelles : quels enjeux ? », *Statistique et société*, vol. 2, n°4, p. 27.

[3] Avis sur les techniques d'anonymisation émis par le groupe de travail du G29 sur la protection des personnes à l'égard du traitement des données à caractère personnel, https://www.cnil.fr/sites/default/files/atoms/files/wp216_fr.pdf

[4] « Cette vie juridique, vouée à traduire, conformément à la justice, les besoins de l'homme, en y adaptant, dans leur continuel mouvement, les structures de fait, ne peut, sans se tarir, méconnaître les constantes nécessaires, soit de l'homme, soit du droit » – SAVATIER René (1950), *Réalisme et irréalisme en droit civil d'aujourd'hui. Structures matérielles et structures juridiques*, Mélanges Ripert, Tome 1, p. 75.

d'objectif de l'article 1er inchangé et toujours actuel de la loi du 6 janvier 1978 modifiée[5].

Il faut attendre les années 1970 et l'émergence des nouvelles technologies de l'information pour voir la notion d'information à caractère personnel se détacher avec netteté et qu'une définition juridique en soit donnée ; on parlait alors d'informations nominatives. À l'heure où les développements technologiques et le large phénomène de décloisonnement qui les accompagne conduisent à l'écrasement des spécificités et à la circulation sans entrave d'informations sans ancrage avec leur contenu substantiel, il faut promouvoir, avec force, l'approche personnaliste de la protection des données personnelles, laquelle est justifiée par l'unité du genre[6]. C'est également la position du Conseil d'État qui considère que « s'il convient de renforcer la dimension de l'individu acteur dans le droit de la protection des données, c'est en envisageant celui-ci comme un droit à l'autodétermination plutôt que comme un droit de propriété »[7].

Il faut donc, tout en favorisant l'usage des données personnelles dans l'intérêt de la santé publique, maintenir un régime de protection qui les considère fondamentalement dans leur lien avec la personne.

Développement d'une dimension participative de la recherche

Les droits des personnes concernées sur leurs données sont réaffirmés et renforcés par le RGPD. Ils participent des principes de transparence et de loyauté qui visent à leur conférer une meilleure maîtrise de leurs données personnelles (RGPD, art. 12 et s.).

Les chercheurs sont conscients de travailler avec des « données confiées » ; les volontaires entendent pouvoir décider de participer ou non aux différents projets en fonction de l'intérêt qu'ils en comprennent. Cela dépasse le seul respect du droit au consentement éclairé et relève bien plus d'un véritable « partenariat » de recherche, en particulier pour la mise en place de bases de données longitudinales, dont les données et/ou échantillons biologiques associés ont vocation à être réutilisés à des fins d'études, de recherches ou d'évaluations dans le domaine de la santé[8].

Nécessité d'un accord et d'une information honnête et accessible réaffirmée par les textes

Il est rappelé, à titre liminaire, que le consentement des personnes n'est pas systématiquement requis pour les traitements de données personnelles réalisés à des fins de recherche scientifique. Le RGPD prévoit que chaque traitement de données personnelles repose sur une ou plusieurs conditions qui en fondent la licéité, des conditions dont la liste est précisée à l'article 6. Le consentement est l'une de ces bases légales au même titre que l'exécution d'une mission d'intérêt public dont est investi le responsable du traitement, tout comme le sont le respect d'une obligation légale ou l'intérêt légitime du responsable du traitement. S'agissant du traitement des données de santé, l'objectif de recherche scientifique est un critère de licéité indépendant du consentement (art. 9.2.j).

Le point de savoir si le consentement de la personne est requis dépend donc de la qualification réglementaire de la recherche. Mais le consentement à la participation à la recherche lorsqu'il est requis ne constitue pas pour autant nécessairement la base légale du traitement.

Rappelons qu'un consentement à la participation à la recherche (*opt-in*) est requis pour les recherches impliquant la personne humaine (RIPH)[9] relevant des catégories 1°[10] ou 2°[11], des recherches qui présentent par définition des risques ou des contraintes, et pour les recherches nécessitant l'analyse des caractéristiques génétiques. Les études non interventionnelles (RIPH relevant de la catégorie 3[12]), de même que les recherches portant exclusivement sur des données ou des échantillons biologiques déjà acquis (NRIPH) sont soumises, quant à elles, à un régime d'opposition de la personne (*op-out*).

Enfin, si à l'occasion d'une recherche, il est envisagé de constituer un « entrepôt de données »[13], c'est-à-dire une base de données ouverte à des fins de recherches

[5] « L'informatique – on parlerait aujourd'hui davantage des technologies numériques – doit être au service de chaque citoyen. […]. Elle ne doit porter atteinte ni à l'identité humaine, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques. »

[6] LESAULNIER F. (2005), *L'information nominative*, thèse, Paris II.

[7] Conseil d'État, « Le numérique et les droits fondamentaux », Études 2014, p. 264.

[8] Comité d'éthique de l'Inserm (2022), « Note d'étape sur le Health Data Hub, les entrepôts de données de santé et les questions éthiques posées par la collecte et le traitement de données de santé dites "massives" ».

[9] Cette expression introduite par la loi n°2012-300 du 5 mars 2012 désigne les recherches organisées et pratiquées sur l'être humain en vue du développement des connaissances biologiques ou médicales (art. L. 1121-1 du CSP).

[10] La catégorie 1 concerne les recherches interventionnelles qui comportent une intervention sur la personne non justifiée par sa prise en charge habituelle.

[11] La catégorie 2 concerne les recherches interventionnelles ne comportant que des risques et des contraintes minimes (liste fixée par arrêté du 12 avril 2018). Exemples de ce type de recherches : prélèvements de sang ou d'échantillons biologiques effectués pour les besoins de la recherche, réalisation de scanners, entretiens ou questionnaires dont les résultats peuvent conduire à la modification de la prise en charge médicale habituelle du patient.

[12] La catégorie 3 se rapporte aux recherches non interventionnelles, c'est-à-dire celles qui ne comportent aucun risque ni contrainte et dans lesquelles tous les actes sont pratiqués et les produits utilisés de manière habituelle (liste fixée par arrêté du 12 avril 2018). Quelques exemples : recueil d'éléments ou produits du corps humain ne présentant aucun caractère invasif, enregistrements audio et/ou vidéo, questionnaires ne pouvant conduire à la modification de la prise en charge des personnes.

[13] C'est l'expression utilisée par la CNIL qui a publié, en 2021, un référentiel relatif aux traitements de données à caractère personnel mis en œuvre à des fins de création d'entrepôts de données dans le domaine de la santé.

menées ultérieurement, il est recommandé de recueillir le consentement exprès des personnes concernées par cette réutilisation secondaire des données et/ou échantillons biologiques. En effet, si l'entrepôt n'est pas fondé sur le recueil du consentement, une autorisation de la CNIL sera nécessaire (LIL, art. 65.2°), à défaut de conformité au référentiel homologué par cette autorité. Le consentement, s'il est recueilli, ne couvrira pas les recherches ultérieures qui nécessiteront une information spécifique préalable assortie d'un droit d'opposition, mais il sera de nature à fonder la constitution de l'« entrepôt ».

Quels que soient la qualification réglementaire de la recherche et le régime du recueil de l'accord des personnes, l'information de ces dernières est essentielle. En matière de recherche en santé humaine, cette transparence est une garantie reconnue aux participants en contrepartie de la levée du secret professionnel. Elle leur permet de comprendre les objectifs de la recherche, les modalités de leur participation, la portée de l'accord qu'ils donnent et de maîtriser l'utilisation qui sera faite de leurs données. Elle conditionne et facilite l'exercice effectif de leurs droits. Elle permet de développer la compréhension et d'instaurer une relation de confiance avec les chercheurs et l'organisme dont ils dépendent.

Par principe, l'information doit être délivrée individuellement à chaque personne participant à la recherche, que les données soient recueillies directement auprès d'elle ou par l'intermédiaire de tiers (LIL, art. 58). L'information individuelle doit être doublée d'une information générale dans les structures de soins qui transmettent des données personnelles de leurs patients pour permettre des activités de recherche (affichage dans leurs locaux, mention dans le livret d'accueil, etc.).

Si le RGPD favorise l'utilisation de ces données à des fins de recherches scientifiques en posant le principe d'une présomption de compatibilité des traitements ultérieurs (RGPD, art. 5 b), il impose cependant que les personnes concernées en soient individuellement informées[14]. L'information doit être réalisée pour chaque projet auquel le patient participe ou pour lequel les données du patient feront l'objet d'un traitement.

Le rôle des patients ou des personnes à l'origine des données est majeur et leurs droits à être informés et à donner leur accord à la réutilisation de ces données et échantillons doivent être respectés. Pourtant, l'exigence d'une information individuelle et spécifique à chaque projet issu de ces données et/ou échantillons est difficilement compatible avec le fonctionnement des biobanques, des cohortes, notamment celles financées par le programme d'investissements d'avenir, qui ont vocation à créer une ressource pour la communauté des chercheurs.

À la faveur du numérique, des modalités d'exercice des droits plus souples et plus dynamiques

Comme le souligne Georges Dagher au sujet des biobanques, « […] Il est temps de repenser le rôle des personnes-sources plus largement en termes de participation et de contribution à la recherche. Le nouveau paradigme développé par l'utilisation des collections biologiques et visant à créer une ressource pour la recherche invite à une évolution du cadre réglementaire et éthique qui régit la question de la participation des patients aux projets de recherche »[15].

De plus en plus, les patients ou leurs représentants au travers des associations intègrent les organes de gouvernance des infrastructures qui permettent l'utilisation secondaire des données (cohortes, biobanques, entrepôts de données…). Ils contribuent ainsi à la définition des orientations stratégiques, des principes directeurs et des procédures d'accès aux données qui seront mises en place. Ils participent également à la relecture des documents remis aux participants et sont informés régulièrement des résultats globaux des recherches. Les responsables de cohortes soulignent l'importance de cette dimension participative dans le fonctionnement en routine des cohortes, une dimension qui résulte du contrat de confiance qui lie le scientifique aux volontaires.

En outre, il n'est pas toujours réaliste ni éthique de recontacter trop fréquemment les personnes concernées, en particulier quand elles sont atteintes de maladies graves ou incurables, pour les informer de chaque utilisation de leurs données et leur permettre, à cette occasion, de s'y opposer. Une telle information ne permet pas à une personne constamment sollicitée de disposer d'une vue d'ensemble de l'utilisation des données qui la concernent.

À la faveur du numérique, d'autres modes d'information et de recueil du consentement voient le jour et sont développés, tels que la mise en place dans la durée de « portails de transparence », qui sont assortis d'une communication régulière sur l'utilisation des données et d'un droit de retrait dynamique, permettant d'adapter, le cas échéant, le consentement initial de façon flexible.

Il faut saluer le pragmatisme de la CNIL sur ce point, illustré par la rédaction de la méthodologie de référence MR004[16]. Celle-ci admet que des données et/ou des échantillons biologiques puissent faire l'objet d'une réutilisation et que l'information puisse être considérée comme valablement délivrée sans qu'il soit besoin de procéder à une nouvelle information individuelle des personnes : cela nécessite que l'information délivrée lors de la collecte initiale des données mentionne clairement la possibilité de les réutiliser et renvoie à un support d'information dynamique tel qu'un site Internet,

[14] CEPD (Comité européen de la protection des données), lignes directrices 5/2020 sur le consentement au sens du règlement (UE) 2016/679, p. 35. « La transparence [ultérieure] constitue une garantie complémentaire lorsque les circonstances de la recherche ne permettent pas un consentement spécifique. »

[15] Article paru dans le supplément « Sciences & Santé » du journal *Le Monde* du mercredi 8 juillet 2015.

[16] Délibération n°2018-155 du 3 mai 2018 portant homologation de la méthodologie de référence relative aux traitements de données à caractère personnel mis en œuvre dans le cadre des recherches n'impliquant pas la personne humaine, des études et évaluations dans le domaine de la santé (MR-004).

centralisant les informations relatives à l'ensemble des projets menés et leurs caractéristiques et auquel les personnes pourront se reporter pour, notamment, exercer leur droit de retrait s'ils le souhaitent. Ce type d'approche, conforme au rapport du 15 septembre 2017 du Comité international de bioéthique (CIB)[17], apparaît de nature à favoriser l'utilisation des données en matière de recherche, tout en préservant l'autonomie des patients qui deviennent des parties prenantes au projet.

Cette souplesse suppose toutefois que le devenir des données personnelles et des échantillons biologiques collectés dans le cadre d'un projet initial ait été anticipé et que les personnes aient été renvoyées vers un « portail de transparence », dont elles ont été informées de la localisation de façon claire et précise.

À défaut, il faudra recontacter les personnes pour les informer de l'utilisation secondaire des données. Une dérogation à l'obligation d'information sera possible, mais une absence d'information des personnes rendra les recherches subséquentes non éligibles aux mesures facilitatrices concernant les traitements jugés les plus courants, que sont les méthodologies de référence. Une demande d'autorisation, spécialement motivée, devra être effectuée auprès de la CNIL, ce qui sera de nature à retarder la mise en œuvre à l'échelle nationale ou internationale de projets de collaboration qui présentent des enjeux de santé publique.

Dans un contexte évolutif et longitudinal où prévaut l'hétérogénéité des acteurs, il est également nécessaire d'envisager d'autres modes de consentement pour assurer durablement l'équilibre entre le respect des droits des personnes et la dynamique des usages des données personnelles.

Une réflexion en lien avec les chercheurs et les partenaires doit être menée sur l'objet du consentement à la réutilisation des données/échantillons et les modalités de son recueil et de son maintien dans le temps[18]. Le RGPD permet le recueil d'un consentement « pour une ou plusieurs finalités spécifiques » (art. 6.1.a), ce qui suppose que les finalités aient été déterminées et que la personne concernée en ait été informées. Il résulte de la lecture du considérant 33 du Règlement qu'une finalité spécifique est compatible avec un consentement global. Ce considérant prévoit en effet que les personnes concernées devraient pouvoir donner leur consentement « pour ce qui concerne certains domaines de la recherche scientifique dans le respect des normes éthiques reconnues en matière de recherche scientifique. »

Les attributs de la personne humaine étant indisponibles, une personne ne devrait pas pouvoir consentir à l'utilisation de ses données à des fins de recherche scientifique sans autre précision[19]. En revanche, la personne concernée pourrait être en mesure d'accepter que ses données soient utilisées dans le cadre de différents projets de recherche susceptibles d'être menés dans une branche ou un domaine particuliers pour lesquels elle autorise l'utilisation de ses données (par exemple, la recherche en oncologie) ou encore dans le cadre d'une infrastructure qu'elle aura choisie selon des modalités de partage définies[20].

On pourrait également imaginer un consentement à options, la personne pouvant, par exemple, accepter un partage de ses données avec des organismes académiques, mais pas avec des industriels ; ou un partage de ses données quand leur utilisation reste au sein de l'Union européenne, mais pas quand elles circulent en dehors de cet espace.

Les algorithmes sont susceptibles de constituer de précieux outils d'aide à la décision[21]. Ils constituent également un sujet de réflexion pour la recherche. Outre les questions des données personnelles et de la qualité et de la fiabilité des sources des données issues d'un traitement algorithmique, se pose la question de la clarté et de l'intelligibilité nécessaires à l'explicabilité qui pèse sur les concepteurs, afin que les professionnels de santé utilisateurs puissent effectivement exercer un contrôle sur les résultats du traitement dans le respect de la garantie qu'apporte une intervention humaine (CSP, art. L. 4001-3).

Le développement d'une gouvernance des données

L'implication des personnes à l'origine des données est essentielle. Mais la protection des données dépend également de la capacité à développer des écosystèmes facilitant l'accès et l'exploitation de données pseudonymisées avec toute la rigueur et la qualité scientifiques qui s'imposent, ainsi que dans le respect de la vie privée et de la confidentialité des données.

La protection des données est un sujet de gouvernance transversale qui demande aux organismes de recherche, en lien avec leurs partenaires, d'assurer une meilleure coordination de l'action des communautés de recherche et de leur proposer un accompagnement. Comme le souligne Romain Boidin, « les contraintes réglementaires sont l'occasion de se prémunir contre les risques liés aux fuites ou aux pertes de données,

[17] UNESCO, rapport du CIB sur les mégadonnées et la santé, Bibliothèque numérique, https://unesdoc.unesco.org/ark:/48223/pf0000248724_fre

[18] CCNE, avis n°130, 28 mai 2019.

[19] Voir, en ce sens : le G29 indique que mentionner « des fins de recherche » n'est pas suffisamment clair, car le type de recherches visées n'est pas précisé. Lignes directrices sur la transparence au sens du règlement (UE) 2016/679.

[20] Voir, en ce sens : le CCNE (2018), *Numérique & santé : quels enjeux éthiques pour quelles régulations ?,* en ligne sur Internet : « Il serait donc de bonne pratique, dans toute la mesure du possible, de recueillir un consentement qui permette aux personnes d'autoriser le partage de leurs données en sachant comment elles vont être partagées (plan de partage), plutôt que pourquoi (par qui et pour quelle recherche). »

[21] Pilotée par Stanley Durrleman, Thomas Nedelec, chercheur à l'Institut du cerveau, et Carole Dufouil, directrice de recherche à l'Inserm, une étude permettant d'identifier les facteurs de risque de démence due à la maladie d'Alzheimer grâce à l'analyse de données massives vient d'être publiée dans la revue *The Lancet Digital Health*, le but étant de détecter le plus tôt possible les signes précoces et facteurs de risque pour retarder le début de la maladie.

et d'enclencher, dès aujourd'hui, la transition numérique du soin et de la recherche »[22].

Le développement de solutions d'hébergement et de mise à disposition de données sécurisées et mutualisées permettant de garder leur maîtrise et présentant des garanties de conformité technique et réglementaire certifiées est indispensable. Il nécessite le développement d'architectures techniques appropriées et régulièrement réévaluées. De nombreux organismes sont actuellement confrontés sur ce point à la question du choix de solutions techniquement robustes non soumises exclusivement aux juridictions européennes en l'absence d'alternatives françaises ou européennes.

De même, doivent être mises en place en concertation avec l'ensemble des partenaires, des procédures et des chartes d'accès aux données permettant de faciliter l'évaluation scientifique et éthique des projets de recherche, selon des procédures de partage respectueuses du cadre légal et réglementaire.

Renforcer l'acculturation et l'appui aux chercheurs

Il n'est pas simple pour les chercheurs et les institutions qui les accompagnent de comprendre et de respecter les exigences de la réglementation.

Un environnement juridique complexe

La lisibilité du droit pour les acteurs de la recherche n'est pas facile. Le RGPD doit être articulé avec la loi « Informatique et libertés » française[23] qui maintient un régime d'autorisation pour les recherches, études ou évaluations réalisées dans le domaine de la santé, dès lors qu'elles ne sont pas conformes à une méthodologie de référence (art. 66 III). Cette même loi doit aussi être articulée avec d'autres dispositions de l'UE et nationales applicables à la recherche[24], notamment les dispositions du Code de la santé publique applicables aux recherches impliquant la personne humaine et au système national des données de santé (SNDS), et les dispositions du Code civil. De nouveaux règlements européens sont en cours d'élaboration qui devront, eux aussi, s'articuler avec le cadre juridique existant[25].

La préparation et l'instruction réglementaire des dossiers de demandes d'autorisation pour des études sont devenues plus complexes ; elles requièrent une expertise juridique et technique toujours plus importante. Les procédures administratives demeurent complexes, faisant intervenir une multiplicité d'organismes selon que la recherche implique ou non la personne humaine.

Dans ce contexte, les chercheurs doivent pouvoir bénéficier d'un accompagnement expert qui leur assure, en amont de la soumission, une conformité de leurs dossiers à la réglementation et aux attentes de la CNIL et qui, finalement, facilite l'obtention des autorisations dans des délais compatibles avec ceux parfois très contraints des recherches (projets sur l'étude de la pandémie de Covid-19…). En parallèle de cette évolution, le respect des contraintes réglementaires évoquées précédemment devient un élément déterminant pour la participation à des projets internationaux et européens.

Mise en place d'un dispositif de formation des chercheurs et d'autorégulation adapté aux spécificités du RGPD

Les organismes de recherche doivent se doter d'équipes pluridisciplinaires dédiées, à même de relever les défis juridiques et techniques posés par la réglementation. La fonction de délégué à la protection des données (DPD, ou DPO pour *Data Protection officer*) a été rendue obligatoire dans tous les organismes dont l'activité est d'effectuer des traitements à grande échelle de données sensibles. Son rôle est de contribuer, en lien avec l'ensemble des acteurs présents dans l'organisme et les partenaires de celui-ci, à l'élaboration et à la déclinaison de politiques institutionnelles favorisant la protection des données. Il est également de faciliter la tâche des scientifiques pour leur permettre de se consacrer à la production de connaissances utiles à la santé de tous. L'objectif est de diffuser une culture et de bons réflexes, l'enjeu étant la déclinaison opérationnelle de la protection des données personnelles au sein de l'organisme.

Pour ce faire, des campagnes d'information et de formation au RGPD et à ses contraintes doivent être réalisées régulièrement afin de sensibiliser les personnels, chercheurs comme administratifs.

Par ailleurs, la délégation à la protection des données de chaque organisme doit proposer des supports pédagogiques permettant aux scientifiques de s'approprier le cadre juridique et de les guider de façon didactique dans leur démarche de mise en conformité, tels que des guides de bonnes pratiques assortis de modèles permettant une auto-régulation par la communauté scientifique.

Enfin, les dossiers complexes, stratégiques et innovants, qui laissent plus difficilement place à une standardisation des procédures, doivent être évalués en recourant à des compétences scientifiques et juridiques croisées. Il faut capitaliser sur l'expérience acquise pendant la pandémie de Covid-19 et associer les délégations à la protection des données au montage des projets de recherche, et ce dès les premières

---

[22] BOIDIN R. (2019), *Protection et gouvernance des données dans la recherche en santé*, thèse soutenue pour l'obtention de son doctorat d'État en pharmacie.

[23] Loi n°78-17 du 6 janvier 1978 modifiée relative à l'informatique, aux fichiers et aux libertés.

[24] Notamment le règlement (UE) 536/2014 du 16 avril 2014 relatif aux essais cliniques de médicaments à usage humain et abrogeant la directive 2001/20/CE.

[25] Le 21 avril 2021, la Commission européenne a publié sa proposition de règlement établissant des règles harmonisées en matière d'intelligence artificielle. Le 3 mai 2022, la Commission européenne a présenté une proposition de règlement dans le but d'instituer l'espace européen des données de santé (EHDS), https://ec.europa.eu/health/publications/proposal-regulation-european-health-data-space_fr

réunions de cadrage[26]. Ce travail d'anticipation réalisé en équipe, dans l'esprit du RGPD et en conformité avec les lignes directrices du Comité européen de protection des données[27], est essentiel pour éviter les « erreurs d'aiguillage », qui obligent à des allers-retours et à des modifications ultérieures coûteuses en temps et en énergie. Cette écoute en amont des scientifiques permet de bien comprendre leurs besoins et de trouver avec eux les adaptations nécessaires des protocoles permettant de concilier ces besoins avec les exigences de la réglementation. Bien plus, cela permet aux chercheurs de faire des choix méthodologiques mieux éclairés, qui contribueront à l'élaboration plus rapide de dossiers solides permettant ultérieurement une valorisation agile des données personnelles dans le respect de la transparence à l'égard des personnes concernées. Cela facilite également le dialogue qu'entretiennent les DPO avec la CNIL, les partenaires, les participants aux études et les autres acteurs.

[26] LESAULNIER F. (2021), « La protection des données personnelles au cœur de la recherche Covid-19 à l'Inserm », *JDSAM,* n°29, p. 33.

[27] Lignes directrices 03/2020 sur le traitement de données concernant la santé à des fins de recherche scientifique dans le contexte de la pandémie de Covid-19. Adoptées le 21 avril 2020.

Mise en place d'un réseau de proximité constitué de référents spécialisés

Pour accompagner les scientifiques au niveau local, il est utile de mettre en place dans les organismes – au sein des centres qui traitent à grande échelle des données sensibles, des plateformes de services, des services support, des directions métier, etc. – des référents spécialisés dans la protection des données personnelles qui constituent autant de relais de proximité et qu'il appartient à la DPO de former et d'animer. Ces référents, dont les compétences diverses contribuent à la richesse du réseau, seront autant de points d'appui au niveau local pour diffuser les bonnes pratiques et faire remonter les questions qui permettront de construire des outils basés sur des retours d'expérience et qui seront adaptés aux réalités du terrain.

Face aux contraintes imposées par la réglementation et aux enjeux de la mise en place de politiques institutionnelles de protection des données, un besoin de renforcement des délégations à la protection des données, qui interviennent aux côtés des chercheurs, et de leur gouvernance se fait sentir, tout comme le besoin de la création d'instances locales d'appui à la recherche. Ce nécessaire renforcement s'impose aussi au niveau des services de la CNIL. Tous ensemble, nous pouvons fluidifier l'activité de recherche au bénéfice de la santé des populations, tout en assurant la promotion d'une recherche en santé humaine responsable.

La recherche internationale et la protection des données

Par Gaëlle BUJAN
Déléguée à la protection des données, CNRS

La recherche évolue à l'échelle mondiale dans ses fondements, ses valeurs et ses réseaux d'experts. À l'ère du numérique, l'accélération des échanges et les perspectives de recherche accompagnent une diffusion des savoirs, dont les limites sont sans cesse repoussées. Simultanément, les risques sont nouveaux et démultipliés ; ils sont associés à des impératifs d'intégrité, d'éthique et de respect des personnes. La réglementation s'adapte à ce mouvement : elle vise à responsabiliser les acteurs, à contribuer à la confiance de la société dans la recherche et devient de plus en plus protectrice des droits et des personnes.
Dans cet environnement évolutif et d'apparence contraignant, les organismes de recherche ont mis en place des politiques favorisant la culture de la protection des données et la réalisation de leurs objectifs à l'échelle internationale. Recherche internationale et réglementation relative à la protection des données ne sont ainsi pas incompatibles.

Tout établissement impliqué dans des activités de recherche a besoin quotidiennement de ressources humaines, financières et matérielles. Ces dernières recouvrent notamment les données de la recherche issues d'expériences, de calculs complexes… ; elles sont constituées de textes, de logiciels… Selon les disciplines scientifiques, les natures et les supports des données sont variables. Leur sensibilité peut s'entendre dans une logique de concurrence industrielle, être liée aux intérêts fondamentaux de l'État, à la vie privée et à des dispositions réglementaires.

Parmi ces données de recherche, une partie constitue des données à caractère personnel (c'est-à-dire toute donnée permettant d'identifier directement ou indirectement une personne[1]).

Dans cet article, nous centrons notre propos sur ces seules catégories de données qui requièrent, pour pouvoir être utilisées et/ou réutilisées à des fins de recherche scientifique, des mesures techniques et organisationnelles adaptées à la protection de la vie privée des personnes.

Le Règlement européen sur la protection des données personnelles (RGPD) consacre de manière large le traitement de données personnelles à des fins de recherche scientifique[2]. L'article 89 du RGPD prévoit les « garanties et dérogations applicables aux traitements à des fins de recherche scientifique ou historique. »

Il s'agit donc de faciliter la recherche, pour laquelle la donnée personnelle constitue le matériau essentiel. D'ailleurs, la Commission nationale de l'informatique et des libertés (CNIL) accompagne cette évolution en mettant à la disposition de la communauté scientifique un ensemble de guides et de fiches pratiques.

Au niveau du Centre national de la recherche scientifique (CNRS), dont la mission est « d'identifier, d'effectuer ou de faire effectuer, seul ou avec ses partenaires, toutes les recherches présentant un intérêt pour la science ainsi que pour le progrès technologique, social et culturel du pays »[3], la protection des données personnelles est associée à l'éthique, à la déontologie et à l'intégrité scientifique, qui sont des valeurs fortes de l'établissement. Il porte une réelle responsabilité sociale lorsque des données lui sont confiées par des citoyens pour conduire sa mission originelle de recherche.

Dans son expertise et son accompagnement, le Centre prend en compte l'environnement évolutif de la recherche et la sensibilité des données. La politique de protection des données mise en œuvre par le CNRS concerne tous les traitements, bien que cela peut parfois s'avérer complexe au regard du caractère international de la recherche.

L'environnement évolutif de la recherche

Les progrès de la connaissance et les changements technologiques, économiques, sociaux, environnementaux et culturels contribuent à l'évolution constante des conditions et du contexte d'exercice de l'activité de recherche.

Ces différents facteurs forment des opportunités ou, au contraire, apparaissent comme des contraintes.

(1) Article 4 du RGPD (Règlement UE 2016/679 du Parlement européen et du Conseil, du 27 avril 2016).

(2) Considérants 156, 157 et 159 du RGPD.

(3) Décret n°82-993 modifié du 24 novembre 1982.

Le numérique

Les technologies de l'information et de la communication ont connu une accélération conséquente au cours des vingt dernières années : le développement des réseaux sociaux et la frontière devenue extrêmement ténue entre les sphères privée et publique modifient profondément les comportements individuels ou collectifs. L'accélération de la diffusion et de la circulation de l'information ouvre de nouvelles opportunités scientifiques.

Les progrès de la connaissance prennent appui sur de nouveaux objets de recherche, de nouvelles modalités d'accès aux données et des modalités d'expérimentation différentes. La situation sanitaire récente a transformé les possibilités de réalisation des expériences, lesquelles s'opèrent à domicile et non plus uniquement au sein des laboratoires. Les protocoles sont plus aisés à élaborer du point de vue sanitaire, mais sont plus intrusifs en termes de collecte des données personnelles.

Le volume des informations disponibles et accessibles, les capacités des grands calculateurs et des infrastructures de recherche, ainsi que les potentialités des algorithmes ouvrent des champs nouveaux à l'échelle mondiale dans les domaines de l'intelligence artificielle, du quantique…

Les données et les bases de données construites et exploitées pour des finalités de recherche sont de plus en plus des informations économiquement stratégiques et, de fait, qui s'inscrivent dans une compétition scientifique internationale accrue. Les risques de captation, de piratage, de potentielles réutilisations à des fins mercantiles sont réels et imposent la mise en œuvre de politiques de protection des données, voire des personnes. Les établissements publics déploient à cette fin la politique de protection des systèmes d'information (PSSI) élaborée par l'État. Comme dans les autres organismes de recherche, cette PSSI s'applique au CNRS, mais en tenant compte des spécificités de ses missions, notamment le nomadisme, la compétition internationale et la protection du patrimoine scientifique et technique. Pour l'ensemble de ses activités, la stratégie numérique de l'État est donc prise en compte.

L'ouverture de la science

La communauté scientifique est également confrontée au changement de paradigme induit par l'ouverture de la science. Cette nouvelle orientation vise à renforcer l'efficacité de la recherche, favoriser la transparence et la reproductibilité de la recherche et à l'intégrer dans la compétition internationale.

Impulsées par Frédérique Vidal, ministre de l'Enseignement supérieur, de la Recherche et de l'Innovation[4], des initiatives conséquentes ont été prises pour accélérer l'évolution de la mise en place des plans d'action dédiés décidés au niveau du ministère précité, au CNRS (plans d'ouverture des données, science ouverte)…

Le double objectif d'ouverture et d'accessibilité est à intégrer dès le démarrage d'un projet et à introduire dans la préparation des plans de gestion de données et le cycle de vie des données notamment en vue de leur réutilisation.

Les évolutions réglementaires

Pour accompagner les évolutions dans l'organisation du travail et la plus grande circulation des biens et des services, la protection des biens et des personnes est renforcée. L'éthique, la déontologie et l'intégrité scientifique sont des valeurs fortes portées par les établissements en charge de la recherche et elles sont aujourd'hui inscrites dans les textes réglementaires.

La sensibilité des données

Les données et jeux de données présentent un intérêt économique réel si leur utilisation est détournée de leur finalité de recherche. Cette sensibilité est ici liée à leur valeur économique potentielle.

Des catégories particulières de données

Selon l'article 9 du RGPD, des traitements portant sur des catégories particulières de données personnelles sont possibles, mais sous certaines conditions.

Il s'agit des « données qui révèlent l'origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement de données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique. »

Les données peuvent concerner le domaine de la santé : données relatives à la santé physique ou morale passée, présente ou future, qui donnent une indication sur l'état de santé de la personne. Leur traitement et les procédures dépendent de la finalité de la recherche ainsi que précisé dans le guide pour la recherche[5] (Institut des sciences humaines et sociales du CNRS).

D'autres données font l'objet de dispositions spécifiques :

- le numéro de sécurité sociale est une donnée identifiante, dont l'utilisation est très encadrée ; elle est circonscrite à quelques finalités définies par la CNIL ;
- les données d'infraction ou liées à des condamnations relèvent pour leur exploitation de dispositions spécifiques prévues par la loi Informatique et libertés modifiée.

Des données hautement personnelles

Elles peuvent présenter une forte sensibilité pour la personne ; il peut s'agir de données liées aux revenus des ménages, au patrimoine individuel, aux activités professionnelles… Des dispositifs spécifiques et protecteurs de la vie privée permettent à la communauté scientifique d'utiliser ces données à des fins de

[4] https://www.enseignementsup-recherche.gouv.fr/fr/le-plan-national-pour-la-science-ouverte-les-resultats-de-la-recherche-scientifique-ouverts-tous-49241

[5] https://www.inshs.cnrs.fr/sites/institut_inshs/files/pdf/Guide_rgpd_2021.pdf

recherche. Par exemple, le Centre d'accès sécurisé aux données (CASD) permet d'accéder à distance à une infrastructure sécurisée assurant ainsi la confidentialité des données.

Des données concernant des personnes vulnérables

La notion de personne vulnérable n'est pas définie précisément par la réglementation. Elle peut concerner les mineurs, les personnes âgées ou des catégories particulières de population.

Pour leurs activités de recherche et en fonction de leur métier, les chercheurs, ingénieurs ou techniciens apportent une forte attention à la protection de leurs données : ce sont, par exemple, des personnes travaillant dans des animaleries, ayant accès à des équipements dangereux ou à risque, des spécialistes dans des domaines confidentiels, stratégiques ou liés à la défense des intérêts de la nation… Dans ces situations, une identification de la personne et des coordonnées de celle-ci deviennent des données sensibles à protéger (en ne mentionnant pas d'informations identifiantes dans les annuaires publics, notamment).

La politique de protection des données au CNRS

Le Centre national de la recherche scientifique[6] est une institution de recherche parmi les plus importantes au monde ; elle œuvre dans tous les champs de la connaissance (voir la Figure 1 ci-dessous).

[6] https://images.cnrs.fr/video/7343

L'implication du CNRS dans l'apport d'un haut niveau de sécurité en matière d'utilisation des données est renforcée depuis plus de dix ans. Au regard des risques encourus en termes de disponibilité, de fiabilité et d'intégrité des données et, par conséquent, de l'impossibilité éventuelle pour lui de conduire des projets de recherche, le CNRS a internalisé des fonctions d'analyse de la conformité des traitements de ses données à la loi Informatique et libertés.

Outre les aspects liés aux potentielles sanctions pénales et financières associées à l'absence d'application de la réglementation, les enjeux relèvent pour le CNRS de l'éthique et de l'intégrité scientifique. Pour le CNRS et la communauté scientifique, la protection de la donnée traduit leur souhait de conserver la confiance de la société dans l'activité scientifique et de crédibiliser leur action mise en œuvre essentiellement avec des ressources publiques.

Le CNRS a décidé de consacrer des ressources conséquentes à la protection des données et développe une culture de la protection de la donnée auprès de l'ensemble des communautés de l'établissement.

Son expertise est forte dans l'analyse des traitements à finalité de recherche scientifique, s'appuyant sur un réseau de compétences réparties dans les différentes directions du CNRS et ses unités de recherche. Des compétences dans les domaines juridique et de la sécurité des systèmes d'information sont aussi mobilisées.

Le choix fait d'une forte mobilisation pour sensibiliser et former à la protection des données personnelles a favorisé la prise en compte de cette dimension dans l'élaboration des projets et des protocoles de recherche.

Figure 1 : Les chiffres clés du CNRS – Source : Rapport d'activité 2021 du CNRS.

L'approche par les risques est retenue pour chaque traitement de données conduisant, selon les situations, à des analyses d'impact sur la vie privée. Au CNRS, ces moments de réflexion sont menés de manière collaborative entre les chercheurs, les juristes et les spécialistes de la sécurité des systèmes d'information et se fondent sur la méthodologie proposée par l'autorité de contrôle qu'est la CNIL.

La recherche internationale et la protection des données

Une caractéristique intrinsèque de la recherche est qu'elle est, par essence, internationale. La science se nourrit des avancées faites par une communauté qui travaille en réseau, par-delà les frontières.

Chaque chercheur est confronté à des réglementations qui diffèrent selon les pays : protection des données, du patrimoine scientifique, accords internationaux, principes éthiques de la recherche, protocoles spécifiques aux disciplines, mesures de sécurité dont la sécurité des systèmes d'information.

Le RGPD est très protecteur de la vie privée et, plus largement, des personnes. Il comporte de nombreuses orientations qui ont nécessité une évolution des réglementations nationales. Ainsi, dans de nombreux pays, la protection des données est une véritable préoccupation qui se traduit par un nombre conséquent de textes, dont la communauté doit tenir compte[7]. En cas de conflit entre les réglementations, il est souvent préconisé de retenir les dispositions les plus protectrices des droits des personnes.

Les projets financés par l'Union européenne

La dimension protection des données est affirmée et exigée par l'UE pour tous les projets qu'elle accompagne. Ainsi, chaque acteur a l'obligation sur chacun de ses sujets de recherche d'apporter des informations détaillées sur l'utilisation et le devenir des données personnelles[8].

L'Union européenne a élaboré un guide sur l'éthique et la protection des données pour aider les chercheurs dans la définition de leurs données sensibles et à porter à celles-ci une grande attention selon leurs thématiques de recherche.

Les démarches de conformité guidées par le terrain de recherche

Quel que soit le sujet de recherche, l'objectif de protection des personnes s'appuie sur des principes généraux que sont les finalités des traitements, la proportionnalité des données, les droits des personnes, les durées de conservation et les mesures de sécurité adaptées.

Pour chaque projet et chaque collaboration de recherche, des questions similaires sont soulevées : quels acteurs, quels rôles, quelles responsabilités, quelles données, comment partager, quel hébergement, quelle organisation adopter ?

La réglementation et les dispositions à retenir sont déterminées par l'analyse du projet et par le profil des acteurs concernés (leur pays d'origine, leur structure d'appartenance…). Les critères de localisation des responsables des traitements des données et de ciblage de ces traitements permettent de déterminer si le RGPD s'applique ou non.

Conclusion

La protection des données personnelles pour les finalités de la recherche s'insère dans les bonnes pratiques et les valeurs fondamentales de la recherche, quel que soit le périmètre national, européen ou international sur lequel elle porte. Il est un fait que le RGPD diffuse largement et que son application est de plus en plus internationale. Cette dimension est devenue cruciale pour le secteur de la recherche.

[7] https://www.cnil.fr/fr/la-protection-des-donnees-dans-le-monde

[8] Guide européen, voir : https://cache.media.education.gouv.fr/file/2018/54/9/h2020_hi_ethics-data-protection_en_1046549.pdf

Aligning access to microbiome data and privacy considerations for better solutions for health and wellbeing of society and environments

By Frederik COPPENS
VIB-UGent, ELIXIR Belgium, Gent Belgium

Lene LANGE
BioEconomy Research & Advisory, Copenhagen, Denmark

And Kathleen D'HONDT
Department Economy, Science and Innovation, Flemish Government, Brussels, Belgium

There is a growing body of evidence that underpins the importance of microbiomes in biology. Understanding the functioning of microbiomes and their interaction with the environments will allow to develop novel interventions to support human, animal, and plant health as well as the environment. The potential that microbiomes can have to prevent the onset of non-communicable diseases is huge. This can only be developed when studying the impact of lifestyle, nutrition and environment in the context of the genetic content. As human microbiomes have been shown to be stable over time and can allow to identify the "carrier" of the microbiome, access to microbiome data has been questioned in the light of privacy protection and the General Data Protection Regulation. In this paper we discuss the potential of microbiomes in different areas and how microbiome data may be shared to support the concept of doing good.

Introduction

Microbiome research has taken a giant leap forward over the last years. It is now generally accepted that microbiomes are an integral part of our body and all biological ecosystems. The role of microbiomes on our health, wellbeing, development and to the health and resilience of whole ecosystems also becomes better documented.

The growing insights on the impact of microbiomes has created an important potential for the development of novel intervention, both in human and in animal health, as well as for ag-food production.

Understanding microbiome functioning is based on the analyses of large data sets, combined with learning from insights across different expertise domains. The importance of learning across microbiomes of man, animals, plants, soil and food was underlined by the calls from researchers to initiate global microbiome initiatives to accelerate the technologies for analyses and agree on standardisation in the different aspects of microbiome research (Alivisatos *et al.*, 2015; Dubilier *et al.* 2015).

Microbiomal fingerprints

Nevertheless, open access to microbiome data has become an issue of discussion as it was shown that metagenomic microbial fingerprints are sufficiently stable to be linked to individuals, even with follow-up samples (Franzosa *et al.*, 2015). The gut microbiome proved exceptionally stable, where over 80% of samples uniquely matched in follow up samples. Other body site-specific microbiomes reached over 30% positive matches. This makes microbiome-related findings a powerful clinical tool for care management, but raises privacy concerns (Chuong *et al.*, 2017). Indeed, associations between human host genotype and gut microbiome have been found (Hughes *et al.*, 2020), and the skin microbiome leaves a signature that allows to identify the person that last touched a surface (Elhaik *et al.*, 2021).

Privacy issues also arise when studying the health impact of lifestyle and diet through the microbiome. When combining such data to microbiome data and health information, analyses may reveal unexpected correlations on metabolic diseases, allergies or intolerance to certain foods. Understanding the interplay of

nutrition, lifestyle, microbiome, and genome is believed to open novel manners for better health management and prevention. Nutrition and lifestyle have a major impact on the occurrence of non-communicable diseases (NCDs), which continue to have an increasing impact across the globe[1].

Combining nutrition and lifestyle information with genomes and microbiomes data

In the aftermath of the human genome mapping, the field of nutrigenomics was launched to understand the relationship between nutrients, diet and gene expression. More recently it became clear that the associated microbiome may be seen as an intermediate that translates food, nutrition and lifestyle into health effects. However, the multifactorial and long-term impact of diet and lifestyle on health, based on the interplay of genome-epigenome-microbiome is too complex for straightforward conclusions.

It needs large data sets gathered during lifetime to characterise how nutrition and lifestyle exert health impacts. Advanced functional prediction and artificial intelligence are expected to lead to novel pathways and interventions to prevent or postpone NCDs. Ultimately, personalised diets for better health outcomes may be defined and guidance for healthy diets targeting different types of societal groupings at different stages in life – from new-borne, childhood, adulthood to senior – as well as related to gender and/or ethnicities may be provided. The role of the gut microbiome for the onset and severity of NCDs, including mental conditions has become increasingly clear.

Accelerating this field will transform healthcare from disease to health management thereby contributing to the sustainability of our healthcare systems. In this respect, the concept of preventive personalised health, emphasising on the role of lifestyle, diet and microbiomes to realise health impacts, has been launched.

Industry perspectives based on microbiome insights

At the same time embracing the potential of microbiome research is also opening a huge potential for the industry, such as the diagnostics or food, feed and diet industry. Microbiome-based applications are also being developed in the bioag sector, mainly in the area of plant protection, soil resilience and improvement, yield improvements, waste management, and contributing to circular economy and organic farming. Finally, this data intensive field also creates extra activities for service providers.

Market prognoses by different consultancy companies estimate the global microbiome-based interventions for health to reach roughly over USD 1-2 billion with a compound annual growth rate (CAGR) of between 15%-20% or even up to 30% for the next 10 years, dependent on the inclusion criteria[2].

Also the global bioag industry is expected to be an important growth market with an expected CAGR of around 14%, to reach over USD 10 billion by 2025-2027[3].

Open big data effort needs

To realise these expectations, different types of data sets should be accessible and combined. The information buried in combined data sets holds enormous promise to develop personalised nutrition, lifestyle approaches or health interventions to conserve health and contribute to better wellbeing.

Open access to data has accelerated research and development considerably (Burgelman *et al.*, 2019). Data sets covering different microbiome domains and integrating the different data resources will lead to more profound insights and contribute to better solutions to address challenges in public health, climate change and food security (D'Hondt *et al.*, 2021).

Not only within the health field it is important to link the different microbiomes and view these as a continuum, also holds true for environmental microbiomes in soil, on and in animals, plants and crops. In view of the one-health concept, all microbiomes are in continuum interacting. The interaction among the microbiomes from different environments will give new insights in the biology and biodiversity. In addition, research is accelerated in a shared knowledge environment covering different microbiome areas. Open access to data resources is a strong means to accelerate knowledge distribution, but as outlined higher, the sharing of human microbiome data in open access is under discussion. As even summary statistics risk to expose privacy sensitive information, this warrants consideration and the development of guidelines for data release (Cho, 2021).

Open science and privacy protection are two sides of a coin that are in conflict to each other. These considerations are manifest in particular, in the realm of personalised medicine. Comparisons of genomes, documented with associated health data is instrumental to develop personalised medicine. The risk of such data being used by insurers, banks, employers to inform decisions whether or not to grant an insurance a loan or a job should be avoided. Several countries have adopted specific restrictions for the use of genomic information for crime investigation. Access to such information is granted on a case by case approach and jurisprudence.

[1] www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

[2] https://www.marketsandmarkets.com/Market-Reports/human-microbiome-market-37621904.html; https://www.mordorintelligence.com/industry-reports/human-microbiome-market; https://www.businesswire.com/news/home/20211221005468/en/Global-Human-Microbiome-Industry-Landscape-2021-2028---New-Drug-Requirements-for-Faecal-Microbiota-Presents-Opportunities---ResearchAndMarkets.com

[3] https://www.fortunebusinessinsights.com/industry-reports/agricultural-microbial-market-100412; https://www.marketsandmarkets.com/Market-Reports/agricultural-microbial-market-15455593.html

Privacy protection

Different options may be envisioned to deal with the privacy issues linked to the sharing of microbiome data. Microbiome fingerprints may be handled like a genomic passport or profile, i.e. for healthcare treatment this data should be available.

For secondary use, the sharing of sensitive data is currently typically based on contracts, after evaluation by a data access committee. As complete anonymisation is not possible, like in genomic analyses, rather than sharing the data, the analysis may be brought to the data. Through a federated infrastructure, the exact same analysis is performed locally on each dataset, and only the resulting (aggregated) data is shared, reducing or even eliminating privacy concerns[4]. This method can also alleviate the technical challenges associated to transferring the ever increasing volumes of data. However, this also imposes limitations which may require development of new algorithms. Trusted research environments are being established to provide a secure computing environment to access sensitive data [5] and [6].

Alternatively, analyses can be performed on encrypted data (Senf *et al.*, 2021). The Global Alliance for Genomics and Health, is developing standards to enable such innovative approaches[7]. These methodologies are being developed and deployed in the context of the 1-Million Genomes Project[8].

The field may also be advanced without compromising human health data protection, by drawing lessons from studies on farmed animals. In this setting feeding, age, gender and genetic background can be accurately tracked, while the impact on the microbiome composition and function can be analysed. The data collected in this way will increase the understanding of the functioning of microbiomes and role for animal health and welfare. It may allow to extrapolate insights from lifestock to approaches for human health.

The risk of privacy intrusions should not block research and hence the benefits that sharing of data may bring. Restricting access to human microbiome data will also impact other microbiome areas, as research on human linked microbiomes is most advanced, due to the potential to develop novel health interventions.

Policy and decision makers should engage to establish the right measures for correct use of data, in support of R&I, while preventing all types of misuse. Creating the right conditions to grant access to data is essential to ensure trust among citizens to agree on the sharing of their data.

[4] https://ega-archive.org/federated

[5] www.hdruk.ac.uk/access-to-health-data/trusted-research-environments/

[6] https://research.csc.fi/sensitive-data-services-for-research

[7] www.ga4gh.org

[8] https://digital-strategy.ec.europa.eu/en/policies/1-million-genomes

Ownership of data

Another important aspect is the ownership of data and biological samples. According to the current legislation people own their health data, including their genome data and microbiome data. The governance of this data is with healthcare professionals, health institutions, insurers or governments. The individual is requested to give consent to use his or her data, but often lacks information about later use of the data. Nevertheless, people are most often interested to share their data and bio-sample, thereby supporting research as a token of good citizenship. Ownership of biobank samples is also a matter of debate.

There is no consensus yet whether microbiomes should be considered as part of the human body and hence be treated like human tissue. In particular, for faecal samples, which are generally considered waste in the first place, it becomes difficult, since it was shown that microbiomes are sufficiently unique to be linked to individuals.

Citizens should at all times have governance over their data, including the microbiome data. Informing people on what their data will be used for and on the outcomes of studies should become mandatory. At the same time, the raw data should not be the source for profit purposes. Data is referred to as the new gold, while it may be questionable whether it is acceptable to use the combined data that form the basis of our common co-evolution with microbiomes for profit and hence the basis to better understand the underlying biology, which will lead to novel interventions and treatments for better wellbeing. In contrast, novel interventions based on more profound understanding of the functioning of microbiomes may be offered for profit.

Building the right framework conditions to protect against misuse of data and ensure awareness building and informed citizens are essential.

Ethical considerations

Restrictions for use of microbiome data, due to the (interpretation of the) General Data Protection Regulation[9], data ownership and privacy interferes with getting to better understanding of microbiome compositions, roles and functions and represents an ethical dilemma as it limits the goals of doing good, based on the growing body of knowledge of the biology and co-evolution including the microbiomes in and around them.

The co-evolution and continuity of microbiomes is well illustrated in the bioag field. Enrichments of plant microbiomes of the root, leaf, or seed, may protect plants against pests and hence reduce the need for pesticides. In this way, microbiomes may contribute to higher yields and even more robustness against stress conditions.

In lifestock breeding, improvement of the gut microbiomes of farmed animals reduces the risk of inflammatory gut, which reduces the need for antibiotics.

[9] https://eur-lex.europa.eu/eli/reg/2016/679/oj

This contributes not only to better animal health, but more importantly, may significantly reduce the need for antibiotics and thus reduce the risk of antibiotic resistance. The spread of antibiotic resistant bugs is one of the most serious threats for future pandemics.

Microbiomes have potential to bring benefits in bioag, human health and environment. The protection of data should not be the limiting factor as it may be considered un-ethical not to share such knowledge.

Conclusions

Integrating the potential of microbiomes for human and animal, environmental and planetary health offers huge opportunities for health and wellbeing. The sharing of data is crucial to accelerate the insights on how microbiomes function and interact in the environments. It will lead to understanding what determines resilience and health of microbiomes and how this impacts health of living organisms and the environment.

However, access to health data should not be misused for unintended purposes. The data needed to build the knowledge should not be the source of profit unlike the development of novel interventions based on the analyses of these data.

Human microbiome data should be treated with similar privacy concerns like genomic data. Working in secured data computing environments, bringing the analyses to the data stored in federated infrastructures and sharing only the resulting aggregated data and using encrypted data may provide novel approaches to ensure privacy protection.

Bibliography

ALIVISATOS A. P., BLASER M. J., BRODIE E. L., CHUN M., DANGL J. L., DONOHUE T. J., DORRESTEIN P. C., GILBERT J. A., GREEN J. L., JANSSON J. K., KNIGHT R., MAXON M. E., MCFALL-NGAI M. J., MILLER J. F., POLLARD K. S., RUBY E. G. & TAHA S. A. (2015) – Unified Microbiome Initiative Consortium, "A unified initiative to harness Earth's microbiomes", *Sciences* 350, pp. 507-508, doi: 10.1126/science.aac8480. Epub 2015 Oct 28. PMID: 26511287.

BURGELMAN J. C., PASCU C., SZKUTA K., VONSCHOMBERG R., KARALOPOULOS A., REPANAS K. & SCHOUPPE M. (2019), "Open Science, Open Data, and Open Scholarship: European Policies to Make Science Fit for the Twenty-First Century", *Frontiers in big data* 2, p. 43, https://doi.org/10.3389/fdata.2019.00043

CHO J. C. (2021), "Human microbiome privacy risks associated with summary statistics", *PloS one* 16, e0249528, https://doi.org/10.1371/journal.pone.0249528

CHUONG K. H., HWANG D. M., TULLIS D. E., WATERS V. J., YAU Y. C., GUTTMAN D. S. & O'DOHERTY K. C. (2017), "Navigating social and ethical challenges of biobanking for human microbiome research", *BMC medical ethics* 18, p. 1, https://doi.org/10.1186/s12910-016-0160-y

D'HONDT K., KOSTIC T., MCDOWELL R., EUDES F., SINGH B. K., SARKAR S., MARKAKIS M., SCHELKLE B., MAGUIN E. & SESSITSCH A. (2021), "Microbiome innovations for a sustainable future", *Nat. Microbiol.* 6, pp. 138-142, doi: 10.1038/s41564-020-00857-w. PMID: 33510435.

DUBILIER N., MCFALL-NGAI M. & ZHAO L. (2015), "Microbiology: Create a global microbiome effort", *Nature* 526, pp. 631-634, https://doi.org/10.1038/526631a

ELHAIK E., AHSANUDDIN S., ROBINSON J. M., FOSTER E. M. & MASON C. E. (2021), "The impact of cross-kingdom molecular forensics on genetic privacy", *Microbiome* 9, p. 114, https://doi.org/10.1186/s40168-021-01076-z

FRANZOSA E. A., HUANG K., MEADOW J. F., GEVERS D., LEMON K. P., BOHANNAN B. J. & HUTTENHOWER C. (2015), "Identifying personal microbiomes using metagenomic codes", *Proceedings of the National Academy of Sciences of the United States of America* 112, pp. E2930-E2938, https://doi.org/10.1073/pnas.1423854112

HUGHES D. A., BACIGALUPE R., WANG J., RÜHLEMANN M. C., TITO R. Y., FALONY G., JOOSSENS M., VIEIRA-SILVA S., HENCKAERTS L., RYMENANS L., VERSPECHT C., RING S., FRANKE A., WADE K. H., TIMPSON N. J. & RAES J. (2020), "Genome-wide associations of human gut microbiome variation and implications for causal inference analyses", *Nature microbiology* 5, pp. 1079-1087, https://doi.org/10.1038/s41564-020-0743-8

SENF A., DAVIES R., HAZIZA F., MARSHALL J., TRONCOSO-PASTORIZA J., HOFMANN O. & KEANE T. M. (2021), "Crypt4GH: a file format standard enabling native access to encrypted data", *Bioinformatics (Oxford, England)* 37, pp. 2753-2754. Advance online publication, https://doi.org/10.1093/bioinformatics/btab087

Enjeux épistémologiques de la science des données

Par Jean-Gabriel GANASCIA
Spécialiste en intelligence artificielle

Après avoir rappelé ce qui fait la singularité des « masses de données », laquelle ne tient pas uniquement à leur volume, mais aussi à leur évolutivité et à leur variabilité, nous montrerons que tant leur accumulation que leur exploitation se sont révélées nécessaires pour les grands acteurs du Web et que cela tient à trois raisons liées à la spécificité des industries du numérique. Nous amorcerons ensuite une réflexion sur la science des données et sur l'opposition entre, d'un côté, ceux qui affirment que désormais les corrélations suffisent et, de l'autre, ceux qui s'en tiennent toujours à l'emploi de modèles et à la fonction épistémologique clef qu'ils occupent dans la démarche scientifique. Nous conclurons sur l'absence actuelle de cadre théorique mathématique de la science des données, tout en évoquant les théories anciennes, celles qui existaient dans les années 1990, et en ouvrant sur des progrès en ce sens.

Données et masses de données

Ancienneté des données

La notion de données n'est pas nouvelle et leur recueil systématique ne l'est pas plus. Comme l'ont déjà rappelé Viktor Mayer-Schonberger et Cukier (2013), dès la plus haute antiquité, au temps des Assyriens, on procédait à des statistiques ; par ailleurs, la Bible évoque le recensement ordonné par le roi David de la population de son royaume en vue de lever l'impôt. Depuis l'avènement des premiers ordinateurs, les données représentées sous forme numérique binaire, comme des suites de 0 et de 1, se manipulent de façon automatique. Cette numérisation ne se limite pas à des données quantifiées ; des textes, des images, des sons, des vidéos, voire des sensations kinesthésiques se traduisent, eux aussi, en flux d'informations et sont traités comme tels. Là encore, rien de nouveau : il y a bientôt quatre-vingts ans, Vannevar Bush, Alan Turing, Norbert Wiener et d'autres encore en eurent l'intuition. Bref, les notions de données, de recueil systématique, de numérisation et de traitement automatique de celles-ci paraissent aujourd'hui bien anciennes.

Innovation et masses de données

L'innovation tient à l'automatisation et à la massification du recueil des données par le truchement de machines. Aujourd'hui, des capteurs robotisés enregistrent et stockent des quantités de données. Des observations de toutes sortes – images, sons, températures – se font sans intervention humaine. À cette automatisation de la saisie s'ajoute la production d'annotations collectives centralisées grâce à l'interconnexion de la population sur le Web. Ce que l'on appelle le *crowdsourcing* – littéralement la « collecte par la foule» –, met à contribution de très nombreuses personnes qui collaborent, consciemment ou inconsciemment, à ce recueil de données. Ainsi, chaque requête faite *via* un moteur de recherche comme Google (et, *a fortiori*, chaque achat) est automatiquement exploitée.

Ordres de grandeur

Pour nous faire une idée du changement quantitatif intervenu, donnons quelques ordres de grandeur en prenant pour référence deux unités bien connues : le livre et la bibliothèque. Si l'on considère, en faisant abstraction des images, qu'un livre compte à peu près un million de caractères typographiques – estimation plutôt généreuse –, et qu'un caractère typographique se code sur un octet, on peut dire qu'il « pèse » un million d'octets d'informations, soit un mégaoctet (1 Mo = 10^6 octets). Le catalogue des livres et imprimés de la Bibliothèque nationale de France comprend à peu près quatorze millions d'ouvrages. En reprenant cette estimation de 10^6 octets pour un livre, le poids de l'ensemble des ouvrages référencés au catalogue de la BNF correspond donc à peu près à 14 x 10^{12} octets, soit quatorze téraoctets (To). Ce chiffre assez conséquent correspondait pour l'homme éclairé du XX^{e} siècle à l'horizon ultime du savoir. Mettons-le en perspective avec les capacités de stockage actuelles : un disque dur externe de vingt téraoctets valant moins de cinq cents euros aujourd'hui, il est donc désormais possible de posséder chez soi l'équivalent numérique de la BNF, pour un coût dérisoire comparé au prix de ce grand bâtiment ! Quant au volume total des données du Web, il a été évalué en 2020 à environ 47 zettaoctets (1 Zo = 10^{21} octets), ce qui fait 47 milliards de téraoctets, c'est-à-dire un peu plus de trois milliards de BNF.

Et ce chiffre s'accroît démesurément. Chaque jour, 500 millions de « gazouillis » sont échangés sur Twitter, ce qui, à raison de 140 caractères par message, engendre environ 25 To de données par an, soit bien

plus que le fonds documentaire de la Bibliothèque nationale de France ! Et encore, nous ne considérons ici que le texte des *tweets*, faisant donc abstraction des images et des sons que l'on y associe souvent. En outre, ceux-ci ne rassemblent qu'une très faible partie de ce que l'on échange sur Internet…

Ces éléments aident à se faire une idée grossière des caractéristiques quantitatives de ce que l'on appelle les *Big Data*. Mais le fait d'emmagasiner de grandes masses d'informations n'est pas leur seule caractéristique. On les définit souvent par la formule des « 3 V » – pour volume, vélocité et variété. Le volume, c'est-à-dire la quantité proprement dite, oscille entre le téraoctet (1 To = 10^{12} octets) et le pétaoctet (1 Po = 10^{15} octets = 1 000 To), à savoir entre un dixième du fonds de la BNF et l'équivalent de cent BNF. La vélocité renvoie au fait que cette masse de données se renouvelle en permanence. Enfin, les données sont variées au sens où elles sont hétérogènes : elles peuvent contenir du texte, des images, des sons, etc. Le texte lui-même peut intégrer différentes langues, divers systèmes d'abréviations, etc.

Particularités de l'économie du numérique

À ce contexte technique de l'automatisation du recueil et à la quantité vertigineuse des données collectées s'ajoutent trois caractéristiques de l'univers numérique qui rendent nécessaire l'accumulation de grandes masses d'informations. Elles tiennent aux spécificités de l'industrie du logiciel, à la vulnérabilité de l'information, sujette aux rumeurs (aux *fake news*) et à l'aspiration à la gratuité qui fut à l'origine de l'essor du Web. Précisons ces trois points pour mieux comprendre la situation actuelle.

Retours d'usage

À la différence des objets techniques traditionnels, comme les voitures, les machines à laver ou les montres, les logiciels apparaissent d'une complexité telle que les ingénieurs qui les conçoivent ne parviennent jamais à les réaliser parfaitement, d'un seul coup, en envisageant toutes les situations dans lesquelles on les emploiera. Pour les aider, sont impliqués les utilisateurs afin qu'ils contribuent à l'amélioration des logiciels en faisant part de leurs impressions et en indiquant les erreurs de fonctionnement qu'ils ont constatées. Ce faisant, les industries du logiciel récupèrent ce que l'on appelle les « retours d'usage », c'est-à-dire tous les dysfonctionnements déplorés par leurs clients, et ce à l'échelle planétaire. Elles amoncellent alors des masses considérables d'informations qu'il leur faut absolument être en mesure de traiter.

Détection de signaux faibles

Alors que les empires industriels d'antan reposaient sur des équipements coûteux – hauts fourneaux, mines, fabriques, usines, etc. –, les industries du numérique recourent essentiellement à la matière grise d'ingénieurs bien formés. L'investissement paraît donc désormais très léger dès lors qu'il repose plus sur des hommes que sur des infrastructures. En contrepoint, de nouveaux acteurs apparaissent rapidement et d'autres disparaissent tout aussi rapidement, car la réputation et la pression sociale jouent un rôle considérable pour assurer la fidélité des utilisateurs : on recourt à tel moteur de recherche et l'on ouvre un compte sur tel réseau social simplement parce que d'autres, en qui nous avons confiance, l'ont fait. Si des informations tendent à discréditer tel ou tel acteur, nous pouvons très rapidement le quitter ayant perdu confiance en lui. De ce fait, toute rumeur qu'on laisse enfler risque potentiellement de déstabiliser les empires les plus puissants. Il est de nombreux exemples de telles déstabilisations, comme celui de la société Dell qui a perdu sa position dominante sur le marché des micro-ordinateurs, parce qu'elle n'a pas su écouter les récriminations des utilisateurs de ses matériels (Jarvis, 2011). En conséquence, il importe aujourd'hui pour un industriel du numérique de récupérer tous les bruits qui le concernent et de les traiter au plus tôt pour mettre en œuvre des stratégies de communication visant à apporter les réponses attendues et à le faire savoir, par exemple en laissant entendre qu'il va changer sa façon de traiter les données personnelles dans le but de prendre soin de ses utilisateurs. Cette auscultation permanente du cyberespace pour parer aux commentaires désobligeants émanant soit de clients sincères, soit d'adversaires résolus, conduit, là encore, à une accumulation considérable de données qu'il faut être en mesure de traiter.

Économie de la gratuité

Enfin, la troisième caractéristique de l'univers du Web vient de la part importante qu'y prend la gratuité. Rappelons qu'avec le Minitel, la France disposait d'une avance technologique dans la mise en place des réseaux de télécommunications. C'est dans ce contexte qu'elle a appris à facturer des services, comme la consultation des banques de données, des horaires SNCF ou des sites de rencontre. L'engouement pour le Web, dans la seconde moitié des années 1990, tint donc non pas à l'absence de services équivalents, puisqu'ils existaient auparavant, du moins en France, mais à leur gratuité : grâce au Web, nous pouvions tout faire, sans rien débourser ! Cependant, pour rentabiliser leurs activités, les acteurs se devaient de trouver des ressources sur un modèle compatible avec la gratuité, ce qui paraissait une gageure. Très tôt, on fit de la publicité. Mais cette publicité tous azimuts restait peu rémunératrice ; elle l'était d'autant moins qu'elle s'adressait à des consommateurs de tous âges et de tous pays, puisque le Web était mondial, et que son accumulation la rendait illisible. Apparut alors le besoin de cibler les annonces. À cette fin, on récupéra toutes les informations utiles pour déterminer automatiquement le profil du consommateur qui sommeille en chacun de nous et, par là même, accroître dans des proportions considérables l'efficacité des messages publicitaires. Là encore, l'accumulation et le traitement de masses de données considérables paraissaient indispensables pour les grands acteurs du Web, et les résultats obtenus à l'aide de techniques d'intelligence artificielle en démontrent aujourd'hui toute l'efficacité.

Science sans causalité, ni modèle

Fin de la théorie ?

Un journaliste très influent, Chris Anderson (2008), proclamait en 2008, dans un article publié par la revue *Wired,* dont il était le directeur de rédaction, « la fin de la théorie » et, par voie de conséquence, l'obsolescence de la méthode scientifique attachée à la preuve et à la compréhension. Son argumentation ne relevait pas vraiment de la démonstration, mais plutôt de l'hyperbole : selon lui, les masses immenses de données collectées permettaient aux grands acteurs du Web, comme Google, de conquérir sans coup férir le monde de la publicité. Face à ce succès sans précédent, d'autres champs de l'activité humaine, en particulier la science, devaient dès lors « tomber » sous la coupe des méthodes de l'intelligence artificielle. De ce fait, on rangera bientôt le raisonnement, la déduction, la logique et la pensée au magasin des oubliettes comme autant de vieilleries inutiles. Il ne faut pas rire de ce constat ! L'article a en effet été cité à maintes reprises comme révélateur d'un tournant épistémologique majeur de la modernité. Et, la preuve étant dépassée, la popularité de cet article atteste de sa plausibilité, et cela seul suffit.

Utilité pratique des corrélations

Les techniques d'apprentissage machine détectent très efficacement des corrélations entre différents paramètres, ce qui rend de grands services dans des secteurs comme la publicité, puisqu'il suffit d'établir des liens entre les comportements des consommateurs et leurs achats pour accroître l'efficacité des annonces publicitaires. De même, ces techniques aident à traduire des textes en constatant que, la plupart du temps, telle locution placée dans tel contexte et se rattachant à telle langue se traduit par telle autre locution dans une autre langue. Elles peuvent aussi inspirer, à titre heuristique, les scientifiques en mettant en évidence telle ou telle hypothèse de travail, par exemple, en suggérant, par la détection de corrélations présentes dans d'immenses quantités d'observations, les effets secondaires de tel médicament ou les facteurs de risques associés à tel ou tel comportement, ou encore en réalisant un diagnostic médical à partir d'une image, comme celui d'un mélanome à partir de photographies de grains de beauté (Esteva *et al*., 2017).

Nécessité de preuves

Cependant, comme le dit Raymond Aron dans sa préface de l'ouvrage de Max Weber intitulé *Le Savant et le Politique*, « la vocation de la science est inconditionnellement la vérité ». Et cette quête de vérité ne saurait se satisfaire de simples corrélations, fussent-elles souvent vérifiées ; les relations causales, les mécanismes explicatifs et, surtout, les preuves demeurent essentiels à la compréhension ; à défaut, nous ne disposerons que de conjectures.

Ainsi, et contrairement à ce qu'affirme Chris Anderson, la notion de preuve n'a jamais été remise en question par le traitement de grandes masses de données. Un exemple en convaincra aisément : il existe une corrélation avérée entre l'application réitérée de crèmes solaires et l'apparition de cancers de la peau. Doit-on pour autant interdire les crèmes solaires ? Sans une expérimentation contrôlée effectuée dans les règles, donc en maîtrisant tous les facteurs, en particulier l'exposition au soleil, une telle conclusion ne pourrait être réfutée, ce qui se révélerait néfaste, car plus personne n'oserait dès lors utiliser des crèmes solaires pour se protéger…

Besoin de théorie

Les faits ne couvrent jamais l'intégralité de l'espace des possibles : le monde brut se donne par l'intermédiaire d'une représentation à travers laquelle on le décrit ; il se donne aussi en fonction de facteurs spécifiques qui restreignent la répartition des observations. À titre d'illustration, si l'on se contentait de collationner des faits bruts, il serait difficile d'imaginer que nous observions autant d'exemples de personnes qui mettent des crèmes solaires sans s'exposer au soleil que de personnes qui en mettent et s'exposent ensuite au soleil… Il existe toujours, qu'on le veuille ou non, que l'on en soit ou non conscient, un biais dans les données. Pour corriger celui-ci, on doit formuler clairement des hypothèses théoriques et rassembler les observations, c'est-à-dire les données, eu égard à ces hypothèses, afin de les valider.

À l'évidence, le traitement de grandes masses de données par les ordinateurs transforme la méthode scientifique. Mais l'affirmation selon laquelle la corrélation supplante la causation n'apparaît pas fondée. Et il en va de même des affirmations selon lesquelles la science avance désormais sans s'appuyer sur des modèles cohérents, sans une quête de théories unifiées et sans recours à des mécanismes explicatifs.

Approches théoriques de la science des données

Alors que, dans le courant des années 1990, l'apprentissage machine reposait sur des théories mathématiques, comme la théorie de l'apprentissage statistique de Vladimir Vapnik (1998) ou l'approche PAC – probably approximately correct – de Leslie Valiant (1984), aujourd'hui l'apprentissage profond (*deep-learning*) ne dispose pas d'autre justification que la constatation empirique d'une remarquable efficacité statistique que jusqu'ici personne n'est parvenu à expliquer clairement en termes mathématiques. À cela s'ajoute l'opacité des conclusions obtenues par les machines entraînées avec de l'apprentissage profond. En effet, il est, la plupart du temps, impossible de comprendre dans chaque situation, ce qui justifie ces conclusions, car celles-ci se présentent comme dérivant de combinaisons pondérées d'un très grand nombre de facteurs. Cette double incapacité à interpréter tant la capacité des machines à apprendre que leurs conclusions fait écho au caractère inexplicable de la science contemporaine qu'évoque Chris Anderson.

Tout cela paraît symptomatique d'une tendance actuelle visant à réduire la vérité à un calcul si complexe que seuls des ordinateurs parviendraient à l'exécuter. En affirmant que nous passons de l'humanisme, c'est-à-dire de la religion de l'humain, au « dataïsme », à savoir la domination de l'homme par des ordinateurs abreuvés de *data*, Noam Yuval Harari (2017), dans son livre *Homo Deus,* révèle un penchant analogue. Dans l'un et l'autre cas, cela revient à se laisser dominer par des machines, en se rangeant systématiquement à leurs conclusions, sans les discuter et, par là même, à renoncer à la raison, à savoir à la capacité de l'homme à appréhender le réel par sa pensée, à renoncer à l'argumentation et au débat face à ces oracles que seraient devenus les ordinateurs programmés à l'aide de techniques d'intelligence artificielle… Or, derrière ces renoncements, plus qu'une conception épistémologique nouvelle, se cache surtout une démission politique face aux pouvoirs des nouveaux acteurs qui maîtrisent ces machines ! Et rien ne dit qu'une théorie mathématique formelle de l'induction, à partir de grandes masses de données, n'adviendra pas dans le futur…

Références bibliographiques

ANDERSON C. (2008), "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", *Wired*, June 23, https://www.wired.com/2008/06/pb-theory/

ESTEVA A., KUPREL B., NOVOA R. A., KO J., SWETTER S. M., BLAU H. M. & THRUN S. (2017), "Dermatologist-level classification of skin cancer with deep neural networks", *Nature* 542(7639), pp. 115-118.

HARARI Y. N. (2017), *Homo deus : une brève histoire du futur*, traduction de l'anglais de Pierre-Emmanuel Dauzat, Albin Michel.

JARVIS J. (2011), *La méthode Google : que ferait Google à votre place ?*, Pocket.

MAYER-SCHONBERGER V. & CUKIER K. (2013), *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, John Murray, traduction française : *Big Data : la révolution des données est en marche*, Robert Laffon.

VALIANT L. G. (1984), "A Theory of the Learnable", *Communication of the ACM*, vol. 27, n°11, November, pp. 1134-1142.

VAPNIK V. (1998), *Statistical Learning Theory*, New York, Wiley-Interscience.

Le chiffrement, ou l'apport de la cryptologie à la sécurisation du stockage, de la transmission et du traitement des données

Par Louis GOUBIN
Professeur à l'Université Versailles Saint-Quentin-en-Yvelines – Université Paris-Saclay, directeur du groupe de recherche « Cryptologie et sécurité de l'information » au laboratoire LMV (UMR CNRS 8100)

Le développement des techniques nécessaires au stockage, à la transmission et au traitement des données numériques crée un besoin de plus en plus aigu de sécurisation de ces données. La cryptologie, que l'on appelle souvent la science du secret, apporte des réponses solides, souvent des preuves mathématiques, à la question de la confidentialité des données, et par là même à la protection de la vie privée. Nous proposons ici une excursion à travers la problématique du chiffrement des données, depuis les principes cryptographiques de base jusqu'à des applications plus complexes nécessitant de pouvoir effectuer des calculs sur les données chiffrées. Des perspectives nouvelles s'ouvrent grâce à des techniques récentes ; l'un des grands défis à relever en la matière est d'arriver à articuler sécurité et conformité réglementaire, notamment lorsqu'il s'agit de chaînes d'approvisionnement dans l'industrie ou de plateformes mettant en œuvre une analyse parfois inquiétante des données personnelles de leurs utilisateurs.

Le contexte du chiffrement

« Le système doit être matériellement, sinon mathématiquement, indéchiffrable. Il faut qu'il n'exige pas le secret, et qu'il puisse sans inconvénient tomber entre les mains de l'ennemi. [...] Il faut qu'il soit portatif, et que son maniement ou son fonctionnement n'exigent pas le concours de plusieurs personnes » (Auguste Kerckhoffs, *La cryptographie militaire*, 1883 [6]).

On peut considérer que les principes édictés par Kerckhoffs à la fin du XIX^{e} siècle marquent la naissance de la cryptologie, au sens moderne du terme. On peut en tout cas les apercevoir en filigrane tout au long du développement de la théorie et de la pratique du chiffrement, qui a accompagné l'invention de techniques de plus en plus cruciales pour le stockage, la transmission et le traitement des données. Le texte qui suit invite à une excursion – forcément trop brève ! – à travers la problématique du chiffrement, depuis les principes cryptographiques de base jusqu'à des applications plus complexes nécessitant de pouvoir effectuer des calculs sur des données chiffrées.

Tout d'abord, intéressons-nous à la cryptologie. Il est maintenant naturel de postuler que le système « puisse sans inconvénient tomber entre les mains de l'ennemi », et donc de considérer des algorithmes cryptographiques connus de tous, les informations secrètes étant concentrées dans une clé. Longtemps limité à la cryptographie symétrique, ce principe – poussé à l'extrême – a donné naissance à la notion de cryptographie asymétrique, sachant que la nature ouverte des systèmes les rend en outre bien plus vulnérables, puisque l'attaquant peut dans certains cas avoir un contrôle complet sur la plateforme d'exécution et l'implémentation logicielle elle-même.

« Il faut qu'il soit portatif » : la cryptographie est de plus en plus déployée dans les applications exécutées sur des périphériques portables, tels que des ordinateurs portables, des tablettes ou des *smartphones*. On ne pourrait trouver plus belle illustration de ce contexte que la carte à microprocesseur (la « carte à puce »[1]). C'est encore le meilleur moyen que l'on ait trouvé pour garder secrète une clé au sein d'un dispositif embarqué, tout en proposant des capacités réelles de calcul cryptographique.

[1] Rendons ici hommage à Michel Ugon, disparu le 28 décembre 2021. Ingénieur de génie et véritable père de la carte à puce, c'est en grande partie lui qui a fait de la France le berceau incontesté de cette technologie qui a inondé le monde.

« Le système doit être matériellement, sinon mathématiquement indéchiffrable » : cette phrase préfigure la théorie de la complexité, qui conduit à identifier des problèmes mathématiques « difficiles », au sens informatique du terme. Ce sont les briques de base à partir desquelles on peut construire de nouveaux cryptosystèmes aux propriétés inédites, notamment pour tenir compte des contraintes de mémoire ou de temps de calcul propres aux dispositifs mis en jeu. Le défi des chercheurs est alors de trouver un compromis acceptable entre le niveau de sécurité et le niveau de flexibilité du chiffrement.

La cryptologie

La cryptographie et la cryptanalyse

Pour analyser le sens du mot « cryptologie », il est utile de se référer à son étymologie, qui s'appuie sur les mots grecs χρυπτός (*kryptos* = caché) et λόγος (*logos* = discours, traité). On peut ainsi définir la cryptologie comme la science du secret : elle s'appuie sur la possibilité de transmettre un message, tout en dissimulant son contenu pour un observateur indiscret.

Plus généralement, dans sa conception moderne, la cryptologie a essentiellement pour objet l'étude de trois problématiques qui en constituent les véritables piliers [15] : la confidentialité, l'authenticité et l'intégrité de l'information. Pour répondre à ces trois exigences, trois concepts importants ont émergé au fil du temps et sont considérés aujourd'hui comme les fonctionnalités fondamentales de la cryptologie :

- le chiffrement, qui permet de cacher l'information contenue dans un message ;
- la signature électronique, qui permet de prouver l'identité de l'auteur d'un message et de garantir la non-répudiation ;
- l'authentification, qui permet de prouver son identité lors d'un contrôle d'accès.

Pour compléter cette typologie, remarquons qu'il est d'usage de distinguer deux facettes de la cryptologie :

- La cryptographie, qui consiste à concevoir et à mettre au point des mécanismes cryptologiques adaptés afin d'assurer une ou plusieurs des trois notions de la sécurité décrites précédemment[2]. Concrètement, ces mécanismes sont en général décrits de façon algorithmique et font appel à des notions mathématiques, allant des probabilités à la théorie des nombres, en passant par la théorie de la complexité, la combinatoire, la théorie des codes correcteurs d'erreurs, la théorie de la réduction des réseaux, les corps finis, les polynômes multivariables, les courbes elliptiques et hyperelliptiques, la géométrie algébrique, etc.
- La cryptanalyse, qui consiste à évaluer la résistance des méthodes mises au point par la cryptographie pour contrer les attaques. Une partie consiste à définir des scénarios d'attaque et à les appliquer pour mettre à l'épreuve les algorithmes. Il peut s'agir d'attaques purement mathématiques ou bien d'attaques faisant, en outre, intervenir la façon dont l'algorithme est implanté dans un système électronique (on parle alors d'attaques physiques).

Le chiffrement symétrique

Dans la cryptographie symétrique, les clés de chiffrement et de déchiffrement sont identiques (voir la Figure 1 ci-dessous). Elles doivent donc toutes deux être gardées secrètes, ce qui fait que l'on parle également ici de cryptographie à clé secrète.

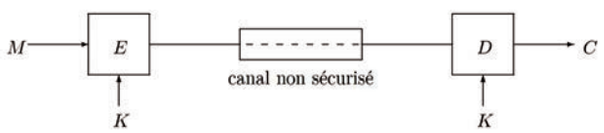


Figure 1 : Le chiffrement symétrique – Schéma général (la clé secrète commune est notée *K*).

C'est le domaine le plus ancien de la cryptographie (il est aussi connu sous le nom de cryptographie conventionnelle). De nombreux algorithmes appartiennent à cette catégorie, dont les deux plus importants actuellement sont évoqués ci-après.

En 1977, le gouvernement américain[3] a publié et standardisé l'algorithme DES (Data Encryption Standard [8,9]), qui a fait l'objet d'un effort de cryptanalyse considérable depuis sa conception. Il est toujours considéré comme excellent : ainsi, même si la taille de la clé (56 bits) s'avère aujourd'hui trop courte, la variante appelée Triple-DES est encore utilisée pour certaines applications, en particulier pour les transactions bancaires.

L'algorithme AES (Advanced Encryption Standard [1, 10]) a progressivement remplacé le DES depuis sa standardisation en 2001. Conçu pour être à la fois plus rapide et plus polyvalent que le Triple-DES, il a fait l'objet d'une grande attention parmi les chercheurs. À ce jour, aucune attaque efficace contre lui n'a été identifiée.

La cryptographie asymétrique

Un nouveau paradigme

La cryptographie asymétrique repose sur le recours à deux clés nécessairement différentes, d'où le qualificatif d'asymétrique. Dans son principe, cette technologie de cryptographie consiste à rendre publique la clé qui sert à la fonction de chiffrement. La clé de déchiffrement doit, quant à elle, rester bien entendu secrète sous peine de perdre complètement tout espoir d'assurer la confidentialité des messages. Cette branche de la cryptologie est souvent appelée également cryptographie à clé publique.

[2] Ainsi que d'autres objectifs plus spécialisés, tels l'anonymat, le *broadcasting*, le *traitor tracing*, etc.

[3] Plus précisément, le National Bureau of Standards (NBS), qui est l'ancêtre du NIST (le National Institute of Standards and Technology).

Remarquons qu'un attaquant disposant du message chiffré C connaît tout à la fois la fonction qui a servi à chiffrer ce message et la clé (publique) qui a été utilisée (voir la Figure 2 ci-après). C'est même encore plus inquiétant, puisqu'en général, il existe une relation connue entre la clé publique et la clé privée ! La situation semble donc paradoxale ; et avec le recul, on s'aperçoit qu'une solution n'a pu émerger qu'avec le développement de la théorie de la complexité, qui rend plausible l'existence de problèmes mathématiques intrinsèquement difficiles. C'est ainsi qu'en théorie (c'est-à-dire en supposant que l'ennemi dispose d'une puissance de calcul infinie), il n'est pas impossible de reconstituer le message M en clair à partir de son équivalent chiffré C. Mais, en pratique, un tel décryptage sous-entend une puissance de calcul que l'on espère suffisamment grande pour dépasser les capacités supposées de tout attaquant.

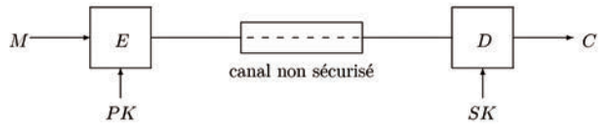


Figure 2 : Le chiffrement asymétrique – Schéma général (la clé publique et la clé privée sont notées respectivement *PK* et *SK*).

Ainsi, ce n'est qu'en 1976, que Whitfield Diffie et Martin Hellman, dans leur célèbre article fondateur [2], ont montré la possibilité théorique de la cryptographie à clé publique, en l'illustrant dans le cas particulier d'un protocole d'échange de clés[4]. L'algorithme RSA est reconnu comme étant le premier algorithme publié[5] de chiffrement asymétrique réellement utilisable.

Par rapport aux systèmes symétriques, le chiffrement à clé publique présente le grand avantage de ne pas nécessiter un accord préalable entre les interlocuteurs qui souhaitent échanger des messages. Mais la plus grande nouveauté apportée par la cryptographie asymétrique est la possibilité de concevoir des protocoles d'authentification et de signature, répondant ainsi de manière spectaculaire aux besoins d'intégrité et d'authenticité.

L'exemple du schéma RSA

L'algorithme RSA, inventé par Ronald Rivest, Adi Shamir et Leonard Adleman [12], a été présenté publiquement pour la première fois dans le numéro d'août 1977 de la revue *Scientific American* [4]. C'est encore actuellement le cryptosystème à clé publique le plus utilisé dans le monde. On le retrouve dans un très grand nombre de produits commerciaux liés à la sécurisation des échanges de données sur Internet, à la protection de la confidentialité et de l'authenticité des courriers électroniques, au paiement électronique au moyen de cartes à puce, etc.

Mathématiquement, on peut décrire l'algorithme RSA de la manière suivante. On commence par choisir l'exposant public e[6]. On utilise ensuite un générateur de nombres aléatoires pour obtenir deux nombres premiers p et q, tels que e soit premier avec p - 1 et avec q - 1. Si l'on pose $n = p \times q$, la clé publique est alors constituée de e et de n, alors que la clé secrète (ou privée) est composée de p et q. La fonction de chiffrement est alors définie par : $f{:}x \rightarrow y = x^e \bmod n$, et la fonction de déchiffrement par : $f^{-1}{:}y \rightarrow x = y^d \bmod n$, où d est une valeur qui doit rester secrète.

La fonction f est donc conçue pour être facilement inversible lorsque l'on connaît la « trappe » d. Casser la fonction RSA consiste à trouver un moyen de calculer $f^{-1}(y)$, alors que l'on ne dispose pas de l'exposant secret d. La seule stratégie d'attaque connue consiste à retrouver p et q à partir de n. De nombreux algorithmes spécialisés ont été inventés pour résoudre ce problème de factorisation. La Figure 3 ci-après illustre la puissance nécessaire[7] pour mettre en œuvre la meilleure méthode connue, en fonction de la taille du modulo n. Empiriquement, on fait généralement l'hypothèse (connue sous le nom de « Loi de Moore ») que la puissance des ordinateurs double tous les 18 mois. On peut alors prédire (s'il n'y a pas de découverte théorique nouvelle concernant les techniques de factorisation[8]) que les clés RSA de 1 024 bits seront cassées vers l'an 2034, et celles de 2 048 bits vers l'an 2079...

Nombre de Mips.ans disponibles	Taille maximale des clés RSA «factorisables»
$4,8 \times 10^{-2}$	251
$4,9 \times 10^{-1}$	292
$4,9$	337
$5,0 \times 10^{1}$	385
$5,0 \times 10^{2}$	438
$5,0 \times 10^{3}$	494
$\mathbf{1,0 \times 10^{4}}$	512
$5,1 \times 10^{4}$	555
$5,1 \times 10^{5}$	620
10^{8}	784
10^{11}	1035
10^{16}	1551
10^{20}	2057

Figure 3 : Puissances nécessaires pour factoriser le modulo RSA.

(4) Notons qu'un protocole présentant les mêmes propriétés avait été décrit un peu auparavant par Ralph Merkle [7]. Les *puzzles de Merkle* affichaient toutefois un moins bon rapport sécurité/performance.

(5) On sait aujourd'hui [3] qu'au sein du service du chiffre britannique, James Ellis avait établi, dès janvier 1970, la possibilité de la cryptographie asymétrique. Par la suite, Clifford Cocks a inventé en 1973 une variante du RSA, avant que Malcolm Williamson décrive à son tour, en 1974, une variante du futur protocole d'échange de clés Diffie-Hellman.

(6) Des exemples courants sont e = 3, e = 17, e = 257 ou e = 65537.

(7) Mips signifie « million d'instructions élémentaires par seconde ». 1 Mips.an représente le nombre d'instructions élémentaires exécutées par une machine qui en exécute 1 million par seconde, et que l'on fait tourner pendant 1 an. Ainsi, 1 Mips.an équivaut plus ou moins à $31{,}5.10^{12}$ instructions élémentaires.

(8) Ni de saut technologique tel que l'apparition d'ordinateurs quantiques qui pourraient factoriser efficacement, comme l'a montré Peter Shor en 1994 [14].

Utilisation du RSA pour le chiffrement

Même si la fonction RSA est solide, la façon dont on l'utilise pour obtenir un cryptosystème capable d'effectuer du chiffrement n'est pas neutre. De manière générale, il doit non seulement être impossible, en pratique, de retrouver un message clair à partir de son équivalent chiffré – hormis, bien entendu, pour l'utilisateur légitime du système –, mais il doit également être impossible de connaître la moindre information sur le message clair (ou sur la signature). C'est ce que l'on appelle la sécurité sémantique.

Il est facile de comprendre que la fonction RSA n'est pas suffisante à elle seule pour garantir cette sécurité. La fonction f jouit en effet d'une propriété de multiplicativité : $f(x.y) = f(x).f(y)$, ce qui ouvre la voie à certaines attaques comme celle élaborée par Johan Håstad en 1988, lequel a montré qu'un attaquant peut toujours retrouver le message M, si celui-ci est envoyé à un nombre suffisant de destinataires utilisant le même exposant public e.

Il faut par conséquent faire très attention à la manière dont on applique la fonction RSA pour chiffrer un message M. Dans la pratique, on commence par « formater » le message M au moyen d'une transformation φ qui fait intervenir un nombre aléatoire. Ainsi, le message chiffré est obtenu sous la forme : $C = f(\varphi(M,r))$, où r est une valeur aléatoire. Pour déchiffrer, on calcule $\varphi(M,r) = f^{-1}(C)$, où φ a été conçue pour qu'il soit facile de retrouver M à partir de $\varphi(M,r)$. La théorie des protocoles de chiffrement RSA est aujourd'hui bien maîtrisée, au point que l'on connaît aujourd'hui des transformations φ qui sont prouvées comme étant « saines ». En d'autres termes, si la fonction f est mathématiquement solide, et que l'on utilise l'une de ces « bonnes » transformations de φ, alors il n'existe pas d'attaque mathématique plus simple que celle consistant à trouver un moyen d'inverser la fonction f[9].

Réaliser des calculs sur des données chiffrées ?

Le chiffrement homomorphe

Les données que l'on considère sont, en général, dans l'un des trois états suivants : au repos, en transit ou en cours d'utilisation. Les algorithmes de chiffrement évoqués jusqu'ici se placent dans les deux premiers cas : les données au repos ou en transit ne changent pas activement ; elles ont la même valeur lorsqu'on les déchiffre que quand on les a chiffrées.

En revanche, les données en cours d'utilisation n'ont pas cette propriété. Presque toutes les opérations que l'on peut envisager sur les messages chiffrés modifient la valeur du message correspondant converti en clair. Il est dès lors difficile de s'assurer que ce message en clair change de la « bonne manière ». Habituellement, il s'agit même d'une exigence de sécurité pour les algorithmes de chiffrement, qui sont conçus pour détruire toute relation entre le message en clair et le message chiffré correspondant.

Inversement, la possibilité d'effectuer des opérations mathématiques sur des données chiffrées implique qu'il doit exister une relation entre les messages en clair et les messages chiffrés. Il doit être possible d'ajouter deux messages chiffrés ou de les multiplier et d'avoir le même résultat que celui chiffré obtenu en effectuant la même opération (addition ou multiplication) sur les deux messages en clair correspondants. En outre, une telle opération doit être réalisée de façon à rester cachée pour un observateur, au sens où l'observation de telles opérations mathématiques réalisées sur des messages chiffrés ne doit rien révéler sur les messages en clair correspondants.

Du point de vue cryptographique, il s'agit de construire un schéma de chiffrement complètement homomorphe (en anglais, *Fully Homomorphic Encryption*, FHE), c'est-à-dire permettant à tout utilisateur de calculer, à partir d'un ensemble de données chiffrées : $c_1, \ldots, c_n$ (correspondant à des données en clair : $m_1, \ldots, m_n$), une donnée chiffrée c correspondant à une certaine fonction $F(m_1, \ldots, m_n)$ des données en clair, sans que cet utilisateur connaisse ces données elles-mêmes (voir la Figure 4 de la page suivante). Bien que Ronald Rivest, Leonard Adleman et Michael Dertouzos aient conjecturé l'existence d'un tel schéma dès 1978 [11], il a fallu attendre 2009 pour que Craig Gentry [5] propose la première solution convaincante[10]. D'autres solutions sont alors apparues rapidement, tirant parti de la notion mathématique de réseaux euclidiens, qui avaient déjà fait leurs preuves pour la conception de nombreuses primitives cryptographiques[11].

Il est toutefois à noter que la tension entre les conditions techniques de ces constructions[12] et la sécurité des problèmes difficiles sous-jacents, oblige à choisir des paramètres très grands pour dimensionner le système, ce qui résulte en des schémas de chiffrement d'une grande complexité[13]. Le défi – encore actuellement – est d'améliorer ces schémas[14] pour les rendre pratiques, tout en restant sûrs.

[9] Précisons que ces résultats sont vrais sous l'hypothèse que les fonctions de hachage utilisées dans la définition de φ sont elles-mêmes, dans un certain sens, « parfaites » (c'est ce que l'on appelle le *modèle de l'oracle aléatoire*).

[10] Remarquons que si l'on se limite à un seul type d'opérations à réaliser sur les chiffrés (on parle de chiffrement partiellement homomorphe), des solutions étaient déjà bien connues, à commencer par la fonction RSA qui est homomorphe pour la multiplication.

[11] Et qui sont à la base – comme les systèmes de polynômes multivariés, les codes correcteurs d'erreurs et les isogénies de courbes elliptiques – de nouveaux algorithmes dits post-quantiques, c'est-à-dire résistants face à d'éventuelles attaques menées par des individus équipés d'ordinateurs quantiques.

[12] En particulier, ceux liés à l'opération de *bootstrapping* introduite par Craig Gentry.

[13] Par exemple, dans la version de la bibliothèque de calcul HElib C++ publiée en 2018 par IBM, les opérations portant sur les chiffrés sont environ 1 million de fois plus lentes que les mêmes opérations réalisées sur les textes en clair correspondants. Un calcul qui prendrait une seconde dans le cas de textes en clair prendrait en moyenne 11,5 jours pour être réalisé en recourant à la version 2018 de HElib.

[14] La bibliothèque de calcul HElib d'IBM a ainsi gagné un facteur 100 millions entre 2015 et 2018.

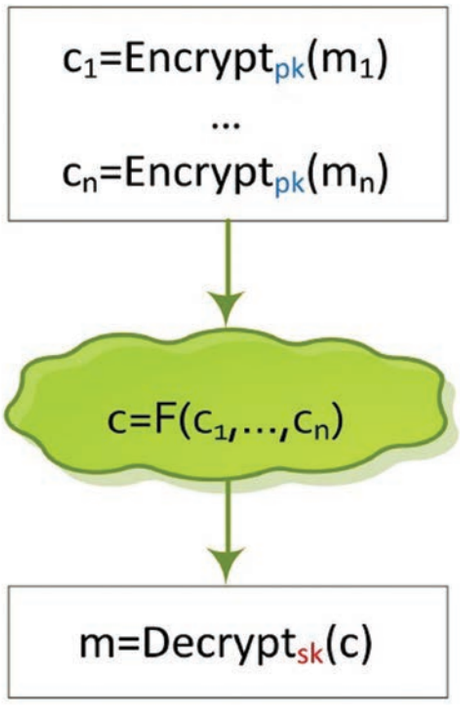


Figure 4 : Principe du chiffrement complètement homomorphe (FHE).

Vers de nouvelles applications

Il est facile de voir qu'un tel schéma peut servir à construire des protocoles respectant la vie privée (*privacy-preserving*) : un utilisateur peut ainsi stocker des données chiffrées sur un serveur[15] et autoriser ce même serveur à effectuer des opérations sur ces données, sans avoir à lui révéler la teneur même de ces données.

Plus concrètement, cette capacité à pouvoir effectuer un traitement sur des données chiffrées a pour potentiel de résoudre de nombreux défis commerciaux majeurs auxquels sont confrontées les entreprises de tous les secteurs.

La plupart des entreprises font appel à des tiers de confiance dans le cadre de l'exercice de leurs activités. Ces sous-traitants, fournisseurs, etc. ont souvent besoin d'accéder aux données sensibles et exclusives de l'entreprise pour pouvoir faire leur travail. Des événements récents ont illustré les risques liés à des chaînes d'approvisionnement non sécurisées et montré comment les cybercriminels ciblent le maillon le plus faible de celles-ci pour atteindre leurs objectifs. Cela signifie pour une organisation que confier ses données sensibles à un partenaire peut l'exposer à des vols de données qui peuvent être pour elle coûteux et dommageables. Le chiffrement homomorphe peut l'aider à se protéger contre ces risques liés à des failles de sécurité dans la chaîne d'approvisionnement : si toutes les données fournies à un tiers de confiance pour opérer un traitement sont chiffrées, un vol de ces données ne présente dès lors qu'un risque minime pour l'entreprise. Cela permet à une organisation d'externaliser le traitement de ses données critiques avec un risque minimal.

Ces dernières années, le paysage réglementaire de la protection des données est devenu de plus en plus complexe. De nouvelles réglementations, telles que le Règlement européen sur la protection des données (RGPD), ont accordé aux personnes concernées de nouveaux droits et imposé des responsabilités et des restrictions supplémentaires aux entreprises. Une règle du RGPD avec laquelle de nombreuses entreprises sont aux prises est l'exigence que les données des citoyens de l'Union européenne (UE) restent au sein de l'UE ou ne puissent être utilisées que dans des pays ou des entreprises dont les normes de sécurité des données sont équivalentes à celles de l'UE. L'arrêt Schrems II de 2020 [13] a invalidé l'un des principaux moyens par lesquels les flux de données entre l'UE et les États-Unis étaient justifiés dans le cadre du RGPD, ce qui a causé d'importants problèmes à de nombreuses entreprises américaines comptant parmi leurs clients des citoyens de l'UE. Des lois comme le RGPD stipulent clairement que leurs exigences ne s'appliquent pas aux données chiffrées. Avec le chiffrement homomorphe, une entreprise pourrait potentiellement stocker et traiter des données en recourant à des systèmes se trouvant en dehors de l'UE, puis les déchiffrer en faisant appel uniquement à des serveurs situés dans des espaces géographiques répondant aux exigences du RGPD.

L'analyse de données est pour de nombreuses entreprises une façon pour elles de générer des revenus. Si des entreprises sont en mesure de fournir des services « gratuits », c'est parce qu'elles collectent des informations sur leurs utilisateurs, qu'elles traitent celles-ci et les vendent à des tiers à des fins de publicité ciblée. Cependant, cette monétisation des données personnelles est controversée. De nombreuses personnes sont mécontentes des pratiques de ces entreprises qui conduisent à créer des profils détaillés les concernant sans qu'elles n'aient de visibilité ni de contrôle sur les données collectées et sur la manière dont elles sont utilisées. Le chiffrement homomorphe fournit une solution potentielle à ce problème : une entreprise pourrait effectuer les analyses de données dont elle a besoin, sans avoir la possibilité de visualiser les données d'origine ou même d'y accéder. Si les clés de chiffrement sont contrôlées par les utilisateurs, alors cela ouvre la possibilité d'une publicité qui soit à la fois privée et ciblée.

Bibliographie

[1] DAEMEN J. & RIJMEN V., *AES proposal: Rijndael,* https://csrc.nist.gov/csrc/media/projects/cryptographic-standards-and-guidelines/documents/aes-development/rijndael-ammended.pdf

[2] DIFFIE W. & HELLMAN M. E. (1976), "New Directions in Cryptography", *IEEE Transactions on Information Theory*, vol. IT-22, pp. 644-654.

[3] ELLIS J., "The story of non-secret encryption", article écrit en 1977 et publié après la mort de l'auteur en 1997, https://cryptome.org/jya/ellisdoc.htm

[4] GARDNER M. (1977), "A new kind of cipher that would take millions of years to break", *Scientific American*, août, pp. 120-124.

[5] GENTRY C. (2009), "Fully homomorphic encryption using ideal lattices", in *Proc. of STOC*, pp. 169-178.

[15] Typiquement dans le contexte du *cloud computing* (ou « informatique en nuage »).

[6] KERCKHOFFS A. (1883), « La cryptographie militaire », *Journal des sciences militaires*, vol. IX, janvier, pp. 5-38, et février, pp. 161-191.

[7] MERKLE R. C. (1978), "Secure Communication over Insecure Channels", *CACM*, vol. 21, n°4, pp. 294-299.

[8] NATIONAL BUREAU OF STANDARDS (NBS), Data Encryption Standard (DES), Federal Information Processing Standards Publication (FIPS PUB) 46, Washington, DC, 1977.

[9] NATIONAL BUREAU OF STANDARDS (NBS), Data Encryption Standard (DES), Federal Information Processing Standards Publication (FIPS PUB) 46-3, Gaithersburg MD, 1999, http://csrc.nist.gov/publications/fips/fips46-3/fips46-3.pdf

[10] NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST), Advanced Encryption Standard, Federal Information Processing Standards Publication (FIPS PUB) 197, décembre 2001, http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf

[11] RIVEST R. L., ADLEMAN L. M. & DERTOUZOS M. L. (1978), *On Data Banks and Privacy Homomorphisms. In Foundations of Secure Computation*, Academia Press.

[12] RIVEST R. L., SHAMIR A. & ADLEMAN L. M. (1978), "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems", *Communications of the ACM*, vol. 21, n°2, pp. 120-126.

[13] Schrems II – Arrêt rendu par la Cour de justice de l'Union européenne dans l'affaire C-311/18 – Data Protection Commissioner contre Facebook Ireland Ltd et Maximillian Schrems, 16 juillet 2020, https://curia.europa.eu/juris/document/document.jsf?text=&docid=228677

[14] SHOR P. W. (1997), "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer", *SIAM Journal on Computing* 26, pp. 1484-1509.

[15] STERN J. (1998), *La Science du secret*, Odile Jacob.

Accompagner les chercheurs pour les aider à mieux gérer leurs données de recherche : le métier de *data librarian*

Par Laetitia BRACCO
Conservatrice des bibliothèques à l'Université de Lorraine

Les données de la recherche sont progressivement considérées comme des productions scientifiques à part entière, dans toutes les communautés disciplinaires. Dans le contexte de la Science ouverte, de plus en plus d'exigences sont émises par les financeurs et par les politiques publiques en général pour mieux produire, structurer, conserver et ouvrir ces données. L'accès aux données sous-jacentes aux publications est, en outre, de plus en plus demandé par les revues scientifiques, dans une logique d'intégrité et de transparence. Les personnels de l'information scientifique et technique, dans les bibliothèques universitaires comme dans d'autres structures, sont ainsi amenés à accompagner et à former les chercheurs à toutes ces problématiques. Pour mener à bien cette mission, une montée en compétences est nécessaire, ainsi que le développement de nouvelles activités, pour celles et ceux qui occupent ce métier émergent qu'est celui de *data librarian*.

Les données de la recherche sont un produit d'étude complexe à définir. La définition de l'OCDE fait néanmoins consensus : ce sont « des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche »[1]. On peut ainsi considérer que les données de la recherche constituent le matériau varié (allant de l'image au tableur, en passant par la vidéo) sur lequel s'appuient les chercheurs pour réaliser et publier leur thèse, leur article ou leur monographie. Quels rôles pourraient bien jouer les bibliothèques dans ce travail de recherche, par essence spécialisé et intimement lié à la discipline scientifique de chacun des chercheurs ?

Pourquoi accompagner les chercheurs dans l'exploitation des données ?

Le métier de *data librarian* est récent dans le paysage français des bibliothèques universitaires. S'il a commencé à émerger vers la moitié des années 2010[2], notamment avec la création de la liste de diffusion *data librarians* en 2017[3], qui compte à ce jour 134 abonnés, c'est en 2019 que les exigences des financeurs en matière de données de la recherche ont émergé et, avec elles, le besoin d'accompagnement des chercheurs.

En effet, le programme de financement européen Horizon 2020 lança, en 2014, l'Open Research Data Pilot[4], qui exigeait que les chercheurs rédigent un plan de gestion des données (ou *data management plan*). Ce pilote, qui ne concernait que certains projets, a été étendu à l'ensemble des projets financés par Horizon 2020 en 2017[5]. L'Agence nationale de la recherche (ANR) a emboîté le pas de l'Union européenne et a demandé à son tour, à partir de 2019, la fourniture d'un plan de gestion des données (PGD) pour tous les projets qu'elle finance[6].

---

[1] OCDE (2007), *Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics*, p. 18, https://www.oecd.org/fr/science/inno/38500823.pdf

[2] ALENGE B. (2015), « Entre *data librarians* et médiateurs du savoir », *Bertrand Calenge* : carnet de *notes*, https://bccn.wordpress.com/2015/02/06/entre-data-librarians-et-mediateurs-du-savoir/

[3] Accès à la liste : https://groupes.renater.fr/sympa/info/datalibrarians

[4] OPENAIRE (2017), *What is the open research data pilot*, https://www.openaire.eu/what-is-the-open-research-data-pilot

[5] EUROPEAN COMMISSION (2017), *H2020 Online Manual, Cross-cutting issues, Data management*, https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

[6] AGENCE NATIONALE DE LA RECHERCHE (2019), *L'ANR met en place un plan de gestion des données pour les projets financés dès 2019*, https://anr.fr/fr/actualites-de-lanr/details/news/lanr-met-en-place-un-plan-de-gestion-des-donnees-pour-les-projets-finances-des-2019/

L'exigence de la production de ce document peut paraître anodine. En réalité, 2019 est une année charnière, qui a rendu le métier de *data librarian* incontournable pour l'accompagnement des chercheurs dans leurs travaux. En effet, le PGD n'est pas un document administratif, mais bien un outil de pilotage, qui nécessite de se poser de nombreuses questions et surtout d'y répondre : quels types de données vais-je produire ? Nécessitent-elles de prendre des précautions particulières ? Comment vais-je les partager après l'aboutissement de mon projet ?

Il est évident que les chercheurs se sont posé ces questions bien avant 2019 et qu'ils n'ont pas attendu les injonctions des financeurs pour organiser leurs données et les partager, notamment dans certaines disciplines comme l'astronomie, qui est pionnière en matière de partage des données[7].

Mais la systématisation de ce document de planification a induit une prise de conscience parmi toutes les communautés scientifiques et a mis en lumière tant les besoins d'infrastructures techniques nécessaires à la bonne gestion de ces données (plateformes de stockage, entrepôts de données…) que les besoins en formation.

En effet, pour compléter un PGD, il est nécessaire de comprendre aussi bien les notions de métadonnées, de thésaurus ou de signalement que celles de données personnelles ou d'archivage pérenne ; autant de domaines qui peuvent sembler obscurs pour un public non averti.

Quelles compétences à acquérir pour les *data librarians* ?

Former et accompagner dans l'élaboration des plans de gestion des données

Le PGD se présente sous la forme de questions, qui interrogent le chercheur sur sa production et sa gestion des données tout au long de son projet. S'il n'est pas question pour les bibliothécaires de se substituer au chercheur, lequel est seul à même d'expliciter les données qu'il produit et leur pertinence au regard du projet scientifique qu'il mène, ils peuvent néanmoins lui apporter leur expertise sur la plupart des questions posées dans le document précité.

En effet, la structuration de l'information, sa description par des métadonnées et des thésaurus, son organisation selon des règles de nommage précises ou encore l'ouverture des données constituent le cœur de métier des bibliothécaires, en particulier dans un contexte de Science ouverte.

Le PGD constitue ainsi une formidable occasion d'amener le chercheur à prendre en considération toutes les bonnes pratiques en matière de gestion des données ; cette sensibilisation est d'autant plus importante pour le futur des PGD, que ceux-ci ont vocation à être lus à l'avenir par des machines et devront donc être élaborés en respectant des standards précis.

Le métier de *data librarian* peut ainsi être défini au prisme des services qu'il apporte aux chercheurs. Outre l'accompagnement dans la rédaction du PGD, les *data librarians* peuvent en effet intervenir à chaque étape du cycle de vie de la donnée, lequel peut être défini comme l'ensemble des phases de transformation de la donnée (laquelle est collectée, décrite, traitée, analysée…) tout au long d'un projet de recherche[8]. Au début d'un projet, il peut notamment être intéressant de chercher des jeux de données préexistants pour compléter l'état de l'art. Ce service peut être apparenté à de la recherche documentaire sur ce nouvel objet que sont les données.

En cours du projet, les *data librarians* peuvent apporter leur expertise pour la plupart des questions qui se posent en matière de gestion des données : choix des formats, organisation des fichiers, solutions de stockage, accompagnement juridique de premier niveau… À la fin du projet, les chercheurs peuvent aussi être aidés dans la sauvegarde et l'ouverture de leurs données par le biais d'un entrepôt adapté, à choisir parmi les milliers qui s'offrent à eux. Les *data stewards*, ou curateurs de données, qui, en général, ont un profil d'ingénieur, peuvent, quant à eux, intervenir sur les données elles-mêmes, disposant des compétences disciplinaires nécessaires.

Induire une montée en compétences progressive des chercheurs et des *data librarians*

En outre, les *data librarians* font généralement preuve d'une forte activité en matière de formation des chercheurs, ingénieurs, doctorants et professionnels de l'information aux enjeux et bonnes pratiques en matière de données, et ce dans le cadre d'ateliers, de rendez-vous individuels ou encore par la production de supports et de guides d'accompagnement, comme le propose, par exemple, l'Université de Lorraine dans sa boîte à outils Science ouverte[9].

Enfin, ce métier comporte une dimension Communication : le cadre réglementaire en matière de données de recherche et les exigences des financeurs évoluent rapidement, il est donc nécessaire de tenir les chercheurs régulièrement informés.

Si de nombreux services liés aux données, tels que la recherche de jeux de données, l'aide à la citation ou la description de données par des métadonnées, font partie des compétences traditionnelles des bibliothécaires intervenant dans le domaine de la Science ouverte, ce n'est pas forcément le cas de tous. Tous les *data librarians* doivent ainsi acquérir de nouvelles compétences.

[7] GENOVA F. (2014), « Du nécessaire partage des données scientifiques : l'exemple de l'astronomie », *Arabesques*, n°73, pp. 12-13, https://publications-prairial.fr/arabesques/index.php?id=999

[8] SCIENCE OUVERTE COUPERIN, *Données de la recherche, définitions*, https://scienceouverte.couperin.org/donnees-recherche-definitions/

[9] Université de Lorraine, *Boîte à outils de la Science ouverte*, https://scienceouverte.univ-lorraine.fr/boite-a-outils/

D'une part, la mise en œuvre du plan de gestion des données amène les bibliothécaires à travailler en étroite collaboration avec d'autres services que ceux de la bibliothèque. En effet, dans le domaine de l'informatique et du juridique, seul un accompagnement de premier niveau peut être apporté : le *data librarian* n'est ni un architecte des systèmes d'information ni un juriste. Aussi, c'est avant tout une excellente connaissance du paysage de la recherche et de ses acteurs qui est attendue du *data librarian*, afin, le cas échéant, de rediriger les demandes vers le bon service. C'est toute la philosophie des guichets uniques qui se mettent en place pour accompagner les chercheurs dans leur recherche de données.

D'autre part, l'objet « donnée » est plus complexe à appréhender que celui de la publication scientifique. Il peut en effet prendre des formes très variées, tant en termes de format que de discipline. Le *data librarian* ne peut être un expert de tous ces domaines scientifiques. Mais l'accompagnement qu'il apporte nécessite pour lui de comprendre, ne serait-ce que de manière superficielle, les données générées par les chercheurs. C'est pourquoi avant tout accompagnement, un entretien individuel est nécessaire pour permettre une vulgarisation scientifique de la part du chercheur et une précisions de ses apports méthodologiques de la part du *data librarian*. Ce moment d'échange est précieux pour forger une relation de confiance mutuelle entre ces deux acteurs et ainsi appréhender les données en mobilisant toutes les palettes de compétences scientifiques et documentaires nécessaires. La gestion des données relève donc bien d'un travail d'équipe.

Quelle est la structuration de l'accompagnement des chercheurs en France ?

Une communauté professionnelle française très développée

Ces nouvelles compétences à acquérir nécessitent la mise en place de formations professionnelles. L'ENSSIB[10], comme les réseaux des CRFCB[11] et les Urfist[12] ont rapidement investi ce terrain et proposent aujourd'hui de nombreuses formations sur ce sujet. Mais pour ce domaine en mutation, une veille professionnelle régulière est indispensable en complément de la formation continue. La communauté professionnelle française est très structurée et foisonnante sur ce point.

Le GTSO Données de Couperin – un groupe de travail composé de professionnels de l'information et de la documentation – propose de nombreux supports à destination des établissements et anime la communauté au travers de webinaires réguliers[13]. En 2021, ce groupe a publié les résultats de son « Enquête sur l'appui à la gestion des données de la recherche en service de documentation et d'information scientifique et technique »[14]. L'objectif de cette démarche était de comprendre la manière dont les services d'accompagnement s'étaient emparés des données de la recherche. Quelques conclusions principales peuvent en être tirées :

- en dépit des nouvelles exigences des financeurs, près de la moitié des répondants ne disposaient alors d'aucun service d'accompagnement ;
- pour les services déjà en place, la plupart étaient encore en phase de préfiguration ;
- les inégalités en matière d'accompagnement sont très fortes entre les petites et les grandes structures ;
- les services d'accompagnement sont davantage d'un niveau de sensibilisation que d'un niveau technique (préparation des données avant dépôt, description des données avec des métadonnées…) ;
- la gestion des données nécessite de travailler en coopération avec de nombreux services, y compris ceux externes à l'établissement.

Un besoin fort de soutien méthodologique a été exprimé par les 82 établissements ayant répondu à l'enquête.

Afin d'y répondre, le GTSO Données a commencé à produire des fiches pratiques thématiques, dont l'objectif est de répondre à une question précise que peuvent se poser les *data librarians* de toute la France : comment animer un atelier PGD ? Comment mettre en place un service d'accompagnement ? Comment former les doctorants ?

Pour animer la communauté, le GTSO Données propose également des webinaires[15], une lettre de veille[16] ou encore le partage de retours d'expérience. D'autres groupes de travail constitués autour des données sont très actifs en France, tels que RDA France, le groupe Données inter-réseaux du CNRS ou encore le Collège Données du Comité pour la Science ouverte. Tous ces groupes de travail sont complémentaires. Le métier de *data librarian* consiste donc également à se former et à s'informer en permanence, une démarche qui est facilitée par l'existence d'une communauté professionnelle nationale soudée ainsi que celle de réseaux internationaux dynamiques.

L'exemple de l'Université de Lorraine

L'accompagnement des chercheurs à la gestion des données de la recherche est une priorité fixée dans le deuxième plan national pour la Science ouverte : « pour accompagner et conseiller les chercheurs tout au long du cycle de vie des données, des "ateliers de la donnée" réunissant une large palette de métiers seront

[10] École nationale supérieure des sciences de l'information et des bibliothèques.

[11] Centre régional de formation aux carrières des bibliothèques.

[12] Unité régionale de formation à l'information scientifique et technique.

[13] COUPERIN, *GTSO Données*, https://www.couperin.org/science-ouverte/ressources-du-gtso/groupe-donnees

[14] FERET R., BRACCO L., LEHOUX E. & AUGOUVERNAIRE M. (2021), *Enquête sur l'appui à la gestion des données de la recherche en service de documentation et d'information scientifique et technique*, https://doi.org/10.5281/zenodo.5078504

[15] GTSO Données, chaîne Youtube, https://www.youtube.com/channel/UCswnouUXeMhHASeZNpYGhMA

[16] BRACCO L., *Quoi de neuf dans les données de la recherche ?,* https://www.getrevue.co/profile/laetitia-bracco

structurés sur tout le territoire »[17]. Une structuration nationale est donc en cours, par le biais d'appels à projets visant la labellisation en « ateliers de la donnée » de cellules existantes d'accompagnement à la gestion des données.

L'Université de Lorraine propose, depuis 2019, un accompagnement des chercheurs dans la gestion de leurs données. Initié par l'aide apportée pour élaborer le plan de gestion des données et la formation des doctorants, il s'est progressivement développé pour couvrir l'ensemble du cycle de vie de la donnée.

La palette de services proposés est large : outre les deux éléments précités, on y trouve le guichet unique de réponses, un accompagnement en matière de stockage, de dépôt dans des entrepôts, de respect du RGPD ou encore de publication d'un *data paper*[18]. Ces services ont vocation à se développer et à se massifier, mais sont déjà reconnus par le label national « Atelier de la donnée ».

Conclusion

Le métier de *data librarian* connaît un fort élan depuis 2019. L'accompagnement des chercheurs dans la gestion de leurs données a pris son essor avec l'exigence de la fourniture d'un PGD pour les projets financés notamment par l'Union européenne et l'ANR ; un essor qui dépasse ce cadre aujourd'hui. La gestion des données ne peut s'envisager qu'au travers d'une combinaison de compétences scientifiques, documentaires et informatiques, un cadre au sein duquel les *data librarians* ont un rôle majeur à jouer.

[17] MESRI (2021), *Deuxième plan national pour la Science ouverte*, p. 13, https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte/

[18] Voir le détail des services proposés sur la page dédiée : http://scienceouverte.univ-lorraine.fr/donnees-de-la-recherche/

Le Health Data Hub, levier pour la valorisation des données de santé

Par Stéphanie COMBES
Directrice du Health Data Hub (HDH)

En cherchant à mobiliser tout le potentiel des données de santé afin d'améliorer le système de santé, le Health Data Hub se situe à la ligne de crête de l'innovation.
Son objectif est de permettre et de simplifier l'accès aux bases de données de santé (principalement au Système national des données de santé (SNDS), la base de données de l'Assurance maladie) afin de les réutiliser à des fins de recherche.
Face aux nombreux enjeux soulevés, la plateforme a pris le parti d'inclure son écosystème et les citoyens dans son projet, ainsi que de s'ouvrir à une perspective internationale : pouvoir aligner le Health Data Hub avec l'ensemble des acteurs impliqués et les perspectives à venir.

La France est dotée de l'une des bases de données médico-administratives les plus volumineuses du monde

Cette base de gestion, liée aux remboursements des actes et des soins des bénéficiaires de l'ensemble des régimes d'assurance maladie obligatoire, a longtemps été sous-exploitée. À la suite de plusieurs rapports, l'appel à une meilleure accessibilité aux données de santé a été entendu en 2016. La loi Modernisation de notre système de santé a créé l'Institut national des données de santé (INDS), qui est chargé de favoriser le dialogue entre les acteurs et de simplifier les procédures d'accès à ces dernières.

Le 29 mars 2018, dans la continuité des recommandations du rapport du député Cédric Villani sur l'intelligence artificielle, le Président de la République annonçait la création d'un *hub* des données de santé, une structure partenariale dont l'objet est de garantir un accès simplifié aux données de santé au travers d'une plateforme technologique sécurisée disposant d'outils d'analyse répondant à l'état de l'art, le tout dans le respect des droits du citoyen. Le projet HDH devient, à ce titre, une des actions de la stratégie nationale en intelligence artificielle.

Pendant tout l'été 2018, à la demande de la ministre des Solidarités et de la Santé, Agnès Buzyn, une mission pilotée par trois experts du domaine des données de santé est conduite qui remettra son rapport le 12 octobre. La direction de la Recherche, des Études, de l'Évaluation et des Statistiques (DREES) se verra alors confier la mise en œuvre de la feuille de route qui y est proposée et, huit mois plus tard, la loi Organisation et transformation du système de santé (OTSS) sera votée. Son article 41 élargit le Système national des données de santé (SNDS) et crée officiellement le HDH.

Le Health Data Hub (HDH) est donc constitué en groupement d'intérêt public, dont la convention constitutive a été approuvée par arrêté ministériel du 29 novembre 2019. L'Assemblée générale du Health Data Hub réunit ses cinquante-six membres, répartis en neuf collèges : l'État, les caisses d'assurance maladie, les organismes d'assurance maladie complémentaires, les organismes de recherche, les établissements de santé, les professionnels de santé, les agences, les opérateurs et l'autorité publique indépendante, les représentants des usagers du système de santé et les industriels.

Le HDH vise donc à garantir un accès aisé et unifié, transparent et sécurisé aux données de santé pour améliorer la qualité des soins et l'accompagnement des patients

Chargé par la loi de réunir, organiser et mettre à disposition les données du SNDS élargi et de promouvoir l'innovation dans l'utilisation des données de santé, le HDH conduit quatre activités principales :

- il est un guichet unique pour accompagner les porteurs de projets dans leurs démarches administratives. Il assure, par ailleurs, le secrétariat du CESREES qui donne un avis sur les projets préalablement à leur transmission à la CNIL ;
- il met à disposition une plateforme sécurisée à l'état de l'art et offrant des capacités avancées de stockage, de calcul, de rapprochement et d'analyse des données ;

• il offre un accès à des bases de données partenaires, constituant le « catalogue de données ». Construit de manière progressive et itérative, ce catalogue est porté par un arrêté pris après avis de la CNIL. La réplication des bases mises au catalogue est posible pour en permettre l'accès et l'interfaçage : leur croisement avec d'autres bases de données, en accord avec les acteurs à l'origine de ces données. Ce catalogue est mis à jour régulièrement ;
• il anime l'écosystème dans le but d'accélérer l'innovation en favorisant le partage des expériences et des connaissances.

L'apport du HDH, service créé à destination des acteurs en charge de la collecte et de la mise à disposition des données et de ceux qui les utilisent, s'est concrétisé deux ans et demi après sa création par la mise en œuvre de :

• deux appels à projets avec plus de 350 répondants ;
• une plateforme technologique homologuée, sécurisée à l'état de l'art et actuellement utilisée ;
• soixante-six projets impliquant pour plus de la moitié des établissements de santé et ayant déjà abouti, pour certains, à de premières publications ;
• une dizaine de bases dans la première version du catalogue ;
• près d'une centaine de partenaires dont certains avec des associations d'usagers du système de santé ;
• et la création d'une direction citoyenne.

Le HDH bénéficie également d'une reconnaissance auprès des acteurs de l'écosystème et des acteurs européens et internationaux, comme la Commission européenne, Harvard ou le MIT.

Focus sur un projet innovant : INNERVE

Le projet INNERVE, porté par la société Quantmetry, a été sélectionné dans le cadre du deuxième appel à projets du HDH, en partenariat avec le Grand défi « Comment améliorer les diagnostics avec l'intelligence artificielle ? ». Il vise à développer un logiciel devant permettre d'affiner le diagnostic des neuropathies des petites fibres. Les petites fibres sont des fibres nerveuses qui permettent de ressentir la douleur et la température, Leur dégradation entraîne des douleurs et une perte de sensibilité. Identifier celle-ci demande un temps considérable au médecin qui doit explorer les images capturées lors de l'examen médical. Le recours à des algorithmes doit permettre, avec une efficacité similaire, de réduire considérablement le temps de détection et de libérer du temps médical. Aujourd'hui, le projet fournit des résultats de 70 % d'efficacité.

En parallèle de la création du HDH, le SNDS a été élargi à toutes les données de santé qui bénéficient d'un financement de la solidarité nationale dans le but de contribuer à une utilisation plus large de ces données. Initialement composé de certaines données médico-administratives telles que les feuilles de soin, la facturation hospitalière et les causes médicales des décès, le SNDS comprend donc désormais des données de registres, de cohortes de recherche, ou d'entrepôts de données hospitalières, etc.

Le HDH a été chargé par un décret publié en juin dernier de la coresponsabilité du traitement de la base principale avec la CNAM et de la responsabilité du traitement du catalogue du SNDS. Ces deux sous-ensembles du SNDS présentent un intérêt majeur pour l'écosystème.

Ainsi, la base principale désigne la réunion de données couvrant l'ensemble de la population : des données en provenance de l'Assurance maladie (base SNIIRAM), des établissements (base PMSI) ; des causes médicales de décès (base du CépiDC de l'Inserm), des données relatives au handicap (en provenance des MDPH – Données de la CNSA) en cible. Elle a vocation à être enrichie en continu avec l'intégration d'autres bases de données nationales et notamment, dans le cadre de ce premier arrêté, des bases relatives à l'épidémie de Covid-19 : vaccin Covid et SI-DEP (système d'information de dépistage).

Le catalogue désigne, quant à lui, une collection de bases de données non parfaitement connue à l'avance. Construit itérativement, il permet de s'adapter aux enjeux et besoins de l'écosystème. Son contenu est fixé par le Comité stratégique des données de santé, qui, créé en juillet dernier, est présidé par le ministre chargé de la Santé, tout comme est fixé l'ajout de nouveaux flux à la base principale. La première composition de ce catalogue a été fixée par un arrêté publié tout récemment.

Il est à noter que seule une copie des données sera transférée au HDH : l'acteur dépositaire de la source initiale la conserve et peut mettre à disposition ces données par ses propres moyens, s'il le souhaite. L'intérêt de copier ces bases dans la plateforme du HDH est de pouvoir enrichir ces bases en les croisant avec la base principale du SNDS et de réduire les délais d'accès en déléguant au HDH les actions d'hébergement, de *data management*, l'enrichissement du système par d'autres sources, la mise à disposition aux personnes habilitées de ces données dans des espaces informatiques maîtrisés, de contractualisation, etc. Les modalités sont précisées par voie de convention entre le dépositaire des données et le HDH.

Le catalogue et la plateforme du HDH permettent de mutualiser les investissements humains, technologiques et financiers inhérents au partage des données de santé dans le respect de la réglementation et des droits des personnes.

Les dix premières bases du catalogue constituent un premier aperçu de l'horizon des possibles en réunissant des bases de nature variée : des bases administratives, des bases nationales, des cohortes ou registres, des entrepôts parfois thématiques pour adresser des priorités de santé publique et également une base d'origine privée qui présente également un grand potentiel.

Des avancées significatives malgré la persistance de freins au partage des données

En dépit d'avancées, le HDH a rencontré différentes difficultés tout au long de sa mise en place, qui ont pu ralentir celle-ci, comme :

• les délais importants de publication des textes et donc des retards dans le transfert de la base principale du SNDS. En l'absence de celle-ci, le HDH met à disposition les données projet par projet, ce qui a pour effet d'augmenter considérablement les délais (environ un an en moyenne, pouvant parfois aller jusqu'à deux ans). Rendre effectif ce transfert en 2022 représente donc un enjeu majeur pour passer à l'échelle et accompagner un nombre croissant de projets ;
• la réticence à un partage large qui continue de guider ceux qui constituent les bases de données, en l'absence de modèles pérennes et lisibles de financement des bases de données de santé malgré leur intérêt aujourd'hui considéré unanimement comme incontestable ;
• l'absence d'un réseau mature d'entrepôts des données de santé dans les établissements de santé ou l'absence d'une base structurée sur les diagnostics médicaux en ville, ce qui privent les différents acteurs d'un accès à un patrimoine de données extrêmement important.

Les freins liés au partage et aux usages de la donnée de santé ont des conséquences pour notre pays. Les dépasser renforcerait notre recherche, augmenterait notre compétitivité, ainsi que notre attractivité et notre souveraineté, et ce dans un contexte de compétition internationale exacerbée. Les échanges entre l'écosystème hospitalier et le HDH se sont d'ailleurs accélérés ces derniers mois dans le but d'organiser une collaboration plus efficace afin de multiplier les projets d'envergure.

Une des pistes d'intérêt pour surmonter ces freins est celle de la dynamique européenne. À l'échelle européenne, la France est considérée comme étant l'un des pays les plus engagés dans le développement des usages des données de santé. La dimension internationale est au cœur de la philosophie du HDH, lequel est fortement impliqué dans la construction de plusieurs textes au niveau européen dans le domaine des données.

Le HDH est également engagé dans deux projets majeurs au niveau européen :

• en consortium avec des établissements de santé (les Hospices civils de Lyon, le Centre Léon Bérard Unicancer, l'hôpital Paris Saint-Joseph, le CHU de Nancy), l'ANS et la plateforme Bordeaux PharmacoEpi (BPE), il a été lauréat d'un appel à projets de l'Agence européenne du médicament visant à constituer une base clinique multicentrique ;
• le HDH est aussi le pilote d'un consortium composé de quinze partenaires européens stratégiques (huit plateformes nationales de données de santé, trois infrastructures de recherche, deux agences européennes (dont l'EMA) et des associations) afin de candidater à la préfiguration d'un espace européen des données de santé (EHDS). Cet EHDS constitue l'une des priorités de la politique de santé européenne ; il est considéré comme fondamental pour développer la recherche et améliorer la santé des citoyens. La réponse est attendue avant l'été.

La coopération européenne apparaît comme une opportunité à saisir pour exister au niveau mondial. En matière de santé, les pays de l'UE sont confrontés à des défis similaires à ceux de la France : transition épidémiologique, pandémies, augmentation des dépenses de santé…

Ces problèmes complexes font de la santé une question transfrontalière et ne seront résolus qu'en mettant à profit un partage des données à grande échelle considéré comme fondamental pour développer la recherche et améliorer nos systèmes de santé. Les collaborations européennes permettent d'atteindre un seuil critique, d'accroître l'avantage concurrentiel pour les entreprises et *start-ups* du secteur et de garantir une résilience et une visibilité accrues au niveau international. Elles peuvent générer des données de santé plus nombreuses et de meilleure qualité, réduire le risque de biais technologique, créer de nouvelles possibilités de croisements entre les données de santé et d'autres données, permettant de construire des algorithmes d'IA plus performants au service de la recherche ou de la performance industrielle. La dimension internationale de PariSanté Campus pourra renforcer la visibilité de la France au niveau international en réunissant ses forces vives derrière une identité unique et lisible (voir l'Encadré de la page suivante).

Conclusion

La France est aujourd'hui considérée comme un acteur légitime et pertinent au niveau européen. Elle dispose de perspectives pour s'imposer pleinement comme un leader naturel sur le sujet de la donnée de santé. Des progrès considérables ont été faits ces dernières années pour consolider la culture de la donnée auprès des acteurs de la santé. Des efforts, à porter avec PariSanté Campus, dont le HDH est membre fondateur, et le soutien de ses opérateurs et de son écosystème, sont encore nécessaires pour consolider un réseau d'acteurs partenaires sur l'ensemble du territoire capable de se mobiliser pour mener des projets collectifs, pour consolider le patrimoine de données de santé au travers de la mise en place d'entrepôts de données dans les établissements hospitaliers ou d'un observatoire de la médecine de ville interopérables avec le Health Data Hub y compris en termes de gouvernance et de valorisation, ou encore pour concevoir et mettre en œuvre une véritable politique publique de la donnée de santé permettant de définir un cadre de partage lisible et opposable au niveau national et tenant compte de la réglementation européenne.

Le HDH, le moteur d'un consortium candidat à la préfiguration de l'Espace européen des données de santé (EHDS)

Depuis 2019, la Commission européenne a identifié la constitution d'un espace européen des données de santé comme l'une des priorités de la politique de santé européenne. Il s'agit de faciliter l'accès aux différents types de données disponibles au sein des États membres dans le but de développer la recherche et d'améliorer la santé des citoyens.

La France est considérée comme étant l'un des pays les plus engagés dans le développement des usages des données de santé et dans de multiples projets. Aujourd'hui, le Health Data Hub, qui a inscrit parmi ses quatre axes stratégiques l'objectif de « positionner la France comme un leader dans l'usage des données de santé », est une structure-exemple dans la construction de plusieurs textes clés au niveau européen dans le domaine des données.

Il s'agit notamment :

• du projet de Data Governance Act (ou DGA), qui, adopté par la Commission européenne le 26 octobre 2020, vise à élaborer un cadre de gouvernance des données favorisant leur réutilisation et la création d'un véritable marché unique des données ;
• du projet de règlement sur l'espace européen des données de santé (EHDS) qui aura pour objet de construire un système de gouvernance solide, ainsi que des règles concernant l'échange des données de santé, mais aussi des infrastructures et une interopérabilité solide.

L'EHDS permettra également de répondre aux trois constats qui peuvent être faits à l'échelle européenne lorsqu'il s'agit de l'utilisation des données de santé : la fragmentation de ces données, la multiplicité de leurs conditions d'utilisation et la diversité des modèles de gouvernance permettant d'y accéder.
Le texte, dévoilé par la Commission européenne, le mardi 3 mai 2022, aura un impact majeur sur l'utilisation des données de santé en Europe, en ce qu'il précise les règles européennes en la matière. En ce qui concerne l'utilisation secondaire des données de santé, le projet de règlement met en place un cadre de gouvernance visant à simplifier et à harmoniser les conditions d'accès. Ce cadre de gouvernance comprend notamment des règles et une procédure d'accès aux données imposant des délais pour l'autorisation et la mise à disposition des données, la mise en place d'organismes nationaux chargés de l'accès aux données, la définition des rôles et des responsabilités des acteurs impliqués.
Par ailleurs, le projet de règlement propose également un cadre pour l'utilisation transfrontière des données de santé avec la mise en place d'une infrastructure européenne HealthData@EU, et guide les acteurs pour faciliter l'identification et la réutilisation des données au niveau européen. Enfin, le texte prévoit la mise en place d'une gouvernance européenne à travers la création d'un European Health Data Space Board.

En 2021, la Commission européenne a par ailleurs lancé un appel à projets dans le but de construire une première version test de cet EHDS. Dans cette optique, le HDH a constitué un consortium rassemblant les plateformes nationales de données de santé de plusieurs États membres (la Finlande, la Norvège, le Danemark, l'Allemagne, la Belgique, la Hongrie ou encore la Croatie), ainsi que de certaines agences européennes, comme l'Agence européenne du médicament (EMA) et le Centre européen de prévention et de contrôle des maladies (ECDC), et d'infrastructures de recherche (BBMRI, Elixir).

Financé à hauteur de 5 millions d'euros par l'Union européenne, ce projet permettra de construire et de tester une première version de l'espace européen des données de santé en interconnectant des plateformes de données, qu'il s'agisse de plateformes nationales, agences européennes et infrastructures de recherche, dans un réseau de nœuds. Le projet sera en mesure de tester un parcours utilisateur pour la création, le déploiement et la gestion de projets de recherche au niveau européens sur les données de santé. Pour ce faire, le consortium construira une infrastructure IT reliant l'ensemble des nœuds du réseau et définira des normes communes pour la sécurité, les catalogues de métadonnées, la qualité, l'interopérabilité des données et les exigences juridiques sur la base de cas d'usage proposés à la Commission européenne.

Ces cas d'usage concrets de recherche seront conduits par le consortium à l'échelle européenne, avec pour objectif de démontrer tout le potentiel de la réutilisation transnationale des données de santé pour la recherche, l'innovation, l'élaboration des politiques et de la réglementation. Ces projets de recherche pourraient couvrir des thématiques telles que la lutte contre le cancer, l'étude des maladies rares, l'évaluation de l'impact de la crise de la Covid-19, etc.

L'ensemble de ces travaux, prévus sur une durée de deux ans à partir de septembre 2022, serviront de matière à l'élaboration du règlement sur l'espace européen de données de santé actuellement à l'étude et préfigurent la mise en place pérenne de l'EHDS par la Commission européenne d'ici quelques années.

L'annonce des lauréats de l'appel à projets est attendue pour le début de l'été 2022.

Data science en santé et protection des données du Métavers

Par Adel MEBARKI
Co-fondateur et directeur général de Kap Code

L'usage des réseaux sociaux en santé connaît un essor permanent dans une société de plus en plus connectée. Ces plateformes sont devenues de véritables outils dans le parcours de vie des patients. De la recherche d'informations à la constitution de communautés, les réseaux sociaux s'imposent comme un élément permanent du « parcours de soins numérique » des patients et des parties prenantes en santé. Cette production spontanée de données de la vie réelle fait l'objet d'une multitude de recherches à des fins de santé publique. Mais à l'ère annoncée du Métavers, plusieurs questionnements éthiques et sociétaux se posent autour de l'exploitation de ces données sensibles.

Introduction

Le Web communautaire connaît un essor considérable dans la vie sociale des citoyens. Avec une utilisation moyenne de l'ordre de 2 h 30[1] par jour, force est de constater que ces outils numériques ont drastiquement changé le paradigme social de nos sociétés (voir la Figure 1).

Cette tendance, observée depuis plus d'une décennie, a connu une accélération fulgurante avec la pandémie de Covid-19. Selon une étude d'Harris Interactive[2], 40 % des internautes français déclarent avoir créé un compte sur un réseau social ou téléchargé une application de messagerie instantanée pendant le confinement.

L'usage des réseaux sociaux en matière de santé ne déroge pas à la règle ; ces plateformes sont devenues de véritables outils dans la gestion du parcours de vie des patients. De la recherche d'informations à la constitution de communautés, les réseaux sociaux s'imposent comme un élément permanent du « parcours de soins numérique » des patients et des parties prenantes en santé.

Les patients s'organisent en communautés avec pour objectifs de partager leurs retours d'expérience entre pairs et de mettre en commun leurs connaissances.

(1) https://wearesocial.com/fr/blog/2022/01/digital-2022-une-nouvelle-annee-de-croissance-exceptionnelle/

(2) https://www2.harris-interactive.com/social-life-2020

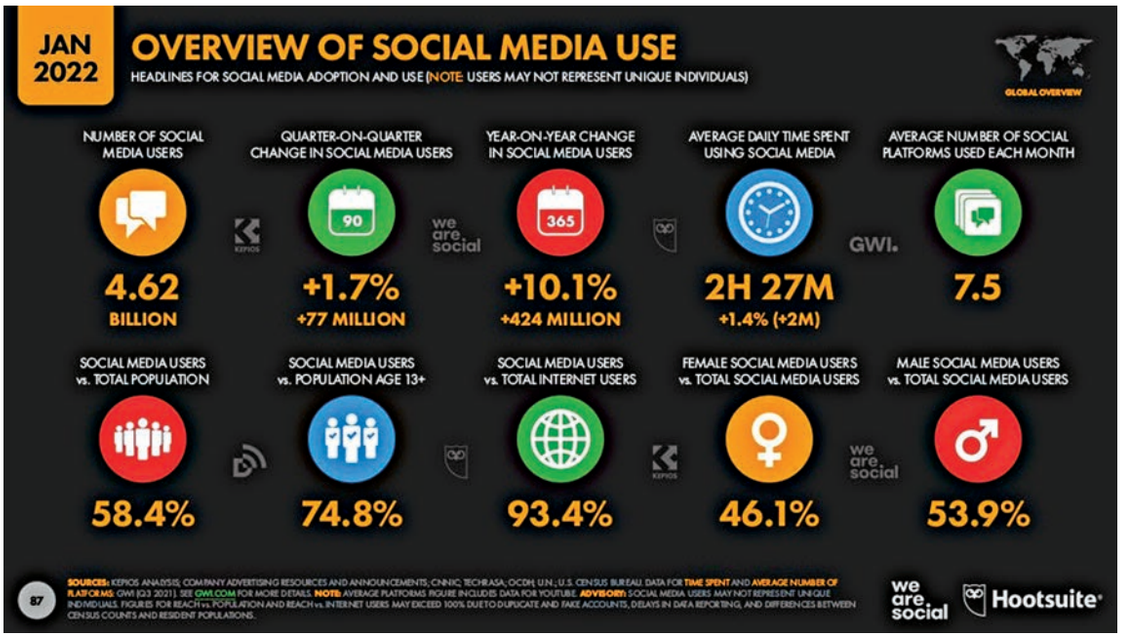


Figure 1 : Aperçu chiffré de l'utilisation des réseaux sociaux – Source : We Are Social.

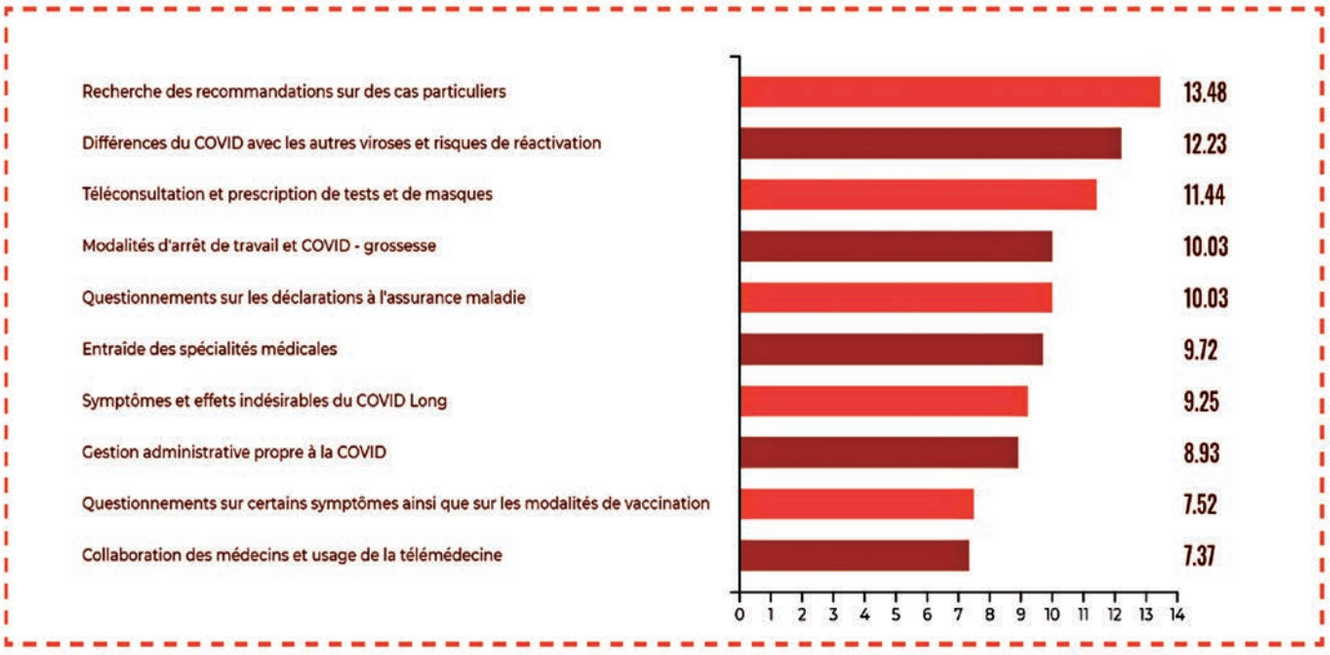


Figure 2 : Usage fait de Twitter par les professionnels de santé durant la pandémie de Covid-19 – Source : Kap Code.

De leur côté, les professionnels de santé se questionnent et s'entraident également, par exemple, *via* le hashtag #DocTocToc sur la plateforme Twitter[3]. Ces interactions virtuelles multiplient les potentiels usages de ces réseaux par les patients, les aidants et les professionnels de santé (voir la Figure 2 ci-dessus).

L'usage fait des réseaux sociaux en matière de santé

En santé, les réseaux sociaux ont deux principaux usages du point de vue du système de soins : la communication ciblée et la recherche observationnelle.

Ces plateformes réunissent un nombre important de patients et des outils performants, ce qui permet un ciblage plus fin des populations par les professionnels de santé. Ainsi, les actions d'information et de prévention font l'objet de campagnes de mieux en mieux ciblées[4]. Ce ciblage permet notamment d'améliorer les recrutements de volontaires dans le cadre d'études observationnelles ou d'essais cliniques. À titre d'exemple, la constitution de la cohorte d'étudiants de l'étude i-Share pour suivre la santé mentale de plus de 30 000 étudiants en France s'est appuyée sur les réseaux sociaux pour leur recrutement et le maintien de l'intérêt des participants tout au long de l'étude[5].

L'avènement de l'infodémiologie

L'usage quotidien de ces plateformes participe activement à la mise en place de ce que le Dr Guy Faggherazzi présente comme le « Digitosome »[6]. Ce concept désigne l'ensemble des données générées en ligne par un individu tout au long de sa vie, que ce soit *via* son *smartphone*, ses objets et dispositifs médicaux connectés ou encore son activité sur les réseaux sociaux. Ainsi, le partage spontané par les patients de leurs données et la disponibilité de celles-ci permettent l'émergence d'une nouvelle discipline de recherche : l'infodémiologie.

Au début des années 2010, le Dr Gunther Eysenbach a défini cette discipline de recherche comme celle construite autour de l'usage des données issues d'Internet dans un objectif de santé publique[7]. Ce terme d'infodémiologie, né de la contraction entre les informations issues de la navigation Web et épidémiologie, désigne la capacité à appréhender les déterminants de santé des populations, grâce aux données qu'elles génèrent en ligne, dans le but de participer à la mise en place d'actions de santé publique.

Cette nouvelle discipline vise notamment à utiliser le « digitosome » des parties prenantes de la santé afin de modéliser les parcours de soins, d'évaluer la qualité de

(3) https://www.kapcode.fr/articles/doctoctoc-covid19-twitter/

(4) https://www.santepubliquefrance.fr/docs/reseaux-sociaux-et-community-management-la-prevention-a-la-rencontre-des-publics

(5) https://presse.inserm.fr/la-sante-de-30-000-etudiants-suivis-pendant-10-ans/7640/

(6) FAGHERAZZI G. *et al*. (2018), « Étude mondiale de la détresse liée au diabète : le potentiel du réseau social Twitter pour la recherche médicale », *Revue d'Épidémiologie et de Santé publique,* 66 :S197-S198.

(7) EYSENBACH G. *et al*. (2009), "Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search", Communication and Publication Behavior on the Internet, *Review J. Med. Internet Res.* 11(1):e11.

vie des patients ou de surveiller les signaux faibles de pharmacovigilance. L'infodémiologie, qui était encore méconnue il y a quelques années, connaît aujourd'hui un boom scientifique. Une équipe de recherche de l'Université du Maryland a dénombré plus de dix usages différents des réseaux sociaux et des données qui en sont issues dans le but de participer à l'amélioration de la santé publique[8].

La disponibilité des données ainsi que l'amélioration des outils de *machine learning* et de traitement automatique du langage permettent le développement de ces nouveaux cas d'usage (voir la Figure 3 ci-après).

Quid de l'éthique et de la protection des données personnelles ?

Cependant, l'évolution sociologique que nous connaissons à travers l'usage quotidien des outils numériques engendre une dimension préoccupante. Déjà présent depuis plus d'une décennie, le phénomène d'infodémie vient rebattre les cartes de l'information de santé sur les réseaux sociaux. Véritable raz de marée qui s'est développé tout au long de la pandémie de Covid-19, les contenus de désinformation ont eu une influence négative non négligeable sur l'acceptabilité des politiques sanitaires partout dans le monde et ont amplifié le phénomène de défiance institutionnelle. Cela pose de véritables questions quant à la pertinence des contenus d'information en santé disponibles sur Internet et quant aux moyens dont disposent les pouvoirs publics pour informer convenablement leurs populations.

L'OMS a notamment initié des travaux portant sur la lutte contre la désinformation en collaborant avec les différentes plateformes afin de mettre en place des fils d'actualités « validés »[9]. Qualifié de menace majeure de la prochaine décennie, le phénomène de la désinformation fait écho à de nombreux questionnements éthiques et réglementaires autour des réseaux sociaux et de leur usage au quotidien.

De même, la disponibilité et les conditions d'accès aux données issues des réseaux sociaux posent question. Le récit du scandale Cambridge Analytica dénoncé en 2018[10] a mis la lumière sur les manquements des politiques des plateformes en matière de mise à disposition des données. Même si, depuis, d'importants changements ont été apportés pour protéger au mieux les citoyens, un certain nombre de ces questions restent en suspens.

Comment peut-on à la fois accepter de partager autant de données avec des opérateurs tels que Facebook ou Twitter, et permettre, en même temps, aux citoyens de protéger leurs données personnelles ? Comment peut-on réguler les finalités et le cadre méthodologique du traitement de ces données ? Comment peut-on œuvrer pour faire en sorte qu'un consentement individuel direct soit possible ?

À l'ère du Métavers

Encore au stade conceptuel, le Métavers est la future évolution du numérique. Il désigne la prochaine évolution d'Internet, qui reposera sur des espaces virtuels, persistants et partagés accessibles *via* une interaction en 3D. Cette vision, si elle se réalisait, changerait la réalité sociale et médicale de demain.

[8] CHEN J. & WANG Y. (2021), "Social Media Use for Health Purposes: Systematic", *Review J. Med. Internet Res.* 23(5):e17917.

[9] https://www.who.int/fr/news-room/spotlight/let-s-flatten-the-infodemic-curve

[10] https://siecledigital.fr/2020/10/08/cambridge-analytica-une-enquete-prouve-que-les-donnees-tirees-de-facebook-ont-permis-de-cibler-les-electeurs/

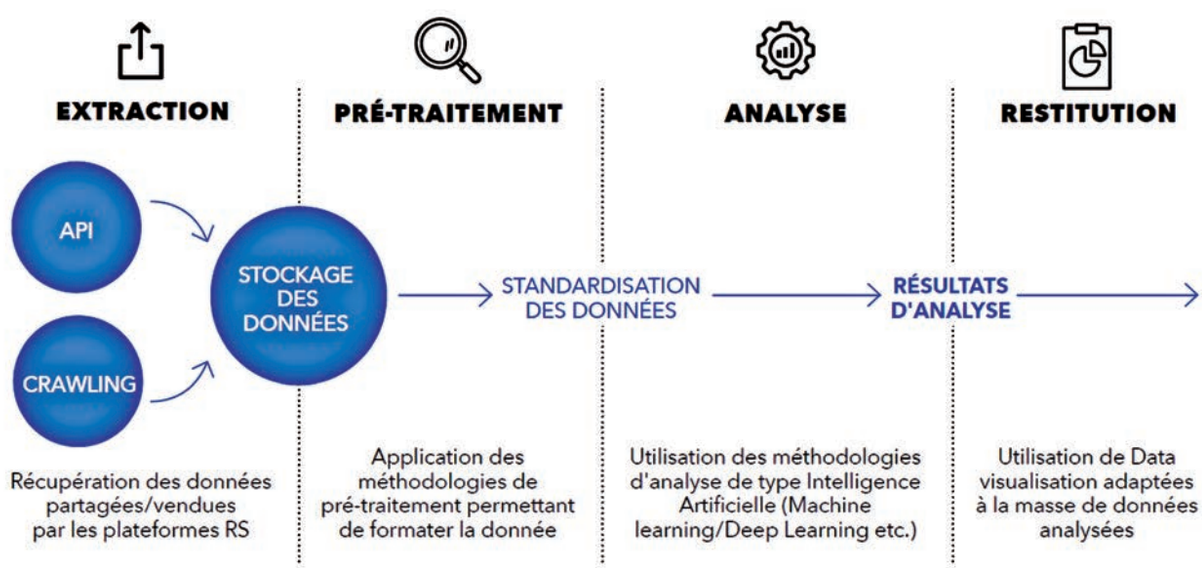


Figure 3 : Schéma d'analyse des traitements des données issues des réseaux sociaux – Source : Heathcare Data Institute.

Il sera notamment possible, dans un monde virtuel permanent, d'avoir accès aux soins et aux expertises médicales en faisant abstraction des contraintes physiques.

Encore au stade embryonnaire, cette vision viendrait décupler considérablement le digitosome des patients et ouvrirait ainsi la voie à une nouvelle forme de médecine personnalisée. La création (ou plutôt la multiplication) de biomarqueurs digitaux viendra alimenter des programmes de pré-diagnostic, de suivi médical, de dépistage ou encore de prédiction de la survenue d'une épidémie. Autant de données qui pourront alimenter les enjeux de santé publique de demain.

Toutefois, cela amènerait aussi potentiellement une centralisation plus importante de la donnée dans les mains d'acteurs déjà largement établis. La dynamique de concentration des données sensibles au sein des GAFAM que l'on connaît aujourd'hui, si elle se poursuit, les mènera à une hégémonie de plus en plus grande. De même, la multiplication de ces données accroîtra considérablement les risques de (ré)identification des personnes, tant elles vont permettre d'isoler des individus uniques au sein de la masse des internautes.

Conclusion

Les évolutions technologiques et la multiplication des données issues des réseaux sociaux offrent de véritables opportunités en matière d'amélioration des politiques de santé publique. Elles s'accompagnent aussi de multiples questionnements quant à la viabilité du modèle numérique de demain.

Le succès du Métavers et son impact positif en matière de santé ne se concrétiseront qu'avec la confiance que lui accorderont les citoyens et le respect par ses acteurs (GAFAM et acteurs publics) des libertés des premiers. Pour cela, une réflexion ambitieuse doit être menée autour de trois axes prioritaires :

- l'éthique : comment réguler et identifier les usages éthiques des réseaux sociaux ? À l'instar des études observationnelles, les comités de protection des personnes peuvent avoir un rôle important à jouer au regard de la validation du caractère éthique des projets de recherche ;
- la méthodologie : comment adopter des méthodologies de référence permettant une validation des résultats de ces recherches ? Au même titre que les essais cliniques, l'infodémiologie a un fort besoin d'un cadre méthodologique pour améliorer son acceptabilité ;
- le réglementaire : comment conjuguer recherche populationnelle et protection des données personnelles ? Au même titre que le Règlement général de la protection des données personnelles a été une démarche ambitieuse de régulation au niveau européen, nous aurons à entamer une réflexion similaire vis-à-vis de la mutation d'Internet et des enjeux de santé de demain.

Les enjeux du Métavers en matière de protection des données personnelles

Par Thomas FAURÉ
Président fondateur de Whaller[1]

Le Métavers annoncé par Mark Zuckerberg sera un univers allant bien au-delà de celui que nous connaissons aujourd'hui ; un univers dans lequel « vous serez capable de faire presque tout ce que vous pouvez imaginer : travailler, apprendre, jouer, acheter, créer (…) » Coupé du monde physique, le « Métanaute » sera totalement immergé dans un nouveau monde numérique. Sans aucune limite, le Métavers aura des conséquences anthropologiques majeures. En parallèle, les données personnelles produites vont être exponentielles ainsi que les difficultés à les protéger. Tout ce que fera l'utilisateur dans ce nouveau monde pourra être exploité pour ou contre lui. Des « données augmentées », biométriques et comportementales, permettront de proposer des publicités encore plus ciblées qu'elles le sont aujourd'hui. Déjà les GAFAM déposent des brevets pour la collecte desdites données et leur exploitation potentielle. Il convient donc, face à ce nouveau monde annoncé où la réglementation actuelle apparaîtra vite dépassée, de réfléchir sans attendre à la future protection des droits fondamentaux de chaque personne.

Le Métavers arrive ; ce nouvel espace virtuel mis en scène, il y a quelques années, par le réalisateur Steven Spielberg dans son film *Ready Player One.* C'est par la voix de son président fondateur, Mark Zuckerberg, que Meta – ex-Facebook – a été la première plateforme à annoncer son engagement dans ce qui sera une nouvelle ère du numérique. Puis, Microsoft, avec Mesh, lui a emboîté le pas, ainsi que de multiples autres acteurs de la *Tech*. Il va de soi que le Métavers, ou le « jumeau numérique de la Terre », produira une quantité astronomique de données. Aussi, après avoir défini le cadre de ce que l'on entend par Métavers, nous tenterons d'identifier les enjeux que ce nouveau monde annoncé comme dépassant l'actuel soulèvera en matière de protection des données personnelles[2].

Le Métavers, un mélange du réel et du virtuel

Historiquement, des métavers existent depuis une trentaine d'années sous la forme d'univers virtuels. Ce sont les univers virtuels des jeux vidéo en trois dimensions (Second Life, Minecraft, etc.). Pourtant, le Métavers annoncé par Mark Zuckerberg franchit une nouvelle étape. Ce changement radical, c'est celui de l'interface homme-machine. Dans un monde où l'écran s'est imposé partout comme l'interface la plus aboutie (surtout depuis qu'il est tactile, mobile et alimenté en longue durée), Zuckerberg veut purement et simplement le supprimer en le remplaçant par une interface totalement immersive : un casque de réalité virtuelle, et pourquoi pas même, une combinaison couvrant entièrement le « Métanaute » et lui permettant de communiquer ou d'entrer dans ce nouveau monde entièrement numérique, mais qui n'en sera pas moins réel.

Le « Métanaute » sera immergé. Ses yeux et même ses autres sens seront coupés du monde physique pour retrouver toute leur plénitude dans ce nouveau monde numérique.

Il me paraît important de nous arrêter quelques instants sur cette révolution, car si nous sommes incapables de comprendre ce qu'elle impliquerait, il nous sera alors impossible de réfléchir plus en avant aux conséquences multiples d'un monde métaversé.

[1] Whaller est une entreprise française proposant des solutions permettant de mettre en place un réseau social d'entreprise, un campus numérique ou un réseau associatif compatible avec plusieurs suites logicielles, telles que Microsoft 365, Google Workspace ou Zoom. Whaller s'est allié avec OVHcloud pour « proposer une solution 100 % souveraine ». Whaller redonne à l'utilisateur le pouvoir et le contrôle sur ses données confidentielles.

[2] Une donnée personnelle correspond à tout ce qui est relié à une personne physique et qui permet, directement ou indirectement, de l'identifier, par recoupement ou par inférence. Ce statut ne change pas avec le contexte.

Cette rupture technologique entraînera des conséquences anthropologiques majeures. Aujourd'hui, le vecteur numérique – « ce qui passe par » – est accessible et partiellement inclus au sein de supports physiques tangibles. Ce sont les terminaux dont les écrans font partie, mais qui sont en réalité composés de beaucoup de briques technologiques issues de cette rapide évolution industrielle que nous vivons depuis cinquante ans : les batteries, les circuits imprimés, la mémoire vive, etc.

Et c'est parce que le numérique est actuellement accessible *via* le monde physique que l'on sait le distinguer dudit monde. Or, nous observons déjà parfois de grandes confusions entre ces deux dimensions, et même une interpénétration du vecteur numérique avec le monde physique.

Ce que l'on veut réaliser à travers le Métavers, c'est une coupure entre les mondes physique et numérique. Comme l'interface physique ne peut pas aujourd'hui disparaître (on ne peut pas se passer des écrans, des haut-parleurs, des capteurs pour interagir avec la technologie numérique), l'idée est alors de faire en sorte que l'interface englobe totalement l'humain pour l'immerger dans un monde entièrement interfacé, et non plus seulement partiellement comme c'est le cas aujourd'hui.

Dans le Métavers, ce mélange entre monde physique et monde numérique disparaîtra. Le Métanaute sera immergé dans le monde virtuel au moyen d'une combinaison sensitive qui reproduira les sensations qu'il éprouve au travers de son corps et de ses sens, notamment celui du touché.

Grâce aux développements techniques, à la réalité virtuelle, Mark Zuckerberg a présenté le Métavers comme un univers dans lequel « vous serez capable de faire presque tout ce que vous pouvez imaginer : travailler, apprendre, jouer, acheter, créer (…). Vous serez capables de vous téléporter instantanément, sous forme d'hologramme, à votre bureau, et donc sans vous déplacer. Vous pourrez assister à un concert avec des amis, ou faire irruption dans le salon de vos parents pour prendre de leurs nouvelles. »

Le Métavers est un méta-univers, c'est-à-dire un univers qui va bien au-delà de celui que nous connaissons actuellement. C'est une véritable mise en abîme ; pour pénétrer dans ce nouveau monde, il faudra se munir d'un casque de réalité virtuelle. Notons qu'il pourra coexister plusieurs métavers. Parmi ceux-ci, il y aura des métavers décentralisés basés sur la *blockchain* (c'est à ce niveau que l'on parle de Web 3.0) et des métavers centralisés, comme celui de Meta/Facebook, s'appuyant sur le *cloud* – qui à mon sens constitueront le véritable Web 3.0, qui sera un dévoiement de l'idée initiale de celui-ci, comme l'a été le Web 2.0 qui se voulait être un Web de partage, de collaboration, mais qui, finalement, est devenu un Web hégémonique, un ogre se nourissant de nos données personnelles.

L'idée finale du Métavers est celle de la recréation d'un monde par les humains qui étendra les capacités du monde réel en partant de ce dernier. Permettant de se projeter où l'on veut, en compagnie de qui l'on souhaite, on peut alors imaginer que les humains pourront faire tout ce qu'ils font déjà dans la vie actuelle, et plus encore.

Par ailleurs, notons qu'il n'y a pas actuellement de barrières techniques qui empêchent le développement du Métavers. Le principe de la réalité augmentée peut se concrétiser au travers de briques ou de couches de technologies qui sont parfaitement maîtrisées aujourd'hui. Elles demandent « simplement » des développements qui feront de ces briques une technologie en elle-même.

Enfin, à l'aune de l'expérience que nous avons des comportements actuels vis-à-vis du numérique et des réseaux sociaux en particulier, on peut imaginer que nombre d'internautes passeront beaucoup de temps dans le Métavers, et même que certains ne le quitteront pas. Dans tous les cas, lorsque l'individu reposera son casque, il vivra alors une dissociation bien plus importante entre les deux mondes, le virtuel et le physique ; une déconnection avec la réalité qui sera une source de troubles encore plus graves que ceux déjà ressentis par beaucoup d'entre nous et que l'on associe aujourd'hui maladroitement à l'addiction. Alors qu'il s'agit probablement davantage d'une préemption pure et simple de l'entièreté de ce qui constitue l'être humain.

Des données personnelles augmentées

Au sujet de la protection des données personnelles, comme nous le constatons actuellement, nous n'arrivons pas, ou avec difficulté, à réguler leur protection sur Internet, et ce qu'elles se présentent sous forme de textes ou d'images. Comment, demain, assurer la protection des droits fondamentaux des utilisateurs quand, notamment, les GAFAM pourront collecter librement leurs données et surveiller sans limites les consommateurs ?

Après les nombreux scandales (Cambridge Analytica, WhatsApp, etc.), Meta/Facebook a annoncé vouloir être exemplaire en respectant la vie privée et en offrant aux utilisateurs un contrôle de leurs données. Néanmoins, cela paraît d'ores et déjà insuffisant, car les données produites vont croître de manière exponentielle, ainsi que leur collecte et leur exploitation potentielle. Dans un environnement où nous utiliserons des interfaces venant de toutes parts pour pénétrer dans les métavers et déterminer la façon dont nous vivrons dans ces derniers, la protection des données deviendra encore plus critique. Qu'en sera-t-il de leur confidentialité réelle ?

À l'heure où j'écris ces lignes, une *start-up* suédoise, BehavioSec, vient d'être rachetée par LexisNexis Risk Solutions, le spécialiste de la lutte contre la fraude[3]. Cette *start-up* édite une solution qui est capable de vérifier l'identité d'une personne en

[3] https://www.usine-digitale.fr/article/lexisnexis-risk-solutions-s-empare-de-behaviosec-specialiste-de-la-biometrie-comportementale.N2001912

analysant son comportement sur son *smartphone*. C'est ce que l'on appelle la biométrie comportementale.

On sait donc identifier formellement une personne uniquement en analysant sa gestuelle sur son téléphone. Les réseaux sociaux, et Facebook en particulier, sont déjà passés maîtres en la matière et savent deviner ce qui retiendra l'utilisateur en analysant en permanence non pas seulement les données personnelles que chacun dépose volontairement (les *likes*, les données de profil, etc.), mais aussi, et surtout, les données secondaires qui sont encore plus importantes en termes de volume : le temps passé à lire un message, les mouvements de la souris, du doigt, la vitesse de défilement, l'attention portée à telle notification ou une autre, etc. Cette exploitation de données personnelles est déjà vertigineuse, que peut-on imaginer de plus dans le Métavers ? Et bien tout. Tous les comportements, le regard, les expressions faciales, la tension artérielle, le rythme cardiaque, les agissements, mais aussi les hésitations, tout est donnée personnelle. Tout ce que fait l'utilisateur pourra être exploité pour…, ou plutôt contre lui.

Issues du comportement physique des utilisateurs, ces nouvelles données personnelles seront captées *via* les casques virtuels ou les combinaisons immersives. Ainsi, les informations disponibles sur chaque individu seront décuplées, englobant les réactions, les émotions et les sensations, qui servent le principe même de la technologie du Métavers, celui de l'interaction avec l'utilisateur.

On parle de « données augmentées » du fait même de leur traitement. Ces données biométriques et comportementales permettront de proposer des publicités encore plus ciblées qu'elles ne le sont aujourd'hui. Meta/Facebook a déjà déposé des brevets en ce sens.

Par ailleurs, et comme c'est déjà le cas pour le jeu vidéo, il existera un marché des avatars avec lesquels les consommateurs « vivront » dans les métavers. Or, à ces avatars seront associées toutes les données historiques. Quand sera-t-il demain avec la création d'avatars (de clones) de personnes auxquelles sont attachées des données biométriques et comportementales ? Une personne gardera-t-elle le droit fondamental à la protection de ses données personnelles, comme c'est le cas en Europe, ou bien, est-ce la philosophie américaine qui s'appliquera, celle du droit de vendre ces données ? Nul ne le sait encore. Mais l'on peut penser que ceux qui créent ces technologies en demeureront les maîtres conformément à l'adage "code is law".

Une autre question surgit à propos des métavers. On peut penser qu'ils relèvent de la politique fiction, mais nous n'en sommes pas certains. Les entreprises privées et autres organisations pourront créer leurs propres métavers, avec leurs règles, leurs frontières, etc. Bref : leurs propres « néo-États ». Et chacun voudra alors y faire régner son propre ordre juridique, à l'intérieur du monde qu'il aura lui-même créé.

Qu'adviendra-t-il alors de l'applicabilité des lois terrestres, et donc de la protection des données ? Avant de répondre à cette question, nous allons regarder l'état actuel du droit.

Étendre le champ d'application du RGPD, du DMA et du DSA

Les différents règlements adoptés par l'Union européenne (RGPD, DMA, DSA) continueront de s'appliquer dans le cadre des métavers proposés à des personnes résidant au sein de l'UE ou voulant interagir avec celles-ci. Pourquoi ? Parce que c'est la philosophie qui s'est imposée au cours des dernières années. Portée par le commissaire européen au Marché intérieur, Thierry Breton, ces textes peuvent se résumer par cet adage : « Tout ce qui est interdit ***offline*** doit l'être *online* ». Ainsi, conséquence importante dans les faits, le RGPD aura vocation à s'appliquer aux métavers qui, quoique virtuels et sans frontières, auront cours dans l'UE. Les éditeurs, les hébergeurs et les prestataires participant au développement des métavers devront se plier aux réglementations européennes s'ils veulent que ces mondes virtuels puissent être utilisés par les Européens.

Néanmoins, compte tenu des volumes faramineux des données personnelles et surtout de cette nouvelle typologie de captation de ces données que les métavers impliquent, la question juridique est plus que jamais à l'ordre du jour.

Il y a fort à parier que le RGPD ne suffira plus. Il sera donc nécessaire de mettre en place une nouvelle législation. Remarquons que l'UE, par son action en matière d'harmonisation législative et sa puissance économique – elle est le premier marché de consommateurs/internautes au monde –, disposera d'un pouvoir fort face aux enjeux de cette régulation qui s'annonce.

Mais avant cela, et dans le but de protéger les données personnelles des utilisateurs, il conviendra de se reposer cette question : qu'est-ce que les opérateurs du Métavers capteront comme données ? Car c'est bien une des difficultés auxquelles nous sommes actuellement confrontés avec les réseaux sociaux, celle de savoir quelles données sont collectées et enregistrées.

Connaître quelles métadonnées seront captées, et sous quel consentement, permettrait déjà d'éclairer de vastes zones sombres et de constituer la base de la construction d'un futur encadrement des métavers.

Continuer à s'opposer à l'hégémonie des GAFAM

Nouvel univers totalement immersif et séduisant, le Métavers est aussi une source d'inquiétude. Car il ne faut pas être naïf, les GAFAM profiteront de cette technologie, comme ils le font déjà, pour étendre leur hégémonie *via* la collecte d'informations personnelles.

Il faudra également lutter pour les contraindre dans leurs agissements. Or, cela peut apparaître comme contre-intuitif, voire attentatoire aux libertés, comme cela a été le cas avec le Web 2.0. Cela pourrait commencer par interdire la commercialisation de casques pour les enfants et les adolescents, afin de les protéger.

Il convient donc d'organiser, et ce quelle que sera la finalité des métavers, leur régulation afin de protéger les droits des personnes, et, en particulier, ceux s'attachant à leurs données personnelles comme nous l'avons vu *supra*. Il faudra même aller plus loin. Étant, avec la réalité augmentée, au début d'une nouvelle ère de l'humanité, il faudra élaborer de nouveaux droits de l'homme tenant compte du fait que celui-ci vit désormais dans une « humanité numérique » en pleine construction. Peut-être, sera-t-il opportun de se poser la question d'adopter une régulation à l'échelle planétaire.

Il n'est pas interdit de penser que la seule « bonne solution » soit que l'Europe sache faire naître en son sein des métavers qui lui soient propres et donc plus faciles à réguler à son niveau. Car une vie meilleure ne s'inscrit pas forcément dans les modèles de métavers que nous proposeront les sociétés de la Silicon Valley, des modèles qui restent étrangers à notre culture européenne et à nos mœurs, mais que, faute d'alternative et, aussi, un peu par faiblesse, nous, Européens, adoptons. C'est avant tout, comme tous les sujets de souveraineté technologique, une question de courage et de volonté politiques.

A decade and a half of OECD action on data governance policy-making[1]

By Elettra RONCHI
Senior Policy Consultant on Data Governance and Digital Health WHO/Europe;
former Head of the Data Governance and Privacy Unit in the Division for Digital Economy Policy at the OECD

And Christian REIMSBACH-KOUNATZE
Information Economist and Policy Analyst in the Division for Digital Economy Policy at the OECD

The OECD has long recognized the need to better understand how to reconcile the risks and benefits of data access and sharing to help governments reap the benefits of data-driven innovation. To guide policy-making, the OECD has produced over the last decade and a half a significant body of analytical work and legal instruments setting out principles and best practices to address sector-or domain-specific challenges in the governance of data. These Recommendations include: the Recommendation concerning Access to Research Data from Public Funding[2]; the Recommendation for Enhanced Access and More Effective Use of Public Sector Information[3]; and the Recommendation on Health Data Governance[4]. In what appears to be the latest strong demonstration of its commitment to the issue, the OECD Council adopted in 2021, the Recommendation on Enhancing Access to and Sharing of Data (EASD Recommendation)[5]. Differently from the preceding ones, the EASD Recommendation provides an overarching set of principles and policy guidance to help governments reconcile potential risks and benefits and unlock the re-use of all types of data across and within sectors, jurisdictions, organisations, and communities. The aim of this paper is to put in context this significant body of work and set out the main policy issues addressed by these OECD Recommendations.

Why Enhance Access and Sharing of Data?

In a world where the effective use of data can help boost productivity and improve or foster new products, processes, services and markets, the ability to access and share data, regardless of geography, has become crucial for securing economic growth and prosperity. Overall, the OECD has estimated that data sharing and re-use can "generate social and economic benefits worth between 0.1% and 1.5% of gross domestic product (GDP), in the case of public-sector data, and between 1% and 2.5% of GDP when also including private-sector data" (OECD, 2019a). Enhancing the sharing of data can also contribute towards societal objectives, help advance the sustainable development goals agenda, transparency and accountability of governments and further democracy.

At the same time there are risks associated with data access and sharing. Data breaches, may violate the privacy of individuals when their personal data is involved and harm the commercial and noncommercial interests of organisations (e.g., through the infringement of intellectual property rights). These risks may make individuals, businesses, and governments reluctant to share data. For example, some individuals may object to their data being re-used due to confidentiality concerns, even if they are aware of the social benefits that such re-use could deliver.

The Covid-19 pandemic has underscored the importance of data sharing. It has also highlighted how trust or the lack thereof can play a significant role in people's willingness to provide their data especially when that data involves their health and other sensitive information as their whereabouts (OECD, 2020a; OECD, 2020b).

Smart, privacy-protective, re-use of health data can, however, improve the safety and quality of care, support better informed health system stewardship and policy making. It can also assist researchers to

[1] The opinions expressed and arguments employed in this article are those of the authors and should not be considered or reported as representing the official views of the OECD or of its member countries.

[2] See: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0347

[3] See: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0362

[4] See: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0433

[5] See: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0463

develop safer and better treatments, and enable more effective disease prevention and public health, resulting in healthier and more productive populations. Yet, health systems remain "data rich but information poor" (OECD, 2019b).

There is arguably no other sector that generates quite as much data and, at the same time, fails to re-use it in effective, useful ways.

There are a range of other factors that can contribute to data under-utilization. Recent OECD reports have highlighted how the legitimate concerns of individuals and organisations may be further compounded by uncertainty in the implementation of laws and policies to protect privacy and to reduce the potential misuse of personal information (OECD, 2019a; OECD, 2019b; OECD, 2021a). Moreover, ensuring that data are handled responsibly and ethically and beyond mere legal compliance is a new challenge for businesses, for the public sector, and society as a whole.

Other barriers may be related to skills. Lack of data-related skills and competences and poor access to computation and storage capacities, for instance, can become bottlenecks preventing the effective re-use of data, even where data are made available via open access (Johnson, 2017). The scale and urgency of the challenge of building digital capacity for data-intensive research appears to be widely under-estimated. Policy initiatives tend to be ad hoc and short-term, with few examples of thorough needs assessments and longer-term strategic initiatives or structural changes to address identified gaps (OECD, 2021b).

Investment can also play an important role. Substantial investments are often required for data cleaning and data curation and for information technology infrastructures to ensure secure data storage, processing, and access. The overall total up-front costs and spending for maintenance can also be very high. The inability to secure returns on these investments can therefore disincentive data sharing.

Finally, one of the most frequently cited barriers to data sharing and re-use is the lack of common standards, or the proliferation of incompatible standards. For example, inconsistent data formats are impediments to the creation of longitudinal data sets, as changes in measurement and collection practices make it hard to compare and aggregate data. This problem is often compounded by the lack of a shared understanding of what quality means in the context of data. Some authors have argued that data quality should be considered a key determinant of trust for data sharing (Sposito, 2017).

Although particularly acute in the health sector (OECD, 2019b), these challenges affect data re-use in all sectors of our economies. Despite the growing need for data and evidence of the economic and social benefits across sectors, including benefits for governments, data access and sharing remains below its potential. Notably, data sharing between companies and with the public sector has not taken off at sufficient scale.

In addition, policy action to promote access to and sharing of data has been uneven at best (OECD, 2019b).

Establishing international consensus on data governance

The OECD has responded directly to these challenges. With its work the organization has influenced and even defined since 2006 data governance policies in OECD countries.

The OECD Recommendation on Access to Research Data from Public Funding (OECD, 2007) (Recommendation on Research Data) represented an important first step in multilateral efforts to create the conditions for opening up access to data in the field of science, technology and innovation (STI).

A review by the OECD in 2009 showed many positive impacts of this recommendation, notably the advancement of science through an accelerated research process, the emergence and development of new avenues of research beyond the initial context in which the data were collected and improved research quality (OECD, 2009). The review however also revealed the need to continue to monitor new developments in the field of STI, which led to the revision of the Recommendation in January 2021.

The Recommendation on Research Data served as inspiration for a host of subsequent multilateral and national policy instruments, including the 2012 European Commission's Recommendation on access to and preservation of scientific information 2012/417/EU, replaced in 2018 by Recommendation (EU) 2018/790[6] and UNESCO's Policy Guidelines for the Development and Promotion of Open Access (UNESCO, 2012), also issued in 2012.

Ten years later, work by the OECD on establishing international consensus about the framework conditions within which health data can be appropriately governed culminated with the 2016 Recommendation on Health Data Governance. OECD Health Ministers welcomed this Recommendation at their meeting in Paris on 17 January 2017[7], along with a request that the OECD undertake further work to support OECD Member and non-Members to further build capacity in this important area. The Recommendation on Health Data Governance applies to the access to, and the processing of, personal health data for health-related public interest purposes, such as improving health care quality, managing health care resources efficiently, contributing to the progress of science and medicine.

The Recommendation on Health Data Governance has provided important guidance to governments during the global COVID-19 pandemic. At the same time the

[6] See: Commission Recommendation (EU) 2018/ 790 – Of 25 April 2018 – On access to and preservation of scientific information (europa.eu).

[7] See: https://www.oecd.org/health/ministerial/ministerial-statement-2017.pdf

pandemic shone a spotlight on the capacity of each countries' health information systems to provide critical information for the public welfare; as well as on aspects of data governance that created obstacles to responding to the pandemic in a timely way (OECD, 2020a).

It particularly underlined the significant need for cross-sectoral re-use of data giving new impetus and direction to OECD's work on the EASD Recommendation. A clear illustration is how during the pandemic digital solutions based on geolocation data emerged as critical tools to help authorities monitor and contain the spread of the virus. Anonymized mobile call data records (CDRs), which provide valuable insights into population movements and are gathered by telecommunications service providers, were shared with governments and re-used to monitor and control the spread of COVID-19 (OECD, 2020c).

Principles and good practice for promoting cross-sectoral reuse of data

The ambitious objective that the OECD has set out to achieve with the EASD Recommendation [OECD/LEGAL/0463] is to facilitate collaboration and the harnessing of all types of new and existing data sources; and to foster innovation across the private and public sectors while protecting the legitimate rights of individuals and organisations. Its development therefore required a co-operative, interdisciplinary and inclusive process involving three OECD Committees: the Committee on Digital Economy Policy (CDEP), the Committee for Scientific and Technological Policy (CSTP), and the Public Governance Committee (PGC). A Joint Steering Group (JSG) of experts was formed to support the work comprising more than ninety experts, including representatives from over thirty OECD Member and partner economies as well as Business at OECD (Business and Industry Advisory Committee, BIAC), the Trade Union Advisory Committee (TUAC), the Civil Society Information Society Advisory Council (CSISAC), and the Internet Technical Advisory Committee (ITAC). In addition, a targeted stakeholder consultation on the Recommendation in draft form was undertaken in February 2021 to seek additional input from major stakeholders in the data ecosystem as well as from academics, whose participation in the JSG was relatively limited.

The Recommendation is divided into three overall sections, covering a total of seven themes:

Section 1 – On "Reinforcing Trust across the Data Ecosystem" deals with: empowering and pro-actively engaging all relevant stakeholders alongside broader efforts to increase the trustworthiness of the data ecosystem; adopting a strategic whole-of-government approach to data access and sharing; and maximising the benefits of data access and sharing, while protecting individuals' and organisations' rights and taking into account other legitimate interests and objectives alongside broader efforts to promote and enable a culture of responsibility for data governance.

Section 2 – On "Stimulating Investment in Data and Incentivising Data Access and Sharing" focusses on: providing coherent incentive mechanisms and promoting conditions for the development and adoption of sustainable business models and markets for data access and sharing;

Section 3 – On "Fostering Effective and Responsible Data Access, Sharing, and Use across Society" deals with: further improving conditions for cross-border data access and sharing with trust; fostering the findability, accessibility, interoperability and reusability of data across organisations, including within and across the public and private sectors; and enhancing the capacity of all stakeholders to effectively use data responsibly along the data value cycle.

In addition to governments, to whom the EASD Recommendation is addressed directly, the Recommendation also encourages data holders, data producers, data intermediaries, and other relevant stakeholders in the data ecosystem to implement or, as appropriate according to their role, support and promote the implementation of this Recommendation.

To help governments and stakeholders in their implementation efforts, the OECD is currently developing a Companion Document to the Recommendation, including guidance on responsible data governance for data access and sharing in the public and private sectors. The Companion Document will provide further clarity on the scope of the Recommendation and it will explain the key concepts that are fundamental for a deep understanding and the effective implementation of the EASD Recommendation. Most importantly, the Companion Document will present country examples on how specific provisions of the EASD Recommendation can be implemented.

A number of governments have already successfully used the EASD Recommendation for the development on their data governance policies. At the High-Level Launch of the Recommendation, held on 10 December 2021 in collaboration with the Danish Business Authority,[8] high-level representatives from the European Commission, Brazil, Norway and Sweden presented their recently adopted policy initiatives on data access and sharing, as a testimony on the role that the EASD Recommendation already had in their policy making. These included in particular the (proposed) EU Data Governance Act (DGA) approved on April 2022 by the European Parliament, which aims to expand the range of public sector data accessible for re-use and create a framework to facilitate data-sharing across the EU (European Commission, 2020); Norway's Data Policy Strategy which articulates national principles for sharing and (re)using data and introduces new approaches to facilitate data sharing such as data factories and regulatory sandboxes; (Norwegian Government, 2021) and Sweden's Data Strategy for Increased Access to Data for AI and Digital innovation, which aims to enhance access and reuse of public sector data, foster sharing and reuse

[8] See www.oecd.org/digital/ieconomy/easd-recommendation-launch-agenda.pdf

of data across the private sector, and businesses and across industries for a responsible, ethical and fair data economy (Swedish Government, 2021).

A clear trend can be observed towards National Data Strategies (NDaS) that help address data governance issues in a comprehensive manner that incorporate a whole-of-government perspective as called for by the Recommendation. NDaS can be instrumental in creating the conditions for effective data governance frameworks to better protect the rights of individuals and organisations, while providing the flexibility needed for all to benefit from data access and sharing. This is why the OECD is currently assessing the potential of NDaS in the context of Phase III of its Going Digital Horizontal Project on Data Governance for Growth and Well-Being, its flagship project that aims to support governments in their efforts to develop, revise or implement coherent data governance policies across sectors and jurisdictions.

Bibliography

EUROPEAN COMMISSION (2020), "Proposal for a regulation of the european parliament and of the council on European data governance (Data Governance Act)", November 25, retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0767

JOHNSON P. E. (2017), "The Cost(s) of Geospatial Open Data", *Transactions in GIS*, vol. 21/3, http://dx.doi.org/10.1111/tgis.12283

NORWEGIAN GOVERNMENT (2021), *Data som ressurs – Datadrevet økonomi og innovasjon,* retrieved from: www.regjeringen.no/no/dokumenter/meld.-st.-22-20202021/id2841118/

OECD. (2007), "OECD Principles and Guidelines for Access to Research Data from Public Funding", OECD Publishing, https://doi.org/10.1787/9789264034020-en-fr, retrieved from: http://www.oecd-ilibrary.org/docserver/download/9207043e.pdf?expires=1505727205&id=id&accname=guest&checksum=8FA3DE023D88AB61F2281628EAC7DC2

OECD (2009), "Access to Research Data: Progress on Implementation of the Council Recommendation OECD", http://www.oecd.org/document/15/0,3343,en_2649_34269_25998799_1_1_1_1,00.html

OECD (2019a), "Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies", OECD Publishing, Paris, https://doi.org/10.1787/276aaca8-en

OECD (2019b), "Health in the 21st Century: Putting Data to Work for Stronger Health Systems", OECD Publishing, Paris, https://doi.org/10.1787/e3b23f8e-en

OECD (2020a), "Beyond Containment: Health systems responses to Covid-19 in the OECD", OECD, https://oecd.dam-broadcast.com/pm_7379_119_119689-ud5comtf84.pdf

OECD (2020b), "Ensuring data privacy as we battle COVID-19- OECD Policy Responses to Coronavirus (COVID-19)", OECD Publishing, Paris, https://doi.org/10.1787/36c2f31e-en

OECD (2020c), "Tracking and Tracing Covid: Protecting privacy and data while using apps and biometrics", OECD, https://read.oecd-ilibrary.org/view/?ref=129_129655-7db0lu7dto&title=Tracking-and-Tracing-COVID-Protecting-privacy-and-data-while-using

OECD (2021a), "Report on the Implementation of the Recommendation of the Council Concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data", doi:https://one.oecd.org/document/C(2021)42/en/pdf

OECD (2021b), "OECD Science, Technology and Innovation Outlook 2021: Times of Crisis and Opportunity", OECD Publishing, Paris, https://doi.org/10.1787/75f79015-en

SPOSITO F. (2017), "What do data curators care about? Data quality, user trust, and the data reuse plan", http://library.ifla.org/1797/1/S06-2017-sposito-en.pdf

SWEDISH GOVERNMENT (2021), "Data – En underutnyttjad resurs för Sverige", retrieved from European Commission (2020, November 25), proposal for a regulation of the european parliament and of the council on European data governance (Data Governance Act), retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0767

JOHNSON P. E. (2017), "The Cost(s) of Geospatial Open Data", *Transactions in GIS*, vol. 21/3, http://dx.doi.org/10.1111/tgis.12283

NORWEGIAN GOVERNMENT (2021), "Data som ressurs – Datadrevet økonomi og innovasjon", retrieved from: www.regjeringen.no/no/dokumenter/meld.-st.-22-20202021/id2841118/

OECD (2007), "OECD Principles and Guidelines for Access to Research Data from Public Funding", OECD Publishing, https://doi.org/10.1787/9789264034020-en-fr. Retrieved from: http://www.oecd-ilibrary.org/docserver/download/9207043e.pdf?expires=1505727205&id=id&accname=guest&checksum=98FA3DE023D88AB61F2281628EAC7DC2

OECD (2009), "Access to Research Data: Progress on Implementation", Recommendation of the Council, OECD, http://www.oecd.org/document/15/0,3343,en_2649_34269_25998799_1_1_1_1,00.html

OECD (2019a), "Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies", OECD Publishing, Paris, https://doi.org/10.1787/276aaca8-en

OECD (2019b), "Health in the 21st Century: Putting Data to Work for Stronger Health Systems", OECD Publishing, Paris, https://doi.org/10.1787/e3b23f8e-en

OECD (2020a), "Beyond Containment: Health systems responses to Covid-19 in the OECD", OECD, https://oecd.dam-broadcast.com/pm_7379_119_119689-ud5comtf84.pdf

OECD (2020b), "Ensuring data privacy as we battle COVID-19- OECD Policy Responses to Coronavirus (COVID-19)", OECD Publishing, Paris, https://doi.org/10.1787/36c2f31e-en

OECD (2020c), "Tracking and Tracing Covid: Protecting privacy and data while using apps and biometrics", OECD, https://read.oecd-ilibrary.org/view/?ref=129_129655-7db0lu7dto&title=Tracking-and-Tracing-COVID-Protecting-privacy-and-data-while-using

OECD (2021a), "Report on the Implementation of the Recommendation of the Council Concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data", doi:https://one.oecd.org/document/C(2021)42/en/pdf

OECD (2021b), "OECD Science, Technology and Innovation Outlook 2021: Times of Crisis and Opportunity", OECD Publishing, Paris, https://doi.org/10.1787/75f79015-en

SPOSITO F. (2017), *What do data curators care about? Data quality, user trust, and the data reuse plan,* http://library.ifla.org/1797/1/S06-2017-sposito-en.pdf

SWEDISH GOVERNMENT (2021), "Data – En underutnyttjad resurs för Sverige", retrieved from: www.regeringen.se/4aa1b8/contentassets/56abce3c5f6447a1a285602718b86ad1/data--en-underutnyttjad-resurs-for-sverige-en-strategi-for-okad-tillgang-av-data-for-bl.a.-artificiell-intelligens-och-digital-innovation#:~:text=Insatsomr%C3%A5de%201%3A%20%C

UNESCO (2012), "Policy guidelines for the development and promotion of open access", UNESCO, https://unesdoc.unesco.org/ark:/48223/pf0000215863

La CNIL face aux enjeux de la construction d'une société numérique de confiance

Par Étienne MAURY
CNIL

Assurer la confiance au temps du numérique répond pour la CNIL à des enjeux juridiques et réglementaires, mais aussi éthiques, technologiques, économiques et sociétaux. Son action et son rôle évoluent, celle-ci devant faire face à des défis multiples, qui sont non seulement nationaux mais aussi européens et internationaux, compte tenu de la géographie de l'économie numérique globalisée. L'appréhension de ces défis constitue le cadre dans lequel la CNIL inscrit sa stratégie pour répondre aux problématiques actuelles et à venir. À la nécessité de garantir l'effectivité du droit fondamental à la protection des données personnelles s'ajoutent des enjeux connexes et intrinsèques à l'évolution de l'écosystème numérique, que ce soit en matière d'accompagnement et d'innovation, mais aussi de développements législatifs et réglementaires au niveau européen.

Créée par la loi Informatique et Libertés du 6 janvier 1978[1], la Commission nationale de l'informatique et des libertés (CNIL) est au cœur des problématiques actuelles. Son rôle de régulateur des données personnelles doit s'appréhender à la lumière des enjeux d'un monde en évolution, se déclinant entre transformation numérique, réponse à la crise sanitaire, risque cyber, développements technologiques et démultiplication des flux de données.

Ce contexte n'est pas conjoncturel. Il est le fruit d'une révolution déjà à l'œuvre et que la CNIL a depuis plusieurs années anticipée et appréhendée. La mission première de cette autorité indépendante, à savoir celle de préserver les libertés des citoyens à l'ère du tout-numérique en accompagnant et en contrôlant l'usage des données personnelles, reste ainsi pleinement d'actualité. Pour aborder la question de la confiance au temps du numérique, il n'est d'ailleurs pas inutile de rappeler les termes de l'article 1er de la loi de 1978 : « L'informatique doit être au service de chaque citoyen. Son développement doit s'opérer dans le cadre de la coopération internationale. Elle ne doit porter atteinte ni à l'identité humaine, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques. »

La CNIL a adopté récemment son nouveau plan stratégique pour 2022-2024, avec l'objectif d'être aux côtés de l'ensemble des personnes et parties prenantes pour parvenir à « la construction d'une société numérique de confiance ». Cette confiance doit bien évidemment s'entendre comme celle manifestée par les individus, dont le quotidien passe désormais de plus en plus par des usages numériques. Mais elle est aussi celle accordée par les acteurs publics, les entreprises et la société dans son ensemble, tant les flux de données sous-tendent aujourd'hui des pans entiers de l'économie et de la vie quotidienne.

Pérenniser la dynamique du RGPD et assurer une protection effective du droit des personnes

Adopté en 2016 et entré en application effective le 25 mai 2018, le Règlement général sur la protection des données (RGPD)[2] a non seulement permis la mise à jour du cadre juridique européen en matière de données personnelles, mais aussi d'engager une dynamique réglementaire allant bien au-delà des seules frontières de l'Union. Il s'agit d'ailleurs d'un des rares textes européens dont l'acronyme est désormais familier pour bon nombre de citoyens de l'UE, mais également pour tout acteur du numérique aux niveaux européen et international. Depuis son adoption, un certain nombre de pays ont d'ailleurs procédé à une mise à jour de leur propre cadre national en matière de protection des données. C'est le cas, par exemple, de la Suisse, du Japon, de la Corée du Sud, du Bénin ou encore de l'Australie. D'autres États ont, quant à eux, adopté, pour la première fois, un cadre juridique général pour les traitements des données personnelles,

[1] Loi n°78-17 du 6 janvier 1978 modifiée relative à l'informatique, aux fichiers et aux libertés.

[2] Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (Règlement général sur la protection des données).

dont les principales dispositions peuvent se rapprocher de celles du RGPD. C'est notamment le cas de la Californie avec le California Consumer Privacy Act adopté en octobre 2018 et entré en application au 1er janvier 2020, ou encore du Brésil avec la Lei Geral de Proteçao de Dados adoptée en 2019.

L'impact du RGPD au niveau international et son rayonnement, en tant qu'instrument du *soft power* européen, ne doivent toutefois pas conduire à considérer que son adoption suffirait à elle seule à permettre une protection effective des personnes dans une économie globalisée. Sa mise en œuvre et le contrôle du respect de ses dispositions sont des enjeux permanents, qui sont au cœur de l'action au quotidien de la CNIL, que ce soit au niveau national, européen ou international.

Assurer une protection effective du droit des personnes, notamment en contrôlant et en sanctionnant, se traduit concrètement pour la CNIL par une activité particulièrement intense, qui est aussi le reflet de ce mouvement engagé avec l'application effective du RGPD. Les plaintes reçues par la CNIL ont ainsi quasiment doublé entre 2016 et 2019, passant de 7 700 saisines annuelles à 14 000 plaintes environ, et se maintenant depuis au même niveau. Ainsi, en 2021, la CNIL a enregistré 14 143 plaintes et en a clôturées 12 522. Elle a par ailleurs reçu plus de 5 000 notifications de violation de données. L'autorité a également procédé, en 2021, à 384 contrôles. Au titre de cette même année, les manquements constatés à l'occasion de certaines des instructions qu'elle a menées, l'ont conduit à prononcer 135 mises en demeure et 18 sanctions, pour un montant d'amendes historique qui, en cumulé, dépasse les 214 millions d'euros[3].

Pour répondre à cette activité particulièrement intense et compte tenu des ressources dont elle dispose, la CNIL s'efforce d'orienter son action de façon stratégique, en concentrant ses efforts sur la clarification du cadre légal du droit précité et le développement d'une offre d'accompagnement adaptée aux besoins des acteurs, tout en mettant en œuvre une réponse répressive proportionnée mais suffisamment dissuasive, un pouvoir de sanction qu'elle mobilise avec discernement à l'encontre des acteurs de toutes tailles et de tous les secteurs confondus. Cette stratégie s'est notamment illustrée à travers le plan d'action dédié aux *cookies* et autres traceurs en ligne. L'autorité a d'abord posé un cadre clair, accompagné une multitude d'acteurs et donné à ceux-ci le temps nécessaire à leur mise en conformité. Puis elle a sanctionné des pratiques non conformes afin d'assurer, pour les internautes, un choix réel et éclairé au regard de la collecte de leurs données en ligne. En s'appuyant sur le cadre juridique applicable et en concentrant son action sur une problématique du quotidien numérique des personnes, la CNIL a ainsi permis une évolution concrète et visible des pratiques au sein de l'écosystème numérique. Elle ambitionne de poursuivre cette politique dans son plan stratégique qui couvrira les deux années à venir en accordant la priorité à des actions de régulation ciblées sur des sujets à fort enjeu pour la vie privée des personnes concernées, et ce afin de faire face à l'intensification et la diversification des usages de données personnelles.

Pour la CNIL, c'est aussi au niveau européen que se joue la prise en compte de la protection des droits des personnes par les grands acteurs du numérique. L'autorité y joue, traditionnellement, un rôle moteur. Elle poursuit ses efforts de manière déterminée pour accroître l'efficacité du mécanisme de coopération et de cohérence (ou « guichet unique ») instauré par le RGPD. Ce système de régulation est inédit au niveau européen et repose sur une logique à la fois intégrée et décentralisée, impliquant chacune des autorités de protection des données de l'UE, qui restent le point d'entrée pour toute personne concernée et assurent une application et une mise en œuvre cohérentes du règlement au sein du Comité européen de la protection des données (CEPD) qu'elles composent. La capacité des autorités de l'UE à mettre en place ce nouveau mécanisme était en soi un véritable défi. Il est aujourd'hui opérationnel. Plus de 800 procédures de coopération ont été mises en œuvre au niveau européen entre 2018 et 2021, à l'issue desquelles près de 300 décisions finales ont été prononcées. Collectivement, les autorité de protection des données ont en quatre ans adopté pas moins de 57 lignes directrices et 6 recommandations. Leur action répressive s'est également intensifiée, avec des amendes s'élevant, en cumulé, à 1,55 milliard d'euros fin 2021.

Ce mécanisme suscite, il est vrai, encore des attentes, en particulier concernant les actions et décisions à prendre vis-à-vis des grandes plateformes, mais aussi au regard des obstacles à une coopération européenne optimale qui demeurent. Les autorités de protection de l'Union européenne sont toutefois collectivement engagées dans la réussite de ce modèle ; elles ont d'ailleurs récemment réaffirmé cette volonté d'assurer une coopération encore plus étroite, en priorisant les cas stratégiques et d'envergure, en facilitant la conduite d'opérations conjointes et en favorisant l'échange d'informations entre les autorités[4]. Au-delà de la mise en œuvre et de la bonne application du droit de l'Union, cette stratégie vise aussi à assurer la crédibilité du modèle européen en matière de régulation du numérique, dont le RGPD a constitué l'une des premières étapes déterminantes.

Favoriser la confiance par l'accompagnement, l'anticipation et l'innovation

Le contrôle du respect du cadre juridique relatif à la protection des données, s'il est essentiel pour assurer une protection effective du droit des personnes, ne saurait à lui seul permettre de répondre à l'ensemble des enjeux présents et émergents dans l'écosystème du numérique. Il doit nécessairement s'appuyer sur une

[3] Commission nationale de l'informatique et des libertés, Rapport annuel 2021, mai 2022.

[4] Comité européen de la protection des données, Déclaration sur la coopération, 28 avril 2022.

approche globale de la régulation ; c'est également en ce sens que la CNIL œuvre à travers ses priorités en matière d'accompagnement et d'innovation.

Pour répondre à la complexité grandissante des textes relatifs à la protection des données personnelles, la CNIL a fait le choix, et ce depuis de nombreuses années, d'accompagner les professionnels des organismes publics et privés dans toute leur diversité : les pouvoirs publics, les responsables de traitements ou leurs sous-traitants, les associations professionnelles ou encore les fournisseurs de solutions techniques, technologiques ou méthodologiques. Cette stratégie se poursuit aujourd'hui au travers de ses efforts renforcés en faveur du développement de nouveaux outils de la conformité et de stratégies dédiées à certains types d'acteurs ou certains secteurs. La CNIL a également choisi d'innover en mettant en place pour la première fois, en 2021, un « bac à sable » en matière d'accompagnement : pour un nombre restreint de projets sélectionnés, elle offre un accompagnement agile et personnalisé aux porteurs de ces projets avec pour objectifs de les aider à mieux connaître le terrain, d'apporter des réponses à des questions juridiques et technologiques nouvelles et, enfin, d'accompagner l'innovation en suivant le rythme des innovateurs eux-mêmes.

L'anticipation et l'innovation constituent également des missions à part entière pour la CNIL, et même une nécessité, compte tenu des évolutions rapides et multiples de l'environnement numérique. Son laboratoire d'innovation numérique (LINC), créé en 2016 et devenu en 2021 un service à part entière, traduit la volonté de cette autorité indépendante de s'ouvrir à de nouveaux enjeux, d'analyser l'émergence de nouvelles technologies et d'étudier de nouveaux usages du numérique, tout en favorisant l'échange, l'expérimentation et les approches pluridisciplinaires. La loi de 2016 pour une République numérique a également confié à la CNIL une mission relative aux questions éthiques et de société posées par les nouvelles technologies. Elle organise ainsi chaque année un débat public, au croisement d'expertises de terrain et scientifiques, autour des nouveaux enjeux du numérique : sont ainsi abordés des sujets comme les droits et libertés numériques au travail, les *civic tech* ou encore l'ouverture et le partage des données.

C'est notamment dans le cadre de cette mission éthique que la CNIL a engagé, dès 2017, une réflexion publique sur les algorithmes et l'intelligence artificielle. Depuis plusieurs années, le déploiement de technologies d'intelligence artificielle (IA) s'est en effet très largement intensifié. Ce développement technologique engendre de nouvelles approches qui bouleversent les façons de faire et soulèvent des questions cruciales et complexes, en particulier en termes de protection des données. Certains des grands principes de la loi Informatique et Libertés et du RGPD sont parfois mis en tension par les présupposés fondateurs de l'IA. La CNIL mène donc d'importants travaux afin de préciser la manière d'assurer la conformité avec le droit des traitements de données recourant à ces systèmes. Très récemment, afin de permettre aux organismes d'évaluer par eux-mêmes la maturité de leurs systèmes d'IA au regard des dispositions du RGPD et des bonnes pratiques dans le domaine, la CNIL a élaboré une grille d'analyse desdits systèmes. L'objectif est d'inviter les organismes prévoyant de mettre en place un traitement utilisant des technologies d'IA, ou ayant déjà initié cette démarche, à se poser les questions qui, en matière de gestion des données personnelles et d'éthique, doivent leur permettre d'assurer leur conformité par rapport au RGPD. Ces nouvelles ressources dédiées à l'IA s'inscrivent aussi dans une stratégie européenne visant à stimuler l'excellence dans ce domaine et à contribuer au débat législatif et réglementaire en cours, avec en particulier la proposition d'un règlement européen sur l'IA.

Assurer la cohérence et contribuer à la gouvernance de la régulation européenne du numérique

L'encadrement réglementaire des nouveaux usages de la donnée et l'effectivité du respect de ce cadre juridique en matière de protection des données personnelles, s'ils sont essentiels, ne sont pas, à eux seuls, à même de répondre à l'ensemble des défis à relever pour construire une société numérique de confiance. Le partage des données, les positions dominantes de marché et le rôle joué par certaines plateformes considérées comme des contrôleurs d'accès (« *gatekeeper* »), ou encore la transparence et la régulation des contenus sont autant de problématiques connexes qu'il convient d'appréhender dans cette perspective. À l'instar de la proposition de règlement pour l'encadrement de l'IA, l'Union européenne a présenté, en l'espace de moins de deux ans, toute une série de propositions législatives visant à mieux encadrer l'environnement numérique – à la fois son marché, ses pratiques et ses dynamiques – avec pour objectif la mise en place d'un véritable modèle européen pour la régulation de l'écosystème numérique : en la matière, l'on peut citer le règlement sur la gouvernance des données (DGA), le règlement sur les services numériques (DSA), le règlement sur les marchés numérique (DMA) ou encore le règlement sur les données (Data Act). Ces initiatives traduisent une réelle volonté politique de l'UE de rééquilibrer les asymétries actuelles dans l'écosystème du numérique, et de pouvoir véritablement garder la main sur cet environnement en évolution constante.

Le rôle de la CNIL et de ses homologues de l'UE, ainsi que les défis auxquels ces autorités s'efforcent de répondre, doivent aussi s'appréhender à la lumière des évolutions réglementaires à venir. Les autorités compétentes en matière de protection des données européennes se sont d'ailleurs positionnées collectivement, au travers de plusieurs avis et déclarations du Comité européen de la protection des données[5]. Elles considèrent en particulier que le RGPD, fondé sur la protection d'un droit fondamental, doit rester un socle pour poursuivre la construction de la régulation

[5] Comité européen de la protection des données, Déclaration sur le paquet « Services numériques » et la stratégie pour les données, 18 novembre 2021.

européenne du numérique. Elles mettent en avant à ce titre deux dimensions d'un même élément clé, la cohérence, pour garantir la bonne articulation entre eux des différents textes applicables.

Viser la cohérence, tout d'abord, pour s'assurer que les futures dispositions de ces nouveaux règlements visant à mieux encadrer l'écosystème numérique n'interfèrent pas avec – ou n'altèrent pas –, dans leur application, les droits et obligations résultant du RGPD. Il s'agit d'un point majeur sur lequel les CNIL européennes ont attiré l'attention des colégislateurs européens, dans le but de garantir le droit fondamental à la protection des données, mais aussi pour assurer la bonne articulation et la lisibilité des différents cadres juridiques.

Cette cohérence doit aussi prévaloir dans le cadre du modèle de gouvernance de la régulation du numérique ainsi mis en place. Il est en effet essentiel que le contrôle de l'application des nouveaux textes à venir se fasse dans le cadre d'une gouvernance intelligente de la régulation, avec la désignation d'autorités compétentes qui soient en capacité d'agir, fortes de leur expertise et de leur connaissance pratique des acteurs et enjeux du secteur. La CNIL et les autorités de protection des données européennes considèrent qu'elles ont un rôle important à jouer dans cette gouvernance, en étant parfois directement compétentes ou en étant intégrées dans des mécanismes d'interrégulation qu'il convient de définir et de rendre effectifs. Comme pour le RGPD, cette coopération et cette articulation réglementaire apparaissent essentielles à l'effectivité des nouveaux instruments juridiques récemment adoptés ou en cours d'adoption, avec pour finalité de répondre à l'objectif d'un meilleur encadrement de l'écosystème numérique, un objectif se fondant sur les valeurs européennes communes.

La singularité de l'action de la CNIL, en tant qu'autorité indépendante garante de la protection d'un droit fondamental, s'inscrit dans cette nouvelle trajectoire de régulation et d'interrégulation, qui doit être à même d'aborder les réalités d'un environnement numérique en pleine transformation, que ce soit au niveau national, européen ou international. Le rôle et l'action de la CNIL se poursuivent pour permettre d'accompagner cette transformation, tout en préservant le droit des personnes et en assurant une ouverture des données qui soit porteuse en termes d'efficacité et d'innovation.

Le *Data* altruisme : comment les données peuvent-elles être mises à contribution pour servir l'intérêt général ?

Par Éric SALOBIR
Président du comité exécutif de la Human technology Foundation et fondateur d'OPTIC

Le volume des données ne cesse de croître, mais pourtant celles-ci sont largement sous-utilisées. Ce paradoxe obère fortement des initiatives œuvrant pourtant au profit de l'intérêt général. Le *Data* altruisme, une innovation en matière de partage des données théorisée par la Commission européenne, pourrait être une solution à ce problème en supprimant la méfiance qui entrave ledit partage. Dans le rapport « Le *Data* altruisme : une initiative européenne, les données au service de l'intérêt général », la Human technology Foundation et l'Exploratoire Sopra Steria Next se sont attachés à montrer comment le concept encore théorique de *Data* altruisme peut devenir une réalité. Dans cet article, nous détaillons les propositions majeures de ce rapport et montrons qu'il est possible de construire un système facilitant grandement la mise à disposition de données pour aider des initiatives œuvrant en faveur de l'intérêt général à se construire.

Introduction

Le volume des données numériques explose. Ce dernier a été multiplié par 30 entre 2010 et 2020, et l'augmentation à venir est encore plus exponentielle. En effet, les prévisions détaillent une croissance fulgurante qui devrait générer 181 zettaoctets de données, rien qu'en 2025[1] (64 zettaoctets ont été créés en 2020). Rappelons qu'un zettaoctet correspond à un milliard de teraoctets ! Les capacités de traitement de ces données vont donc jouer un rôle fondamental pour permettre d'utiliser ce volume toujours plus grand de données.

Pourtant, l'utilisation des données a parfois mauvaise presse en raison de scandales comme les affaires Cambridge Analytica ou IQVIA, lesquels ont montré une exploitation abusive et malveillante de la donnée. Cela génère une méfiance envers le partage des données, ce qui induit une grande sous-utilisation de celles-ci. Ainsi, la présidente de la Commission européenne, Ursula Von der Leyen, parle de « 80 % des données industrielles qui seraient non exploitées ». Ces données sont pourtant précieuses et permettraient de faire de grandes avancées dans le développement d'initiatives en faveur de l'intérêt général, telles que l'amélioration des services publics, la lutte contre le réchauffement climatique ou l'amélioration de la mobilité.

Cette nécessité a ainsi amené à la création du concept de *Data* altruisme par la Commission européenne dans son Data Governance Act (DGA), un règlement approuvé par le Parlement européen et le Conseil de L'Union européenne à la fin novembre 2021.

Le *Data* altruisme n'est encore qu'un concept théorique. Dans cet article, nous reprenons les suggestions que nous avons formulées dans notre rapport « Le *Data* altruisme : une initiative européenne, les données au service de l'intérêt général », coécrit par la Human technology Foundation et l'Exploratoire Sopra Steria Next[2]. Dans un premier temps, nous analyserons le périmètre de ce concept ainsi que ses opportunités et ses limites. Ensuite, nous comparerons plusieurs modes de partage des données existants pour déterminer les meilleures pratiques en vue de l'élaboration de modèles *Data* altruistes. Sur ce point, la Figure 2 de la page suivante présente un récapitulatif des principales mesures que nous recommandons pour rendre concret le *Data* altruisme.

[1] « Le *Big Bang* du *Big Data* » (2021), Statista, https://fr.statista.com/infographie/17800/big-data-evolution-volume-donnees-numeriques-genere-dans-le-monde/

[2] Le rapport « *Data* altruisme : une initiative européenne, les données au service de l'intérêt général », https://www.human-technology-foundation.org/fr-news/rapport-data-altruisme

Figure 1 : Couverture du rapport « Le *Data* altruisme : une initiative européenne, les données au service de l'intérêt général ».

Tout d'abord, qu'est-ce que le *Data* altruisme et comment intervient-il dans le partage des données ?

Le *Data* altruisme est défini par le DGA comme un « partage volontaire de données basé sur le consentement donné [...] ou les autorisations accordées [...] à des fins d'intérêt général, telles que la santé, la lutte contre le changement climatique [...] ou l'élaboration de politiques publiques. »

Il s'agit donc d'un cadre formel de partage des données que décrit le DGA. Pour inciter les acteurs à mettre en œuvre ce partage, le texte autorise des mécanismes de financement pour couvrir la mise en place et la maintenance des infrastructures de partage, ainsi que le recours à des incitations directes ou indirectes pour pousser les contributeurs à partager leurs données.

Il apparaît dès lors nécessaire de définir les différents acteurs impliqués dans un partage de données pour comprendre la nouveauté du *Data* altruisme.

De façon générale, les acteurs sont :

- les sujets (producteurs) des données : par exemple, un individu possédant une montre connectée et qui génère de la donnée *via* son application de sport ;
- le détenteur des jeux de données : par exemple, la société éditrice de l'application utilisée sur la montre connectée ; cette société rassemble les données de tous les utilisateurs de son application ;

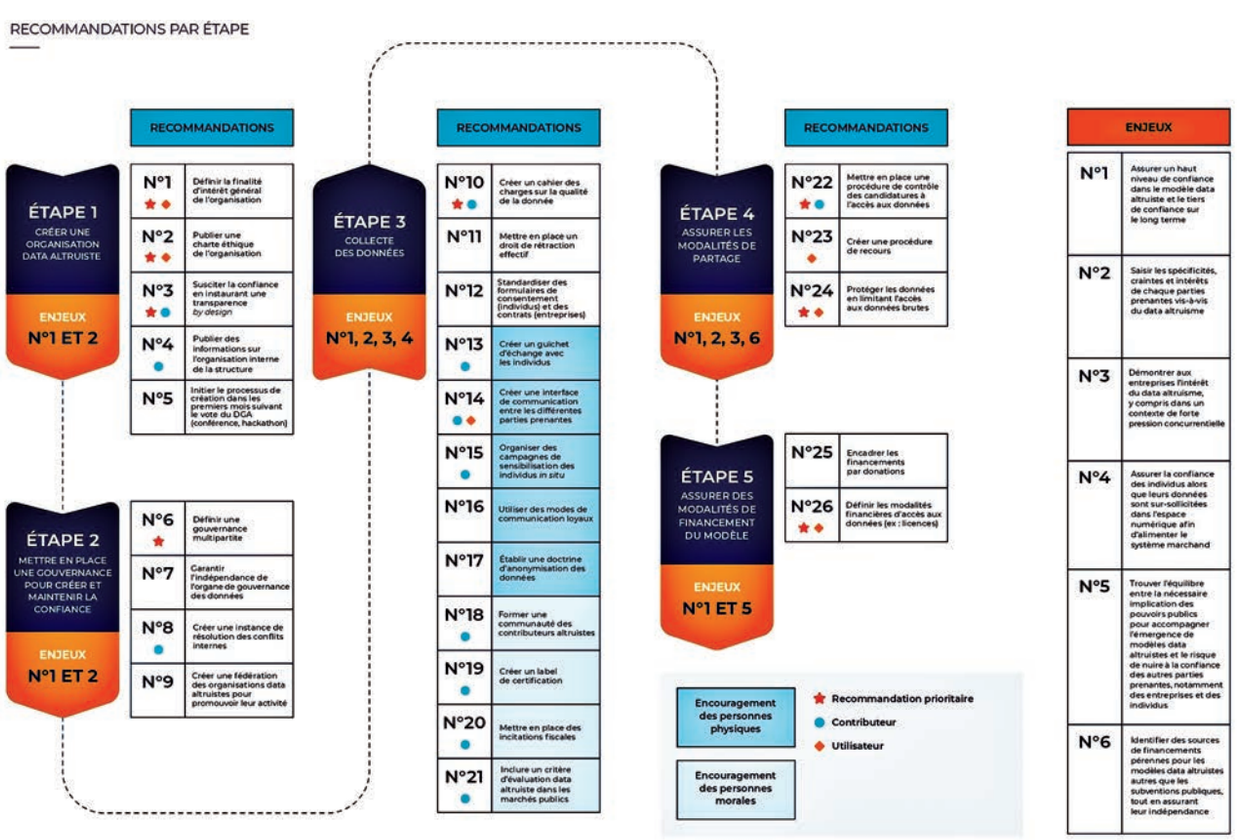


Figure 2 : Principales recommandations du rapport « Le *Data* altruisme : une initiative européenne, les données au service de l'intérêt général ».

• l'utilisateur des jeux de données : ce peut être le détenteur (voir ci-dessus), mais aussi un acteur tiers comme un équipementier sportif qui achèterait au détenteur ses jeux de données pour réaliser des études spécifiques dans le but, par exemple, de créer un nouveau produit ou lancer une campagne publicitaire ;
• l'intermédiaire de données : il s'agit d'un acteur neutre, dont la seule mission est d'opérer techniquement le transfert de données entre un sujet/détenteur et un autre acteur comme un utilisateur ou une organisation *Data* altruiste. Il se contente de faire circuler les données sans les exploiter ;
• Dans le *Data* altruisme introduit par le DGA, il est suggéré la création d'un cinquième acteur qui se placerait entre les détenteurs et les utilisateurs : il serait chargé d'assurer la collecte des données auprès des détenteurs ainsi que le partage de celles-ci avec les utilisateurs, avec le concours de l'intermédiaire de données qui garantit l'opérationnalité des infrastructures de partage.

Ce modèle est innovant, car il s'agit d'une approche spécifiquement européenne qui est radicalement différente de celles adoptées par d'autres grandes puissances comme la Chine ou les États-Unis. L'absence de reconnaissance d'un droit à la protection des données personnelles aux États-Unis ou en Chine comparable au modèle européen explique l'originalité du modèle *Data* altruiste. En effet, ce modèle découle des choix européens faits en matière de protection des données, il cultive donc une différence certaine par rapport aux autres systèmes dominants. La Chine a fait le choix d'un contrôle étroit permettant une supervision des transferts de données. Les données sont considérées comme un bien marchand et peuvent être échangées à ce titre sur le Shanghaï Data Exchange ; cette approche est à l'opposé d'une vision altruiste et est en contradiction avec les règles de protection des données personnelles développées dans l'Union européenne. Les États-Unis, quant à eux, s'appuient sur le principe d'extraterritorialité qui s'applique à leurs réglementations et à leurs multinationales pour collecter auprès d'entreprises ou d'individus étrangers leurs données personnelles. Cette approche est impossible pour l'Europe qui ne dispose pas, contrairement aux États-Unis, de géants des données justifiant d'un quasi-monopole ; cela explique pourquoi la Commission européenne a cherché à promouvoir la stratégie d'un « pas de côté » pour rattraper le retard des pays européens en matière de captation et d'exploitation des données face aux deux modèles hégémoniques que sont la Chine et les États-Unis.

L'innovation vient également du but poursuivi par ce système altruiste, à savoir œuvrer en faveur de l'intérêt général. Il convient donc de définir au préalable cette notion « d'intérêt général » pour pouvoir pleinement appréhender le périmètre attribué au *Data* altruisme.

« L'intérêt général » n'a pas de définition légale officielle, mais certaines de ses caractéristiques sont détaillées dans différents textes. Ainsi, en droit fiscal français, l'intérêt général se conçoit comme une activité non lucrative à caractère philanthropique qui n'est pas mise en œuvre au profit d'un cercle restreint de personnes, et dont la gestion est désintéressée. Sa finalité transcende les intérêts individuels et permet aux individus de participer à la société. Cette notion incite à créer un projet commun incarné par un intérêt supérieur qui permettrait à chaque individu d'apporter sa contribution à la collectivité.

Toutefois, l'approche du *Data* altruisme reste une approche nouvelle et doit donc résoudre un certain nombre de défis pour pouvoir être acceptée par le plus grand nombre.

Le premier défi est celui d'obtenir la confiance des différents acteurs. Plusieurs scandales ont fortement fragilisé la confiance des parties prenantes envers le partage des données. Un chantier sur la confiance est nécessaire ; cela implique de réussir à trouver les bons arguments pour convaincre, notamment en insistant sur les modalités de sécurisation des données, en rappelant la finalité du partage, son caractère altruiste, en décrivant les institutions garantes de la bonne exécution de ce partage, tout en prenant en compte les spécificités culturelles de chacun des pays de l'Union européenne (UE).

Le deuxième défi est l'indépendance des organisations *Data* altruistes, et donc la nécessité de s'assurer qu'elles disposent de sources de financement pérennes.

En effet, nombreux sont les coûts liés à l'activité *Data* altruiste :

• la collecte des données ;
• le stockage de celles-ci ;
• l'enrichissement des données : ajout d'informations supplémentaires ;
• la standardisation : ramener à un même format des données venant de différentes sources ;
• combler les manques d'informations (avec des données pivot, par exemple).

Un modèle de redevances proportionnées ou de frais administratifs pour pouvoir avoir accès aux données mises à disposition par l'organisation est une solution viable prévue par le DGA. En plus de ce modèle économique, il est tout à fait possible d'avoir recours à des financements extérieurs (mécénat, subventions, soutien technique) à condition de garantir que l'organisation maintiendra son indépendance et que les contributeurs ne seront pas exclus de la création de valeur.

Quelles bonnes pratiques parmi les modes de partage existants pourraient servir au développement du *Data* altruisme ? Mais il faut être prudent au regard des dérives liées à ce partage

Dans notre rapport, nous étudions 56 initiatives qui sont divisées en quatre secteurs (santé, transports, environnement et gestion des informations personnelles) afin de comprendre l'écosystème du partage des données. Les modes de partage novateurs étudiés montrent que plusieurs formes de *Data* altruisme sont possibles, avec des orientations tout à fait différentes, mais aussi

qu'un certain nombre de risques existent et menacent la confiance sur laquelle repose tout le mécanisme. Les systèmes étudiés nous renseignent en outre sur les bonnes pratiques à transposer dans la démarche *Data* altruiste ; ces dernières sont détaillées pour chaque secteur.

Enseignements à tirer des dynamiques observées dans le secteur de la santé

Le domaine de la santé est marqué par une quantité importante de données notamment du fait de l'existence de systèmes de sécurité sociale dans de nombreux pays européens. Toutefois, ce volume important ne signifie pas que les données soient facilement utilisables, car cette exploitation se heurte à deux enjeux majeurs : la qualité des données et la question de la sécurité de cette exploitation. Il faut en effet parvenir à avoir des données utilisables qui soient représentatives de tous les cas rencontrés (la qualité), tout en garantissant que ces données ne permettent pas d'identifier un individu (la sécurité).

Le problème crucial pour le *Data* altruisme dans le secteur de la santé serait donc la difficulté à décrire une finalité en amont du projet pour inciter au partage. En effet, la recherche médicale se caractérise par l'existence de besoins non identifiés. Comment alors arriver à convaincre les contributeurs si l'on ne peut définir en amont une finalité suffisamment précise et compréhensible ? Les questions de gouvernance évoquées précédemment sont également cruciales pour rassurer la grande variété des acteurs de la santé dont les intérêts sont souvent différents : l'Assurance maladie, les hôpitaux, les pharmacies, les laboratoires, etc.

Enseignements à tirer des dynamiques observées dans le secteur des transports (la mobilité)

Le secteur de la mobilité est en avance dans le domaine du *Data* altruisme grâce à l'existence de diverses initiatives et propositions émanant des pouvoirs publics. Ainsi, la notion de donnée d'intérêt territorial évoquée dans un rapport parlementaire de 2017[3] aborde déjà l'idée de la transmission des données générées dans le cadre privé vers la sphère publique pour des finalités d'intérêt général. Parmi les initiatives privées, peuvent être cités Strava Metro ou Uber Movement, qui visent à aider les acteurs de la mobilité à améliorer les infrastructures de transport.

Toutefois, ce secteur rencontre des difficultés similaires à celui de la santé, notamment en ce qui concerne la qualité des données. En effet, les différents jeux de données disponibles sont complexes à compiler car ils sont très hétérogènes, notamment dans leur structuration.

Les dynamiques observées dans la mobilité sont en revanche particulièrement portées par les réglementations, comme la portabilité citoyenne qui oblige les acteurs du secteur à transmettre les données générées par leurs utilisateurs à d'autres acteurs si l'un de ces utilisateurs le demande. Ce mécanisme permet de constituer de nouvelles bases de données et d'améliorer la circulation des données sous le contrôle des citoyens.

Enseignements à tirer des dynamiques observées dans le secteur de l'environnement

Deux sous-cas ont été étudiés.

L'énergie

Le recours croissant à des moyens de production non pilotables (solaire, éolien...) a provoqué un plus grand besoin en données pour la production et la distribution de l'électricité. Il faut en la matière parvenir à concilier la demande et l'offre à une granularité de plus en plus fine en tenant compte des aléas propres à ces moyens de production. Le développement des *smart cities* requiert également des systèmes permettant l'envoi et le traitement de données en temps réel, à une maille fine.

La biodiversité

Le *Data* altruisme serait un précieux atout pour la collecte des données portant sur des surfaces et des temporalités longues. Il aiderait en particulier le tiers secteur de la recherche en permettant à un ensemble d'acteurs très divers (associations, collectivités, groupements professionnels...) de pouvoir plus facilement transmettre leurs données.

Enseignements issus d'une perspective canadienne dans le secteur de la gestion des informations personnelles : la fiducie des données

Dans son rapport « Tables de stratégies économiques du Canada » publié en 2018, le gouvernement canadien définit la fiducie de données comme « un organisme formé pour gérer des données pour le compte de ses membres. Ces derniers mettent en commun leurs données et conviennent expressément des conditions afférentes à leur partage ». Il s'agit donc d'un intermédiaire fiable pour le traitement des données s'apparentant au modèle développé en Europe. Cette initiative des fiducies a été encouragée par les pouvoirs publics, par exemple, à travers la Charte numérique du Canada de 2019. Mais ce concept n'a pas encore été réellement adopté. En effet, il manque encore un support clairement défini dans les lois canadiennes pour que ce genre d'entité puisse être créé. Un des échecs connus dans la mise en place de ces fiducies de données est celui de l'Urban Data Trust développé par Sidewalk Labs. Une forte défiance citoyenne, des complications juridiques et une structure parfois ambiguë du fait de son appartenance à Alphabet Inc. ont fait que le projet n'a finalement jamais vu le jour.

La leçon de cet échec est qu'une fiducie de données ne soulève pas uniquement des enjeux technologiques et de sécurité, dès lors que ce modèle vise aussi à bâtir une confiance réciproque entre tous les participants. Cet exemple canadien illustre la nécessité d'avoir un cadre réglementaire et légal clair, notamment en ce qui concerne la protection de la vie privée, pour que les

[3] BELOT Luc (2017), « De la *smart city* au territoire d'intelligence(s) – L'avenir de la *smart city* ».

intermédiaires comme les fiducies de données puissent être considérés comme fiables dans le partage des renseignements d'ordre personnel.

Tous les systèmes de partage étudiés ont également permis de déceler les principaux risques que rencontrerait le *Data* altruisme s'il était mis en pratique :

• la réparation : des entreprises ayant mené des actions néfastes pourraient participer à une démarche *Data* altruiste pour camoufler les dommages qu'elles ont générés en profitant du statut altruiste de l'organisation créée en accord avec les règles du DGA ;

• le cheval de Troie : il s'agit ici d'utiliser les mécanismes du *Data* altruisme de façon malveillante pour capturer les données stratégiques détenues par un concurrent par exemple (cela peut être réalisé de façon directe avec le piratage des serveurs d'une organisation *Data* altruiste ou alors de façon indirecte sous couvert d'un projet *Data* altruiste pour appâter les potentiels contributeurs) ;
• la tromperie : la démarche *Data* altruiste est ici mobilisée pour collecter des données personnelles, notamment celles de consommateurs.

Les données sont-elles devenues le premier enjeu de la cybercriminalité ?

Par Éric FREYSSINET
Officier général de gendarmerie

Les données sont au cœur des enjeux de la cybersécurité, en tant que cible principale des cybercriminels et comme outil non seulement pour ces mêmes délinquants mais aussi pour les défenseurs des systèmes d'information. Cette préoccupation concerne tant les données à caractère personnel que l'ensemble des données sensibles des organisations. Et les risques ne sont pas liés qu'aux célèbres rançongiciels, même s'ils constituent la menace la plus dynamique.

Chaque semaine, plusieurs alertes passent dans les flux d'actualités informant de détournements de données importantes : ainsi, au moment même de la rédaction de cet article, en étaient victimes un groupe de gestion de patrimoine français, une ONG ou encore un casino. Aux États-Unis, un rapport de l'ITRC[1] relevait en 2021 une augmentation de 68 % du nombre de compromissions de données. Au titre de la même année, la CNIL rapportait[2] 2 150 violations de données personnelles.

Derrière ces chiffres se cachent des réalités très variables. Un cas emblématique en est, par exemple, le vol et la diffusion en 2021 des données de 491 840 patients de laboratoires de biologie médicale de l'ouest de la France avec, dans certains cas, non seulement la captation de leurs coordonnées mais également d'informations à caractère médical, telles que les maladies pour lesquelles ils sont suivis ou leurs traitements.

Force est de constater que les données sont devenues le carburant ou, en tous cas, la face visible d'une grande partie des attaques cybercriminelles modernes. Nous nous proposons ici d'explorer les différentes facettes de cette réalité : le cadre légal, l'évolution des pratiques des criminels et les stratégies de ceux qui protègent les systèmes d'information.

La protection des données, un enjeu traduit depuis longtemps dans le droit

Le législateur français ne s'y est pas trompé en faisant de la donnée le cœur de la protection des systèmes d'information. Ainsi, dès 1978, avec la loi Informatique et Libertés[3], le Parlement consacrait l'importance de la protection contre les abus dans l'utilisation des données personnelles : « Article 25 – La collecte de données opérée par tout moyen frauduleux, déloyal ou illicite est interdite. »

Et, dix années plus tard, la loi Godfrain[4] est venue définir les infractions commises à l'encontre des systèmes informatiques en consacrant le concept de « système de traitement automatisé de données », qui se révèle à la fois très général – puisqu'il s'applique non seulement à des ordinateurs, mais également dans les faits à tous les systèmes automatisés qui traitent des données, et donc les réseaux informatiques, les terminaux ou les supports de stockage – et très spécifique, en ce qu'il définit très clairement l'objectif de la loi : protéger les données.

On retrouvera ensuite ce concept dans les traités internationaux qui visent le même objectif, avec, par exemple, la convention du Conseil de l'Europe sur la cybercriminalité qui, adoptée en 2001, définit ainsi un système informatique (lequel doit être protégé par les lois et les enquêtes judiciaires) : « L'expression "système informatique" désigne tout dispositif isolé ou ensemble de dispositifs interconnectés ou apparentés, qui assurent ou dont un ou plusieurs éléments assurent, en exécution d'un programme, un traitement automatisé de données. »

D'autres formes de données bénéficient de protections spécifiques en droit. Ainsi, l'on notera que le Code de la propriété industrielle protège notamment les logiciels informatiques ou les bases de données contre les différentes formes de contrefaçon et d'usage détourné.

Cette approche correspond donc à la fois à une réalité technique – celle des systèmes d'information dont l'objet et les principes de fonctionnement tournent autour de la donnée – et à une réalité opérationnelle – la nécessité de protéger les données et les systèmes qui les traitent. Nous verrons d'ailleurs que les traitements de données sont non seulement une finalité mais aussi un outil pour les attaquants eux-mêmes.

(1) Identity theft resource centre, https://www.idtheftcenter.org/post/identity-theft-resource-center-2021-annual-data-breach-report-sets-new-record-for-number-of-compromises/

(2) Rapport annuel 2021 de la Commission nationale informatique et libertés, https://www.cnil.fr/sites/default/files/atoms/files/cnil_-_42e_rapport_annuel_-_2021.pdf

(3) Loi n°78-17 du 6 janvier 1978 modifiée relative à l'informatique, aux fichiers et aux libertés.

(4) Loi n°88-19 du 5 janvier 1988 relative à la fraude informatique (dite loi Godfrain).

Le rôle des données dans les pratiques des cybercriminels

Les lois sur la cybercriminalité visent donc à protéger les données des atteintes illégales ; des atteintes qui sont en pleine croissance, notamment dans leur forme la plus visible, à savoir les menaces de divulgation de données que profèrent les groupes criminels qui se cachent derrière les rançongiciels.

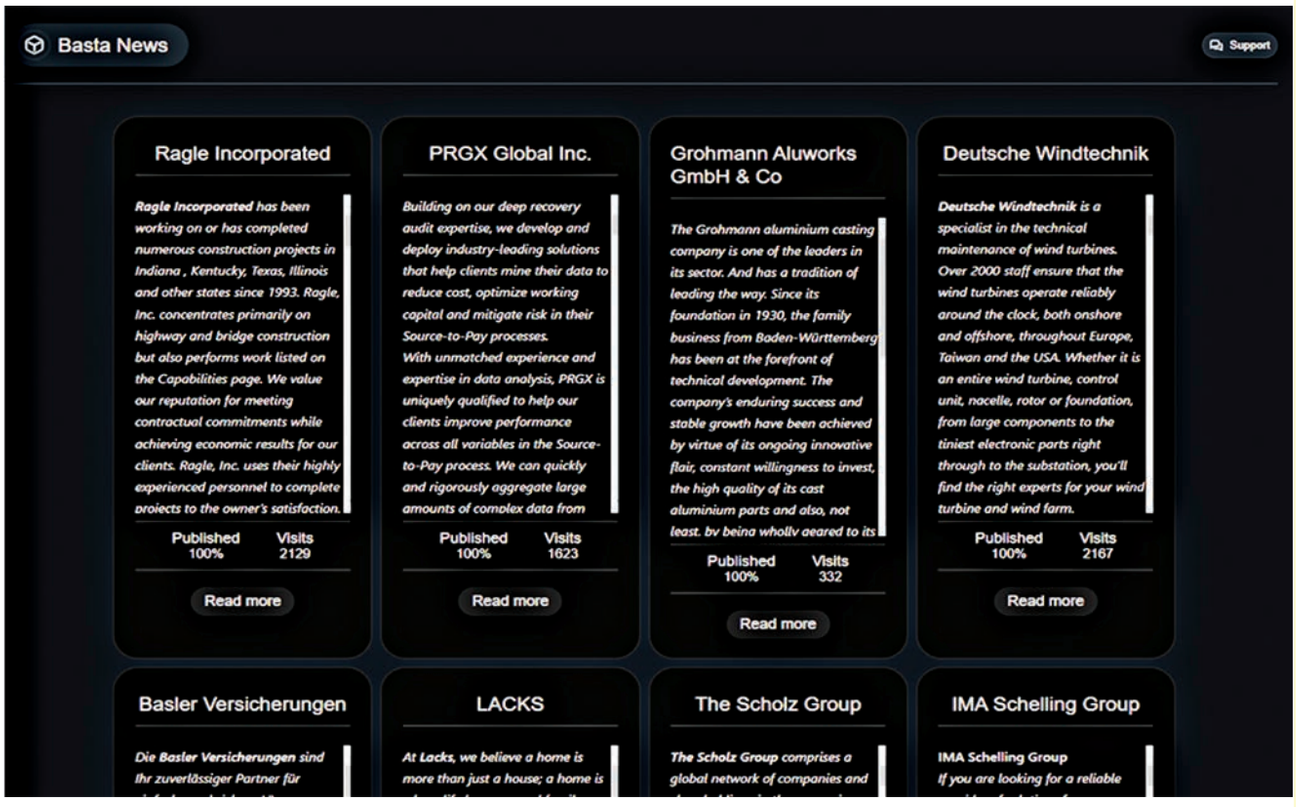
Au nombre d'environ 150, ces groupes de rançonneurs aux organisations variables – certains regroupant quelque dizaines d'affiliés utilisant leur plateforme pour attaquer des victimes spécifiques – utilisent tous aujourd'hui la stratégie dite de la double extorsion, à savoir la menace de ne plus permettre aux utilisateurs légitimes d'accéder à leurs données (le rançongiciel chiffrant les données sur les systèmes informatiques victimes de l'attaque), mais aussi la menace de rendre publiques les données confidentielles de l'organisation ciblée.

La copie d'écran ci-après est celle de la page de l'un des plus récents de ces groupes cybercriminels (surnommé BlackBasta), avec la liste de ses victimes et indiquant le nombre des visiteurs des sites piratés et le volume de données effectivement publiées.

Mais essayons d'avoir une vision plus systématique des différentes stratégies de détournement des données que l'on a pu déjà constater :

- la collecte détournée ou l'*hameçonnage* qui consiste à créer un faux site Web qui reproduit le visuel et le fonctionnement d'un site légitime et dont l'adresse est diffusée par un procédé de masse (typiquement le *spam* diffusé par courrier électronique, SMS ou messagerie instantanée) ;
- une variante de cette technique est, quant à elle, rencontrée dans le monde plus physique des cartes bancaires par le biais d'appareils permettant la capture de copies des pistes magnétiques de ces cartes ou des codes frappés sur les claviers des distributeurs automatiques ou des terminaux ;
- le vol de bases de données à travers l'exploitation d'une faiblesse dans un logiciel d'administration desdites bases (par exemple, grâce à l'injection de commandes dans un formulaire d'un site Web qui normalement contrôle l'accès aux données) ;
- la capture exhaustive d'une base de données en libre accès, dans une logique commerciale en dépit de l'intention de son propriétaire (technique souvent appelée en anglais *Web scraping*), par exemple pour copier un annuaire ou le catalogue d'un concurrent ;
- l'intrusion dans le système d'information de la victime, puis l'extraction de données confidentielles depuis les systèmes attaqués – c'est souvent la première étape de l'attaque mise en œuvre par les groupes criminels de rançonneurs ;
- enfin, une autre variante consiste à utiliser un logiciel malveillant qui va de façon automatique collecter et exfiltrer des données depuis les ordinateurs d'un réseau informatique attaqué.

Mais outre ces actions criminelles, il ne faut pas oublier les divulgations involontaires : certaines organisations ayant mal sécurisé l'accès à leurs données voient celles-ci temporairement ou parfois sur de longues durées librement accessibles à tous les utilisateurs d'Internet. On parlera alors de fuites de données. C'est assez souvent le cas, notamment lorsqu'un système de stockage dans le *Cloud* n'est pas correctement sécurisé.



Les données ciblées sont de différentes natures ; elles peuvent être regroupées dans les catégories suivantes :

- les identifiants et les mots de passe ;
- les données bancaires (y compris les identifiants d'accès aux comptes bancaires) ;
- les bases de données ;
- les fichiers stockés.

Peuvent aussi être ciblées, de façon plus spécifique, des copies d'écran, des images ou des enregistrements sonores détournés à partir des caméras et microphones des ordinateurs ou téléphones de leurs propriétaires.

Toutes ces données ont différents usages et peuvent donc présenter de la valeur pour les attaquants, et ce à différents niveaux :

- les données elles-mêmes peuvent être revendues : chaque catégorie de données a une valeur unitaire qui varie de plusieurs euros à plusieurs centaines d'euros (c'est le cas, en particulier, des identifiants et mots de passe ou des données bancaires) ;
- parfois, la revente concerne plus spécialement un accès détourné à des systèmes d'information (ainsi certaines personnes fournissent à la demande de leurs clients des données qu'ils vont, grâce à un accès légitime, piocher dans un système exploité par la police ou un opérateur de communications) ;
- les données peuvent aussi être utilisées pour démultiplier le potentiel d'une autre attaque : c'est notamment le cas des identifiants de messagerie ou de réseaux sociaux qui peuvent être exploités pour contacter les amis et contacts professionnels d'une première victime dans but d'escroquer de nouvelles victimes mises en confiance par un échange présumé être avec une personne qu'elles connaissent ;
- enfin, le rançonnement à un ou deux niveaux, comme évoqué précédemment.

On observe dans ce champ particulier du marché des données volées une des caractéristiques importantes des activités cybercriminelles modernes, à savoir l'existence d'un véritable écosystème, où des acteurs cybercriminels commercialisent entre eux différents services : il s'agit de développeurs de solutions logicielles (les logiciels malveillants, les plateformes permettant de les piloter), des administrateurs de ces systèmes, de ceux qui les louent pour les mettre en œuvre (souvent appelés affiliés), de spécialistes de l'intrusion, de revendeurs de données et, le cas échéant, de spécialistes du blanchiment des bénéfices retirés de ces différentes opérations.

On le voit donc très clairement, l'exploitation commerciale de la donnée est la motivation principale d'une grande partie des activités cybercriminelles, vraisemblablement sa principale source de financement – à travers la revente ou en favorisant l'extorsion de fonds dans le cadre de chantages.

La donnée peut aussi être utilisée comme un outil pour commettre de nouvelles attaques par rebond vers de nouvelles victimes. D'ailleurs, c'est peut-être dans le cadre de cette dernière dimension que l'on verra s'accroître le nombre des escroqueries dans les années à venir. Car de la même façon que la donnée est indispensable pour protéger les systèmes d'information comme nous l'évoquerons *infra*, elle l'est tout autant pour les attaquants dans leur construction de modèles et de stratégies d'attaque (par une meilleure connaissance de leurs cibles potentielles et de leurs vulnérabilités). Tout comme le seront demain les techniques de l'intelligence artificielle qui nécessiteront pour les cybercriminels l'analyse de gros volumes de données d'apprentissage, mais leur permettront aussi d'étendre et d'accélérer leurs activités. En effet, on peut parfaitement imaginer les cybercriminels exploiter, par exemple, des *chatbots* pour démultiplier le nombre des victimes avec lesquelles ils interagissent.

Les données au cœur des stratégies de cybersécurité

On le voit la première préoccupation d'un responsable d'un système d'information et de ses équipes en charge de la sécurité de celui-ci est donc la protection des données de ce système. Il faut donc être en mesure de maîtriser quels types de données on possède, quelles sont les plus sensibles, quels systèmes les stockent et les traitent, qui y a accès ? Et, au bout du compte, d'être capable de détecter les tentatives d'accès à ces données ou, le cas échéant, les données qui ont pu fuiter.

Cette approche « données » est tout aussi valide dans le cas d'un système informatique servant à réaliser des opérations de production industrielle, car c'est à tout le moins l'intégrité et la disponibilité des données permettant au système industriel de fonctionner que l'on cherchera à préserver et, dans beaucoup de cas, leur confidentialité.

Mais les architectures de protection des systèmes d'information génèrent une couche supplémentaire de données qui servent à protéger lesdits systèmes. Ainsi, on collectera des traces d'activité liées à différentes applications (notamment les traces d'accès aux données), des traces d'interactions sur les réseaux et, de plus en plus souvent, des données liées au fonctionnement des systèmes informatiques eux-mêmes (avec un modèle de sécurité informatique évoluant de l'antivirus vers la notion d'EDR (ou *endpoint detection and response*)).

Au passage, ces données de sécurité sont tout aussi importantes que celles qui étaient initialement protégées ; et la sécurité de leur collecte, de leur stockage et de leur traitement doit elle aussi être préservée.

L'utilisation de ces données doit permettre idéalement de détecter les opérations suspectes et les intrusions en cours. Mais elles peuvent aussi avoir un usage rétrospectif : par exemple, pour comprendre le fonctionnement et l'impact d'une attaque une fois qu'elle a été révélée. C'est l'une des premières opérations que l'on réalisera lorsque l'on découvrira une attaque par un rançongiciel afin, par exemple, de distinguer les systèmes qui sont touchés de ceux qui ne le sont pas.

Les investigations rétrospectives se révèlent également très utiles pour la révélation d'une faille de sécurité inconnue jusque-là (souvent appelées vulnérabilités 0-day). Il s'agit ici de vérifier, grâce à des traces caractéristiques (ou indices de compromission, les IoC), que le système d'information que l'on gère n'a pas été l'objet d'une attaque permettant de l'exploiter. Les mêmes données recueillies dans le cadre de ces investigations pourront bien évidemment être aussi exploitées par les enquêteurs judiciaires.

De même, ces données sont aussi exploitées en masse par les analystes en géostratégie cyberriminelle ou encore par des chercheurs qui s'intéressent au fonctionnement des groupes cybercriminels… Ils vont même jusqu'à exploiter des conversations collectées (ou même qui ont fuité) sur des forums sur lesquels échangent des cybercriminels[5].

Les données sont donc l'objet sur lequel travaillent les responsables de la sécurité des systèmes d'information et sont, dans le même temps, un outil pour eux. Plus que les systèmes eux-mêmes, c'est l'ensemble du cycle de vie des données qu'ils doivent donc maîtriser et savoir exploiter.

[5] C'est un des sujets qui ont plus particulièrement émergé lors de la conférence Botconf, qui s'est tenue à Nantes du 26 au 29 avril 2022 (https://www.botconf.eu/botconf-2021/botconf-2021-22-final-schedule/), comme le relève Louis Adam dans son article *Botconf* : *la nécessité de surveiller les écosystèmes cybercriminels* (https://www.zdnet.fr/actualites/botconf-la-necessite-de-surveiller-les-ecosystemes-cybercriminels-39941429.htm).

Conclusion

On l'a vu, l'enjeu principal de la cybercriminalité peut aujourd'hui se décrire au travers des données qui sont protégées par leurs propriétaires et utilisateurs légitimes, mais qui sont aussi la cible et l'objet de détournements par les délinquants. Ces données sont aussi un outil pour ceux qui sont chargés de les préserver. Demain, la collecte de données permettra tout autant aux défenseurs de connaître les stratégies des attaquants potentiels (*threat intelligence*), qu'aux criminels de développer de nouvelles techniques d'attaque.

Une des conclusions que l'on peut tirer des développements qui précèdent, est qu'il est indispensable aujourd'hui, pour sécuriser les systèmes d'information, de se doter d'une stratégie tournant autour de la donnée et de disposer de bons outils, voire des bons spécialistes du traitement de ces données. Cette vision doit être partagée et transverse (données personnelles et confidentielles, données de production, données de sécurité). C'est à cette fin que de nouveaux métiers ou, en tous cas, de nouvelles compétences se développent dans les équipes de sécurité informatique.

L'enjeu des données pour la cyberdéfense

Par Didier DANET
Maître de conférences (HDR) de l'Université de Rennes 1

L'espace numérique est un champ de conflictualité, où la maîtrise des données est un enjeu dont l'importance ne fait que croître. En effet, contrairement à une vision trop répandue, la cyberdéfense ne s'intéresse pas uniquement à la protection des systèmes d'information interconnectés (le contenant), mais s'intéresse tout autant à celle des contenus informationnels, dont les données forment la matière première. Mais si la maîtrise des données devient ainsi un enjeu central de la cyberdéfense, elle est rendue presque illusoire, notamment du fait de la « datafication » du monde, de la révolution numérique à l'œuvre et de nouvelles pratiques sociales affaiblissant la capacité de contrôle des États sur la production et la circulation des données. Dans cet article, nous suggérons d'approfondir deux pistes de réflexion : la première vise à mieux protéger les stocks de données qui sont en possession des institutions civiles et militaires, et la seconde à responsabiliser les acteurs dans la génération des flux de données.

L'espace numérique est un champ de tensions et de conflits. Les rivalités économiques, politiques, culturelles… s'y déploient, comme elles le font dans les autres champs de la vie sociale. Pour les militaires, l'espace numérique est un champ immatériel, dans lequel les armées doivent être en mesure de se défendre et de s'affirmer (Douzet et Géry, 2020). Tel est l'objet de la cyberdéfense, devenue un domaine prioritaire au regard des investissements et des recrutements opérés par le ministère des Armées depuis plusieurs années (Taillat *et al*., 2018).

Mais, dans ce cadre général de la conflictualité numérique, quel est l'enjeu des données pour la cyberdéfense ?

Les données, un enjeu pour la cyberdéfense

Il serait excessif de dire que la maîtrise des données a été le point de départ du processus qui a abouti à faire de la cyberdéfense un élément prioritaire de la Défense française. La préoccupation initiale était plus la protection des systèmes d'information (le contenant) que de l'information ou des données elles-mêmes (le contenu). Mais il est vite apparu qu'il était impossible de traiter du contenant sans s'inquiéter du contenu, ce qui a conduit à faire des données un enjeu essentiel de la cyberdéfense.

Les données et les différentes couches de l'espace numérique

Parmi les nombreuses missions dont est chargée la cyberdéfense, certaines relèvent de la protection des systèmes d'information, dont l'interconnexion généralisée fait naître la crainte d'une attaque aux effets dévastateurs, le Pearl Harbor numérique. L'enjeu est alors principalement technique : comment empêcher un assaillant de se servir d'une brèche dans l'architecture ou les systèmes d'exploitation du réseau pour y pénétrer et le paralyser ou le saboter ? Dans ce cas de figure, les données contenues dans le système à défendre forment un objet secondaire. L'attaquant a pour objectif de prendre le contrôle d'un système d'information pour s'en servir à son profit (attaque subie par TV5 Monde) ou en faire un outil d'espionnage qui se retourne ainsi contre ses utilisateurs (le virus informatique Flame visant, notamment, certains responsables iraniens). L'attaquant peut encore vouloir le paralyser (attaques lancées en 2007 contre l'Estonie) ou le détruire (Stuxnet employé contre les centrifugeuses iraniennes ; ou KillDisk contre les centrales électriques ukrainiennes à la fin 2015). Du point de vue de la cyberdéfense, la question centrale est bien de protéger l'intégrité du système d'information et de maintenir la connexion entre elles des différentes composantes du système : le défenseur doit empêcher l'intrusion ou, si elle se produit, la détecter, la cantonner et l'éradiquer. Cette dimension conserve toute son importance et son actualité est constamment renouvelée. C'est ainsi que l'invasion de l'Ukraine par la Russie a été précédée d'une attaque visant à détruire les modems utilisés par l'armée ukrainienne pour traiter les signaux passant par le satellite KA-SAT de la société américaine Viasat.

Mais si elles occupent une place relativement secondaire dans ce type d'attaques, les données forment au contraire l'objet principal d'actions qui visent à leur captation, à leur indisponibilité ou à leur destruction. Pour les actions d'espionnage qui se déploient dans l'espace numérique, le but est de s'introduire dans un système d'information pour y dérober des données qui seront ensuite exploitées à des fins diverses : connaître les projets d'un compétiteur, diffuser les éléments qui

le placeront en situation délicate (publication de *mails* confidentiels comme cela a été le cas à la suite du piratage de Sony en 2014 ou du Parti démocrate en 2016) ou encore rendre indisponibles en les chiffrant les données dont le détenteur a besoin pour conduire son activité (Crypto Locker, Petya, Wannacry). Les attaques cyber contre les données peuvent également avoir pour but la destruction pure et simple de ces données. Dans le cadre du conflit ukrainien, plusieurs attaques par "wipers" ont été recensées (HermeticWiper, IsaasWiter, CaddyWiper), la raison d'être de ces logiciels malveillants étant de provoquer l'effacement irrémédiable des données figurant sur les disques qu'ils parviennent à atteindre.

Enfin, les données sont la matière première principale des manœuvres informationnelles qui relèvent de la cyberdéfense. L'accès aux données de terrain est essentiel pour contredire les narratifs adverses ou affirmer les siens. Il a été ainsi relevé que les autorités américaines avaient innové en diffusant largement les données obtenues par les services de renseignement quant à l'accumulation de moyens militaires russes aux frontières de l'Ukraine, établissant ainsi la volonté préméditée de la Russie d'envahir un État souverain aux frontières internationalement reconnues et faisant d'elle, par là même, l'agresseur aux yeux de l'opinion internationale. De même, l'exploitation d'images satellites produites par des compagnies privées et librement accessibles a permis de lever le doute sur la responsabilité de l'armée russe dans les massacres de Boutcha. Il en va de même pour le recensement de fosses communes susceptibles de révéler des crimes de guerre.

Il apparaît donc nettement que la maîtrise des données est, au même titre que la protection des systèmes d'information interconnectés, une préoccupation essentielle de la cyberdéfense.

La maîtrise des données mise au défi de la massification des techniques de communication et des pratiques sociales

La multiplication à l'infini des objets qui captent ou génèrent des données et permettent leur circulation bouleverse les modes de gouvernement des sociétés avec l'émergence de « nouvelles formes d'expression du pouvoir qui se développent par le biais des outils numériques », ce que l'on a pu qualifier fort justement de processus de « datafication généralisée » (Cattaruzza, 2019). Parmi ces outils, les objets communicants, qui prolifèrent dans le domaine civil comme militaire, prennent une place de plus en plus importante et conduisent à la massification de la production de données de tous types. Cette massification soulève en premier lieu des questions intrinsèques de sécurité, c'est-à-dire que le prix modique de nombre de ces objets communicants ne permet pas de couvrir le coût de conception et de mise en place de véritables mesures de sécurité, ce qui laisse ces objets à la merci du premier cyber criminel venu. Surtout, cette massification interroge quant à la maîtrise des données (Danet et Desforges, 2021). En effet, si certaines d'entre elles peuvent ne pas être sensibles en apparence (contenu d'un réfrigérateur, kilométrage et heure d'utilisation d'une trottinette, etc.), d'autres le sont à l'évidence beaucoup plus (données de santé, prestations sociales, sécurité du domicile, par exemple). Ces données sont souvent générées sans même que nous n'en ayons conscience. Or, elles peuvent dévoiler des caractéristiques, voire des failles individuelles ou collectives, offrant ainsi d'énormes opportunités de ciblage à des fins commerciales ou d'espionnage visant des personnes ou des organisations civiles ou militaires.

La « datafication généralisée » doit beaucoup aux avancées de la technologie qui en sont la condition nécessaire. Mais elle n'en doit pas moins aux comportements individuels et aux pratiques sociales qui en forment le véritable moteur. Rien n'oblige véritablement le propriétaire d'une maison à l'équiper d'une sonnette connectée qui va envoyer vingt-quatre heures sur vingt-quatre et sept jours sur sept les images prises par cet équipement connecté sur un *cloud,* où l'opérateur du service, qui s'est réservé le droit d'un accès illimité à ces images, pourra les exploiter à son gré avec tous les moyens de l'intelligence artificielle dont il est l'un des spécialistes mondiaux.

Les conséquences de la « datafication généralisée » ne sauraient mieux être illustrées dans le monde militaire par le cas Strava. Au début de l'année 2018, un étudiant australien, Nathan Ruser, révèle qu'il a découvert l'existence de bases militaires secrètes en Afghanistan, en Syrie et au Niger. Il n'a eu besoin pour ce faire que d'exploiter les données librement disponibles sur Strava, une application permettant d'enregistrer des activités sportives à partir de montres connectées. En croisant le profil des usagers de l'application et les parcours enregistrés par ces derniers, il n'était guère difficile de donner du sens à la concentration de militaires faisant du sport dans des régions en guerre ou dans des endroits où il n'y aurait apparemment aucune infrastructure répertoriée.

Pour les forces armées comme pour les administrations civiles, un véritable changement de paradigme s'est donc produit avec la révolution numérique. La production et la circulation des données échappent désormais très largement à toute forme de contrôle ou de coercition, et la tendance ira en s'accroissant sous l'impulsion de la massification des objets communicants ainsi que de la participation de plus en plus large des individus à toutes les formes de communautés numériques. La masse des données exploitables, sensibles ou non, a donc vocation à se multiplier et à offrir à ceux qui sauront exploiter ces données la possibilité de prendre l'ascendant sur des individus précisément ciblés ou des groupes plus ou moins nombreux, voire des collectivités nationales entières.

Intégrer l'enjeu de la maîtrise des données dans les opérations militaires

L'émergence de l'espace numérique comme champ de conflits a conduit à concevoir et à mettre en place des actions spécifiques, complémentaires des opérations militaires plus conventionnelles. La guerre résultant

de l'invasion par la Russie de l'Ukraine joue comme un révélateur des enjeux actuels de la maîtrise des données dans une situation d'affrontement militaire ; il laisse entrevoir ce que ces enjeux pourraient devenir dans un avenir proche, rendant plus nécessaire que jamais une stratégie de formation à l'usage des outils numériques.

De l'Ukraine à la « guerre cognitive »

Les premiers enseignements de la guerre d'Ukraine ont montré toute l'importance de la maîtrise des données pour des belligérants qui cherchent tous à l'emporter dans le champ informationnel. Mais cette maîtrise pourrait revêtir un caractère encore plus décisif avec le développement du concept chinois de « guerre cognitive » ("intelligentized warfare"). De manière synthétique, ce concept renvoie à l'idée d'agir sur le processus de décision d'individus précisément ciblés (une autorité politique, des groupes appartenant à des forces spéciales…) ou, au contraire, de groupes plus larges de la population, afin de les amener à agir dans le sens des intérêts de leur adversaire (Takagi, 2022). Il s'agit donc moins d'une révolution dans la conduite des affaires militaires que du prolongement et de l'amplification des manœuvres d'influence traditionnelles. Le point intéressant concernant les données est que la clé technologique de la guerre cognitive est l'intelligence artificielle et que celle-ci ne saurait être efficace si elle ne disposait pas de très grands volumes de données sur les cibles visées, notamment des données personnelles de toute nature. Il en résulte une inquiétude au regard de certaines attaques cyber qui ont permis à leurs auteurs de s'emparer de très gros fichiers contenant ce type de données : la double attaque visant l'Office of Personal Management américain à l'été 2014, puis durant l'hiver 2015 ; le piratage des fichiers clients des chaînes d'hôtels Marriott et Hilton ou de ceux de la Marine américaine en 2016.

Les mesures à prendre

La maîtrise des données est donc un enjeu central pour la cyberdéfense. Quelles mesures celle-ci peut-elle s'efforcer de mettre en œuvre pour y répondre ? Deux axes d'effort nous semblent possibles : d'une part, protéger les stocks de données, notamment des données sensibles, que l'institution militaire doit pouvoir mobiliser pour remplir ses missions, et, d'autre part, endiguer les processus indésirables de production des données lorsque la cyberdéfense est en mesure de le faire.

Le renforcement de la protection des stocks de données

Toute institution, civile comme militaire, produit des données qu'elle structure afin de répondre à ses besoins. De cette banale réalité découle pour l'institution concernée une responsabilité particulière en ce qui concerne la protection de ses données contre toute forme d'altération, de captation ou de destruction, ce qui suppose *a minima* le contrôle de l'accès auxdites données. N'importe qui ne doit pas pouvoir accéder à n'importe quelle donnée stockée par une administration ou une entreprise, que cette donnée soit sensible ou non. Quatre questions relatives à cette protection nous semblent utiles à prendre en considération, non seulement pour la cyberdéfense mais aussi pour l'ensemble des acteurs concernés :

- La légitimité de l'ouverture des données publiques : le principe de l'*open data* offre aux acteurs les plus puissants, très souvent étrangers, un accès sans restriction à des données qui, pour être publiques, n'en sont pas moins susceptibles de fournir un atout dans la compétition à ceux qui disposent déjà d'avantages évidents en termes technologiques, commerciaux, industriels…
- La prévention des fuites accidentelles : la protection des données par les institutions qui en sont responsables suppose des mesures adaptées et proportionnées qu'il ne serait pas pertinent de laisser à leur seule libre appréciation. à l'image des règles instaurées en matière de données de santé ; des règles *ad hoc* pourraient être mises en place dans d'autres domaines ou pour d'autres acteurs considérés jusqu'ici comme moins sensibles.
- La lutte contre les pratiques prédatrices : le développement rapide des techniques et des outils du *Big Data* pousse certains prestataires à se spécialiser dans la collecte, la structuration et la revente des données. Certaines pratiques jusqu'ici légales interrogent sur leur légitimité ; une interdiction ou, à tout le moins, un contrôle de celles-ci pourraient être envisagés. C'est ainsi que des entreprises comme Lexis Nexis commercialisent des fichiers de données portant sur des militaires américains (Sherman, 2021). La menace représentée par ce type de produits apparaît assez évidente.
- La prévention de la dépendance à l'égard de prestataires étrangers : le domaine de la collecte et du traitement des données est largement dominé par les acteurs américains du numérique : Amazon ou Microsoft pour le stockage dans le *Cloud*, Palantir pour l'analyse… Certes, ces acteurs sont les plus avancés techniquement et sont les seuls en mesure d'offrir des services de qualité à des coûts très intéressants. Mais sans les accuser de n'être que le faux nez de l'administration américaine, leur positionnement stratégique et la relation qui peut être établie avec eux obligent à tenir compte du risque que leur nationalité représente.

La maîtrise de la génération des flux de données

La production et la circulation des flux de données semblent impossibles à endiguer tant les individus disposent aujourd'hui d'un large éventail de moyens techniques pour ce faire et tant la norme sociale les pousse à s'y adonner. L'empreinte numérique de chacun de nous s'étend constamment et, surtout, elle ne s'efface pas. S'agissant de l'institution militaire, il ne saurait être question pour elle de limiter ce risque en interdisant purement et simplement à ses personnels de participer à la vie numérique de la Nation. Il y aurait là une atteinte inadmissible aux libertés individuelles de plusieurs centaines de milliers d'agents publics. En outre, cette interdiction serait inopérante si elle ne s'appliquait pas également aux conjoints, aux enfants et, *in fine*, aux familles et proches de ces militaires, ce qui reviendrait à vouloir interdire tout accès aux réseaux sociaux à plusieurs millions de citoyens. Il faut donc

compter sur la responsabilité personnelle et collective des personnels militaires et de leur famille pour qu'ils adoptent une conduite responsable dans l'univers numérique. Un *post* ou un *tweet* pouvant mettre en péril une mission, voire la vie des hommes qui la mènent, il ne serait pas exagéré de former les militaires et leurs familles à un usage raisonné des outils numériques.

Conclusion

Après s'être longtemps concentrée sur les risques associés au sabotage des infrastructures et au piratage des logiciels ou des applications, la cyberdéfense a pris toute la mesure de l'enjeu que représente la maîtrise des données pour la conception et la conduite des opérations militaires. La production, la collecte, le stockage, l'exploitation et la diffusion des données forment dès aujourd'hui les facteurs clés du succès des organisations militaires fonctionnant en réseau. Les données seront demain au cœur du combat collaboratif ou des opérations d'influence, voire de la « guerre cognitive ». La guerre en Ukraine a montré à la fois l'importance cruciale de l'enjeu que représentent les données mais aussi la difficulté à les maîtriser, notamment au regard de la multiplicité des parties prenantes et du contrôle limité que les autorités étatiques peuvent exercer sur la production et la circulation des données.

Une politique de maîtrise des données est donc indispensable pour renforcer la résilience de la Nation face à la menace de conflits armés. Elle est en même temps très complexe à définir tant les leviers habituels de la puissance publique ont peu de prise sur l'objet qu'il s'agit de réguler.

Bibliographie

CATTARUZZA A. (2019), *Géopolitique des données numériques*, Paris, Le Cavalier Bleu.

DANET D. & DESFORGES A. (2021), « Des objets connectés aux objets communicants. Les enjeux de souveraineté des objets communicants », *Enjeux numériques – Annales des Mines*, n°16, décembre, pp. 81-85.

DOUZET F. & GERY A. (2020), « Le cyberespace, ça sert, d'abord, à faire la guerre. Prolifération, sécurité et stabilité du cyberespace », *Hérodote*, n°177-178, pp. 329-349.

SHERMAN J. (2021), "Data Brokers Are Advertising Data on US Military Personal", *Lawfare*, August 23, https://www.lawfareblog.com/data-brokers-are-advertising-data-us-military-personnel

TAILLAT S., CATTARUZZA A. & DANET D. (2018), *Cyberdéfense. Politique de l'espace numérique*, Paris, Dunod / Armand Colin, Collection U.

TAKAGI K. (2022), *New Tech, New Concepts: China's Plans for AI and Cognitive Warfare*, War on the Rocks, April 13, https://warontherocks.com/2022/04/new-tech-new-concepts-chinas-plans-for-ai-and-cognitive-warfare/

La responsabilité au cœur de la protection des données : ce que les données disent de l'être humain

Par le Dr Laure TABOUY, PhD
Neuroscientifique et éthicienne, équipe Éthique et épistémologie, CESP-INSERM U1018, Espace éthique APHP, Université de Paris-Saclay

L'accélération des innovations rend indispensable une réflexion sur les enjeux sociétaux, éthiques et juridiques liés à l'exploitation des données, en particulier sur la notion de responsabilité. La conception de garde-fous interdisciplinaires et de systèmes d'évaluation et de suivi, ainsi que la définition d'une gouvernance adaptée aux valeurs sociologiques, éthiques et juridiques des différents pays émergent actuellement dans le monde entier. C'est autour de la nécessité de s'accorder sur la notion de responsabilité sociale que, par exemple, se construit la neuroéthique appelée de ses vœux par le Conseil de l'OCDE à travers sa recommandation n°0457 de 2019 sur l'innovation responsable dans les neurotechnologies. Dans la réflexion sur la notion de responsabilité, la philosophie peut apporter un éclairage non négligeable sur cette question. C'est donc en convoquant Hans Jonas et Hannah Arendt, mais également en utilisant l'éthique de la recherche et la neuroéthique ainsi que les lois et les recommandations existantes, que ce travail autour de la responsabilité sociale concernant les données s'est dessiné.

L'éthique pour parler des données, quel vaste programme…

Aborder la question de la protection des données à la lumière de l'éthique de la recherche, de l'intégrité scientifique, de l'épistémologie et de la neuroéthique n'est pas chose aisée.

Cette dernière contribution qui clôt ce numéro est une invitation à se questionner sur ce que les données disent de l'homme.

Qu'est-ce que cela change dans notre compréhension de l'humain ? Quels sont les enjeux et les limites de l'exploitation de ces données ? Quelles sont les questions et les craintes qu'elle pourrait susciter pour les futures générations ? Comment concevons-nous nos responsabilités individuelles et collectives face aux enjeux éthiques, économiques, politiques, sociétaux et scientifiques que recouvre l'exploitation des données ?

Paul Ricoeur parle de l'éthique comme étant « le mouvement même de la liberté qui cherche une vie bonne, dans la sollicitude envers autrui et dans un juste usage des institutions sociales. »

Réfléchir sur les enjeux éthiques nécessite de s'interroger, de s'ouvrir à la démarche de questionnement sur les valeurs et les finalités de nos actions, les principes et les normes qui les encadrent, les contextes et les conséquences de celles-ci, à des moments où nos actions sont ambiguës, en tension, voire en conflit avec notre environnement.

Une telle réflexion est essentielle à l'innovation responsable qui requiert un temps de réflexion pour que sa concrétisation dans la société soit sûre et respectueuse des valeurs universelles des droits de l'homme, des principes de dignité, de bienfaisance, d'autonomie, de responsabilité, de liberté de penser et de choix, et ce en prenant appui sur l'engagement volontaire des utilisateurs pour en assurer sa réussite.

Les données issues de la finance, des assurances, de la recherche toutes disciplines confondues, de la santé, de l'agroalimentaire, des métiers de l'information et du journalisme, pour ne citer que ces domaines, mais aussi tous les contenus sur Internet et les réseaux sociaux, et informations relatives à l'identité des citoyens et à nos déplacements sont toutes des données précieuses et qui racontent beaucoup de choses sur ce que nous sommes. Le revers de la médaille est la surveillance de la vie des individus, ce qu'ils font et ce qu'ils sont ; il est donc important de garder l'humain au centre des préoccupations que suscite l'exploitation des données. L'urgence est certes de la réguler juridiquement, et c'est la raison d'être de l'action de la Commission européenne (au travers notamment du RGDP), de l'OCDE, de la CNIL et de l'AMF, mais aussi de comprendre ce que signifient ces données pour chacun de nous.

Dans un monde hyper-connecté, comment concilier les enjeux éthiques et épistémologiques que recouvrent les données avec l'innovation et l'accélération des projets d'entreprise et de recherche dans le numérique ?

Car certes, favoriser l'usage, le partage et la réutilisation des données, s'interroger sur les modèles économiques de valorisation de celles-ci, faire confiance à l'innovation sont, en 2022, des enjeux cruciaux, dont l'attractivité de la France dépend. Certes, l'intérêt suscité par les données et les investissements qui y sont consacrés a conduit à la naissance de nombreuses entreprises et à de nombreux projets de recherche français, européens et internationaux visant à faire progresser la connaissance. Mais cette accélération des innovations rend indispensable les réflexions sur les enjeux sociétaux, éthiques et légaux qu'elles soulèvent, ainsi que sur la conception interdisciplinaire de certifications, de normes, d'autorisations, de systèmes d'évaluation, de surveillance et de cadres de gouvernance des données qui soient adaptés aux valeurs sociologiques, éthiques et juridiques de la France et de l'Europe.

Des protections appropriées des données, quelles que soient leur origine et leur catégorie – données de la recherche ou propriété des entreprises ; données cérébrales, identitaires, comportementales, financières, de santé ou métadonnées de contenus –, doivent être mises en œuvre avec pour finalité de nous aider dans notre compréhension des droits de l'homme qui s'y attachent.

Car les données modifient nos conceptions philosophiques et éthiques traditionnelles en apportant des informations sur le fondement biologique de notre comportement moral, nos pensées, nos décisions, nos goûts, nos positions, nos fragilités et nos prédispositions.

La captation des données est une mine d'or pour certaines entreprises, pour leurs commerciaux, mais aussi pour les banquiers, les assureurs soucieux de ne pas perdre d'argent ; ces derniers vont-ils vous assurer en fonction de vos prédispositions génétiques ou de vos risques de développer telle ou telle pathologie ? Les questions de discrimination des personnes à travers leur profil génétique sont déjà très actuelles dans certains pays, mettant en évidence des situations de possible vulnérabilité : fichages policiers, fichiers d'empreintes génétiques et de microbiotes, reconnaissance faciale, recherches généalogiques… Nous sommes tous fichés pour contribuer à un monde qui se veut meilleur pour la santé, la recherche et notre sécurité.

Mais à quel prix ? Les usages détournés de l'utilisation de l'ADN, hors contexte médical, sont une intrusion dans la vie des personnes pouvant entraîner des confrontations à des risques de discrimination, de contrôle social, de manipulation... Il se peut également que des personnes mal attentionnées se saisissent sans votre consentement de vos données pour les vendre sur le *dark Web* ? Qu'en feront-ils, me direz-vous ? Peut-être s'en serviront-ils pour s'approprier votre identité ou vous manipuler, pour prendre le contrôle sur votre personne, pour exercer sur vous des chantages, vous excroquer ?

Un des exemples d'utilisation des données est celui du *nudge*. Ce « coup de pouce » est un concept issu des sciences du comportement ; il est utilisé pour inciter un individu ou un groupe de personnes à changer de comportement, pour les influencer au regard de leurs motivations ou de leurs prises de décisions, sans les mettre sous contrainte ni les soumettre à une obligation, et sans que cela n'implique aucune sanction. Les données récoltées peuvent être utilisées pour mesurer les émotions, l'anxiété, l'attention, la motivation, la vigilance, sans qu'il y ait de véritable consentement éclairé de l'usager. Il ne s'agit plus de convaincre, mais de comprendre, par exemple, les ressorts d'une décision d'achat…

Quelle place donnent les sociétés contemporaines à la vulnérabilité humaine face aux technologies numériques ? L'être humain doit être au centre des enjeux de la protection des données. L'accompagnement des entreprises, des organismes de recherche, des chercheurs, mais aussi des citoyens est un enjeu crucial.

La vulnérabilité nous permet d'entrer en relation avec autrui. C'est une leçon d'humilité, la comprendre permet de prendre soin les uns des autres, dans une juste interdépendance, en assumant une autonomie relationnée. L'important serait d'accepter notre vulnérabilité d'être humain au sein d'une société appelant à une constante « augmentation » de nous-même, ce qui n'est pas chose simple en 2022.

Ce qui est donc en jeu, c'est la confidentialité de ce que nous sommes en tant qu'être humain singulier. La collecte des données interroge sur la notion de consentement et sur les finalités des recherches, mais également sur la vulnérabilité, l'intérêt et la sécurité de la personne et de la nation, qui doivent être au cœur de ces questionnements éthiques.

Pour une recherche et une science fiables

L'éthique en matière de données, par construction, interroge nos actes, nos décisions, nos intuitions, nos choix de recherche, de société, et s'impose dans le champ des innovations responsables. Une recherche responsable, intègre et ouverte est une recherche qui a pour but d'établir des connaissances et de générer des données honnêtes, démontrées et reproductibles.

La science et la recherche, en France et en Europe, sont des espaces de discussion, et pour des raisons multiples et de conflictualité, elles s'exercent dans le cadre des lois, de déclarations, de recommandations et de codes de bonne conduite. Elles sont constitutivement plurielles et conflictuelles par essence ([1] et [2]). Les acteurs de la recherche s'organisent en communautés disciplinaires et écrivent de belles histoires rythmées par la conflictualité, la critique, et qui sont empreintes de la pluralité de la recherche, mais tout en créant un fonds commun, qui peut s'identifier par des paliers d'adhésion ou de consensus.

L'appel à une fiabilité de la science vient de l'idée que celle-ci doit être digne de confiance ([1], [3], [4], et [5]). C'est-à-dire qu'elle doit se situer dans un rapport dynamique et se construire sur des valeurs épistémiques et non épistémiques, autour des idées :

• de robustesse, laquelle ouvre sur le chemin du pluralisme de la recherche à travers la reproductibilité des données, mais passe également par le respect des normes, des lois, des protocoles, des techniques et des méthodes. Ces valeurs épistémiques sont des normes de validité internes au monde de la recherche.
• de la pertinence d'une approche responsable vis-à-vis de la société. Ces valeurs non épistémiques sont des normes de validité externes au monde de la recherche. Cela correspond à la prise en considération par la recherche de la temporalité, des attentes de la société. Les conséquences des recherches et la mise à disposition des données dans des contextes précis seront fonction des cultures sociétales de chaque pays.

Comment pouvons-nous définir la notion de responsabilité à la lumière de la philosophie ?

Parler des enjeux éthiques liés aux données nécessite de prendre un peu de temps pour comprendre ce que signifie « être responsable ». Le détour par la philosophie nous apprend à prendre de la hauteur, à regarder les données et les enjeux qu'elles soulèvent avec un peu plus de recul. Il nous est dès lors possible de définir l'éthique de la recherche, l'intégrité scientifique et la responsabilité sociale des sciences à la lumière de la philosophie.

La responsabilité signifie être responsable de ses actes. C'est être en mesure d'en revendiquer la paternité. L'approche juridique du mot « responsabilité » évoque l'idée d'obligation et de devoir.

Ce mot vient du latin *Respondere*, qui veut dire « se porter garant, répondre de ses intentions et de ses actes devant autrui ou soi-même ». Ces actes sont fonction du rôle joué par une personne, sont des charges qu'elle doit assumer et dont elle doit en supporter toutes les conséquences. Ce mot est également apparenté à celui de *Sponsio*, qui veut dire « promesse », conférant au mot « responsabilité » l'idée de devoir assumer ses promesses.

C'est donc l'obligation de répondre du passé et la volonté ou, à tout le moins, la possibilité d'assumer les conséquences futures d'un acte, une intention que l'on exprime, une décision que l'on prend, dans lesquelles la personne s'engage au nom de certaines valeurs qu'elle partage et en vue d'une finalité déterminée qu'elle s'est fixée.

La responsabilité est donc l'expression d'une volonté, un engagement, mais aussi un devoir, une obligation de répondre, de rendre des comptes. Elle suppose enfin de disposer d'un pouvoir, d'une liberté.

Dans le « Principe responsabilité »[1] d'Hans Jonas, publié en allemand en 1979, la responsabilité émerge comme un principe moral avec toutes ses caractéristiques.

Dans son ouvrage, Jonas se demande dans quel sens la technologie et la technique moderne sont un questionnement pour l'éthique. Pour lui, du bien peut émerger le mal, l'homme a désormais les capacités de s'autodétruire en peu de temps à travers le perfectionnement de la technologie.

Cette définition introduit le concept de responsabilité des générations actuelles vis-à-vis des générations futures. Y est mis l'accent sur la nécessité d'aller jusqu'à interdire les technologies si la préservation des générations futures est compromise ; ce n'est plus une responsabilité tournée vers le passé mais vers l'avenir de l'humanité et de notre planète, la Terre. Ce concept de responsabilité est à la base des principes de développement durable et de précaution aujourd'hui en vogue.

Quelle décision prendre dont je puisse répondre et donc être responsable, compte tenu des lois, des règles et des cadres relatifs aux données ? Comment prendre une décision ou apprécier ma responsabilité quant à l'usage, l'utilisation, la collecte des données ? Qui a le pouvoir et la responsabilité de décider ce qu'elles signifient et comment les protéger ?

Ces questions expriment la dimension politique du mot « responsabilité ».

Cela m'amène à convoquer Hannah Arrendt pour les définitions qu'elle a proposées, lesquelles ont une portée plus politique de la responsabilité collective et individuelle, de la liberté et des visions qu'elle a de la technologie. Dans « Responsabilité et jugement »[2] et « Condition de l'homme moderne »[3], deux ouvrages parus en 1958, elle dénonce la perte de sens liée aux évolutions en particulier techniques, comme l'automatisation du travail. Pour elle, la responsabilité trouve tout son sens dans une démarche politique. C'est une mutation du concept de responsabilité sociale.

En résumé, l'approche juridique que recouvre le « être responsable » consiste à assumer les conséquences de ses actes. L'approche éthique du mot « responsabilité » nécessite une mise en débat au niveau des parties prenantes. L'éthique de la recherche devient alors le pivot réflexif entre la responsabilité sociale et l'intégrité scientifique. Cette approche tient compte des contextes et des conséquences, en fonction des principes et valeurs qu'elle sous-tend et des finalités des actions et des décisions.

C'est la sagesse pratique de Paul Ricoeur, un équilibre réfléchi entre conviction/désirs et morale/obligation.

---

[1] JONAS Hans (1998), *Le principe responsabilité, une éthique pour la civilisation technologique*, Paris, Flammarion, collections « Champs », 450 pages.

[2] ARENDT Hannah (2005), *Responsabilité et jugement,* collection « Petite Bibliothèque Payot », Éditions Payot, 368 pages.

[3] ARENDT Hannah (1958), *Condition de l'homme moderne*, Paris, Calmann-Lévy.

Les infrastructures de données : l'exemple des neurosciences

La qualité des données est le nerf de la guerre pour les entreprises et la recherche aussi bien académique que privée. En outre, le volume de ces données est aujourd'hui colossal ; leur combinaison peut donner lieu à des données devenant sensibles.

Afin de traiter, restituer, stocker et (ré)utiliser les données des *clouds*, des entrepôts et des *hubs* français et européens ont été créés, répondant par là même à des enjeux stratégiques de souveraineté numérique et d'indépendance.

Ces infrastructures ont pour rôle de favoriser le partage et la circulation des données tout en les protégeant, et ce afin de permettre à l'innovation d'avancer et de prendre toute sa place dans notre vie quotidienne, de permettre aux citoyens et aux chercheurs d'accéder aux données des recherches en cours dans le but de faire avancer la science et la recherche.

Prenons l'exemple des neurosciences et des neurotechnologies : qu'en est-il des données cérébrales hors du contexte médical ? Vont-elles être là aussi utilisées au regard de la relation étroite qui existent entre le cerveau et les capacités cognitives propres à l'identité humaine, notamment la capacité singulière de la personne humaine à agir et à rendre compte ?

Les différents projets internationaux actuellement menés, comme le Human Brain Project[4], l'International Brain Initiative[5] ou le NIH Brain Initiative[6], mettent l'accent sur le développement des neurotechnologies, en particulier celles non invasives qui visent à mieux comprendre et à intervenir sur les fonctions cérébrales. Dans le but de faire progresser les neurosciences et les innovations, est intervenue la création de plateformes nationales et internationales spécifiques aux données cérébrales issues de la recherche. Ces plateformes visent à faciliter les échanges entre les chercheurs et ont permis à de nombreuses collaborations de voir le jour depuis quelques années, comme Health DATA Cloud[7] d'EBrains[8] ou le projet européen Human Brain Project.

Le développement de neurotechnologies non invasives qui sont ainsi créées devient plus complexe, ce qui peut conduire à les considérer au regard de leur statut sur le plan moral, compte tenu de leur capacité à enregistrer et à moduler l'activité cérébrale en intervenant sur le fonctionnement neuronal, mais aussi à modifier les personnalités, les états affectifs, les comportements, l'autonomie, la cognition et les facultés d'action. Elles mettent en jeu la conception que nous avons de nous-mêmes en tant que personne libre et responsable, entraînent des conséquences profondes sur l'identité humaine et bouleversent la compréhension des comportements.

L'utilisation des données issues de l'activité cérébrale des utilisateurs de ces neurotechnologies, sans leur consentement éclairé et sans les avoir informés des effets de celles-ci, soulève un certain nombre de questions concernant la responsabilité ([6] [7] [8] [9] et [10]).

Par exemple, elles pourraient être utilisées pour corriger les erreurs de comportement prédites ou constatées ou permettre de savoir comment une personne réagira dans une situation donnée et évaluer les performances de cette personne dans une équipe. Ces questions pourraient être au cœur des préoccupations de recrutements et de gestion des ressources humaines.

Les données cérébrales et l'encadrement de leur utilisation sont aujourd'hui une vraie préoccupation internationale. C'est l'ultime refuge de la privaticité : le cerveau, les pensées, la liberté d'agir, de décider, de choisir, de penser…

C'est la raison pour laquelle le *brainjacking* soulève des questions au regard des garanties apportées en termes de protection efficace de la personne, notamment lorsque l'on procède à un enregistrement des données cérébrales de celle-ci sans qu'il existe aujourd'hui un statut juridique bien défini pour ces données. Cette pratique pose la question du consentement éclairé, et du cadre normatif en vigueur : obligation ou non d'informer l'utilisateur, existence ou non de garanties en matière de liberté individuelle et collective, de penser, de conscience…

La neuroéthique va nous aider à comprendre les enjeux éthiques des neurotechnologies, lesquelles se placent à l'intersection entre les sciences humaines et les neurosciences et qui invitent à discerner – dans les transformations engendrées par les neurotechnologies – entre ce qui est souhaitable et ce qui l'est moins. La neuroéthique incarne l'approche pluridisciplinaire indispensable à l'essor de toute innovation responsable touchant les neurosciences. Elle exige de nous que l'on comprenne comment les connaissances et les recherches sur les neurotechnologies peuvent affecter le futur de la société. Actuellement, la définition de garanties est en cours de discussion au niveau de l'OCDE[9], autour du concept émergent de neurodroit[10] [11], que ce soit en Europe, au Chili [12] ou aux États-Unis[11].

Les enjeux de la science ouverte

Comment peut-on définir les règles de partage des données ? Comment fédérer les acteurs de la recherche et les entreprises autour des enjeux des données ? Comment valoriser de façon intègre, responsable et éthique les données de la recherche, tout en les ouvrant à la société ? Comment promouvoir l'innovation ? Comment mettre en place une bonne protection

(4) https://www.humanbrainproject.eu/en/

(5) https://www.internationalbraininitiative.org/

(6) https://braininitiative.nih.gov/

(7) https://www.healthdatacloud.eu/

(8) https://ebrains.eu/

(9) https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0457

(10) https://neurorightsfoundation.org/

(11) https://en.unesco.org/courier/2022-1/chile-pioneering-protection-neurorights

des données de santé, en particulier dans le contexte d'une utilisation de plus en plus simple de l'IA ?

Ces questions montrent l'importance de penser les notions de partage des données et de science ouverte, de penser l'éthique de la recherche et l'intégrité scientifique au prisme de la responsabilité individuelle et collective des chercheurs, des industriels, des décideurs et des utilisateurs.

L'ouverture de données de la recherche, *via* le mouvement de la science ouverte[12] et [13], est aujourd'hui devenue incontournable, mais est surtout un véritable défi. Cette notion d'ouverture et d'accès libre renvoie à de multiples problématiques. Parmi les enjeux à relever figure la définition de ces données, car les données de la recherche recouvrent de multiples formes selon les disciplines : quels statuts leur donner ? Quelle gestion et législation pour ces données ? Comment les anonymiser ? Comment les valoriser et les protéger dans un monde d'innovation en profonde mutation ?

La recherche est mondiale ; les données transitent donc de pays en pays, de laboratoires en laboratoires, des laboratoires vers les entreprises. C'est l'idée même des sciences qui « s'ouvrent », ce qui renvoie à des pratiques de mise en culture des sciences, d'une présence des sciences dans un grand nombre d'espaces sociaux et sociétaux. Le monde de la recherche, autant privé qu'académique, et celui des entreprises sont-ils capables de rester autonomes et indépendants, tout en travaillant conjointement pour le bien commun et en interdépendance avec la société, et sont-ils prêts à ouvrir largement leurs données ?

La science ouverte suppose également de repenser les pratiques de recherche ; la mise à disposition en libre accès des données massives et la participation de la société à la recherche bouleversent la production du savoir.

En conclusion

Réfléchir sur les enjeux éthiques, juridiques, sociétaux et économiques des données est une invitation à assumer nos responsabilités face à ces enjeux. L'éthique est fondée sur la pratique ; elle s'apprend, évolue, vise une transformation des actions et peut être évaluée.

Il va donc falloir prendre en considération les contextes et les conséquences de la recherche et interroger les finalités de cette collecte massive de données, ainsi que le rapport entre les sciences et les valeurs importantes à défendre qui seront à identifier. La collecte de données peut refléter une certaine réalité à apprécier au prisme d'autres réalités parfois cachées.

Les données sont des biens communs, elles présentent donc un intérêt collectif. Il est donc nécessaire d'en prendre soin, et cela commence au sein même du monde de la recherche, dès le départ, en amont des protocoles, en s'interrogeant sur le champ des possibles qui s'ouvre, sur le futur de la société ; c'est ce que l'on peut appeler l'éthique de l'anticipation[14].

Hannah Arendt soulignait l'idée que notre liberté nous permettait de donner du sens et une raison à notre action. La vie se vit en commun, en société ; nous avons besoin de l'autre, d'agir vers l'autre pour nous déployer pleinement et prendre part à la vie en société. Nous apportons tout ce que nous sommes en tant que personne à nos actions et c'est notre liberté d'être qui nous permet de nous mettre en action. Mais aller vers l'autre, cela nous bouscule, nous confronte à autrui, nous contraint aussi et, de fait, nous fait réagir.

L'utilisation des données demande de reconnaître ma liberté mais aussi la liberté de l'autre, ce qui nécessite conscience et responsabilité.

Références bibliographiques

[1] COUTELLEC L. (2019), « Penser l'indissociabilité de l'éthique de la recherche, de l'intégrité scientifique et de la responsabilité sociale des sciences », *Revue d'anthropologie des connaissances* 13, n°2, 381 pages.

[2] ISRAEL-JOST C. (2021), « Faire du commun dans les sciences », *Médecine/Sciences* 37, pp. 89-95.

[3] VAN ATTEVELDT N., TIJSMA G., JANSSEN T. & KUPPER F. (2019), "Responsible Research and Innovation as a Novel Approach to Guide Educational Impact of Mind, Brain, and Education Research", *Mind, Brain, Educ*. 13, pp. 279-287.

[4] DOUGLAS H., *Scientific Freedom and Social Responsibility,* Chapter XY.

[5] GARDNER J. & WILLIAMS C. (2015), "Responsible research and innovation: A manifesto for empirical ethics?", *Clin. Ethics* 10, pp. 5-12.

[6] IENCA M. *et al*. (2021), "Towards a Governance Framework for Brain Data", *ArXiv preprint* 2109.11960.

[7] WILLIAMSON B. (2019), "Brain Data: Scanning, Scraping and Sculpting the Plastic Learning Brain Through Neurotechnology", *Postdigital Sci. Educ.* 1, pp. 65-86.

[8] GOERING S. *et al*. (2021), *Recommendations for Responsible Development and Application of Neurotechnologies*.

[9] RAINEY S., MARTIN S., CHRISTEN A., MÉGEVAND P. & FOURNERET E. (2020), "Brain Recording, Mind-Reading, and Neurotechnology: Ethical Issues from Consumer Devices to Brain-Based Speech Decoding", *Sci. Eng. Ethics*, doi:10.1007/s11948-020-00218-0.

[10] PFOTENHAUER S. M. *et al*. (2021), "Mobilizing the private sector for responsible innovation in neurotechnology", *Nature Biotechnology,* vol. 39, pp. 661-664.

[11] MARTIN A. *et al*. (2007), "A four-part working bibliography of neuroethics: Part 2 – Neuroscientific studies of morality and ethics", *Philos. Ethics, Humanit. Med.* 10, pp. 21-23.

[12] BUBLITZ J. C. (2022), "Novel Neurorights: From Nonsense to Substance", *Neuroethics* 15, pp. 1-16.

(12) https://www.ouvrirlascience.fr/accueil/

(13) https://www.unesco.org/en/natural-sciences/open-science

(14) COUTELLEC *et al*. (2016), https://doi.org/10.3917/rfeap.002.0014

Propos conclusifs

Par le Dr Laure TABOUY, PhD
Neuroscientifique et éthicienne, équipe Éthique et épistémologie, CESP-INSERM U1018,
Espace éthique APHP, Université de Paris-Saclay

Chers tous, chers contributeurs et chers lecteurs.

Voici quelques lignes de conclusion pour clore ce numéro sur « la protection des données dans une économie globalisée ».

Le projet de ce numéro sur la protection des données est né en août 2021 et se veut une réponse à des questions qui me taraudaient alors : quel monde voulons-nous ? Jusqu'où pouvons-nous et devons-nous aller concernant l'exploitation des données ?

Mon souhait était de traiter de la problématique des données dans les entreprises et au sein du monde de la recherche, en faisant intervenir des personnes qui n'ont pas forcément l'habitude d'être rassemblées pour contribuer à l'élaboration d'un même numéro dédié à la protection des données. Je voulais amener à s'exprimer des personnes qui ont à cœur de travailler dans leur quotidien sur cette thématique si brûlante aujourd'hui.

Présentant moi-même un profil interdisciplinaire, entre le monde de la recherche académique et celui de l'entreprise. Étant à la fois chercheuse et pour une deuxième fois doctorante dans l'équipe de recherche académique Éthique et épistémologie – Espace Éthique de l'Île-de-France, au sein du CESP-INSERM U1018, et chercheuse au sein du Laboratoire interdisciplinaire sur le doctorat (LID), créé chez Adoc Talent Management, une entreprise, j'avais à cœur de faire dialoguer ces deux mondes.

Pari tenu, et j'en suis très heureuse. C'est donc avec une grande fierté et un grand honneur que je voie aboutir mes efforts à travers la parution de ce numéro.

Confiance, fiabilité, responsabilité, éthique, transparence, rigueur, confidentialité, disponibilité, intégrité : voici quelques mots clefs importants pour parler de la protection des données et des enjeux qu'elle recouvre.

Que ce soit les données de la recherche, comme en neurosciences ou en génétique ou, plus largement, toutes les disciplines de recherche confondues, les données relatives à l'identité et à la santé des personnes ou les données venant de la finance, des services de recrutement, de l'accompagnement des projets ou encore celles issues des entreprises, de l'aéronautique, de la cybersécurité, de la consommation, des déplacements, de la Défense, la problématique de leur signification et de leur protection est, en 2022, partout au cœur de notre société.

« Le nerf du métier, c'est la qualité de la donnée », disait Edwin de Barrau dans un article publié sur DogFinance[1]. C'est un avis que je partage réellement, mais je rajouterai à cette phrase que la qualité des données passe par un travail de recherche intègre et responsable.

Ces réflexions sur la qualité, l'intégrité et l'utilisation des données impliquent tous les acteurs de la recherche, qu'elle soit académique ou privée, ainsi que les entreprises. Elles incitent à engager une réflexion en profondeur de la signification même de ces données.

Qu'est-ce que ces données signifient ? Qu'est-ce qu'elles disent de nous en tant qu'être humain ? Comment les protéger et en prendre soin, afin de les utiliser de façon intègre et responsable ? Il est donc intéressant d'envisager l'éthique de la recherche comme pivot réflexif entre l'intégrité scientifique et la responsabilité sociale.

Plusieurs actions, déclarations et recommandations ont participé à la mise en place de la science ouverte à l'échelle nationale et internationale au travers de plateformes et de *hubs*. L'articulation entre science ouverte et intégrité scientifique est centrale pour assurer cette intégrité et baisser la pression en termes de tentatives frauduleuses visant à être publié. C'est un fait, l'ouverture des données de la recherche vient questionner, voir bousculer certaines pratiques, certains fonctionnements erratiques de la recherche.

Rendre la recherche scientifique et les données qu'elle produit intègres et accessibles à tous, et ce à tous les niveaux de la société :

- soulève et met en jeu des enjeux éthiques, sociétaux et juridiques ;
- appelle un changement du regard que nous portons sur tous les processus de la recherche afin d'ouvrir les résultats de celle-ci à la société ;
- ouvre l'opportunité de donner à la recherche toute la place qu'elle mérite dans la société et faire en sorte que les acteurs de la recherche soient pris au sérieux et bénéficient de toute la confiance de la société.

[1] https://www.dogfinance.com/fr/news/zoom-sur-une-equipe-reporting-et-mesure-de-performance

Car améliorer la reproductibilité de la science, en rendant les processus plus transparents, permet d'augmenter la confiance de la société envers les communautés scientifiques.

D'un autre côté, la société comme les entreprises doivent aussi s'ouvrir à la réflexion ; ces dernières doivent faire preuve de plus de transparence et d'intégrité dans l'utilisation des données de leurs clients, de façon à ce que celle-ci soit plus responsable. Elles doivent aussi veiller à ce que ces données soient de meilleure qualité. Tout cela contribuera à augmenter la productivité des entreprises et à accroître la confiance que leur accordent leurs clients.

Il est certain que tous les points de vue n'ont pas pu être exprimés dans ce numéro, ce sont notamment ceux de financiers, de recruteurs ou de personnes travaillant au financement de projets ou encore dans des entrepôts de données de santé. Mais j'espère que les articles qui sont rassemblés dans ce numéro de *Réalités industrielles* permettront à tout un chacun de mieux appréhender à quel point les enjeux de la conservation et de la mise à disposition de l'ensemble des données sont colossaux.

La réutilisation de ces données, leur archivage comme leur protection, et ce quelles que soient les disciplines de recherche, sont au cœur de l'innovation d'aujourd'hui et de demain. Mais cela demande d'annoter les données, de les ranger correctement.

Il va donc falloir prendre en considération les contextes et les conséquences de la recherche et interroger les finalités de cette collecte massive de données, ainsi que le rapport entre les sciences et les valeurs importantes à défendre qui seront à identifier. La collecte de données peut refléter une certaine réalité à apprécier au prisme d'autres réalités parfois cachées.

Les données sont des biens communs, elles présentent donc un intérêt collectif. Il est donc nécessaire d'en prendre soin. C'est à chacun d'entre nous de comprendre que nous sommes tous responsables de nos données et qu'il est important de préserver l'intimité de chacun.

Je souhaiterais, pour clore mon propos, remercier chacun des contributeurs à ce numéro. Je les remercie de m'avoir fait confiance et d'avoir joué le jeu.

J'invite chacun de nous à nourrir sa réflexion à partir du travail de qualité fourni par chacun de ces contributeurs qui ont pris le temps de nous exposer les problématiques que sous-tend l'exploitation des données, et plus particulièrement personnelles.

Merci à tous !

Expériences de bâtiments passifs ou à énergie positive

Par Pascal GONTIER
Professeur à l'École nationale supérieure d'architecture de Nantes et membre titulaire à l'Académie d'architecture

À partir de la fin des années 1990, la prise en compte des économies d'énergie s'invite progressivement dans le débat architectural. Par paliers successifs, les niveaux d'exigence évoluent. Au départ, focalisés sur les consommations de chauffage, ils concernent aujourd'hui l'ensemble des postes de consommation et s'ouvrent à la problématique du carbone.
L'engagement environnemental de mon agence d'architecture, depuis sa création, nous a conduit à accompagner ce mouvement, en devançant les évolutions, lorsque cela est possible, et en allant au-delà des exigences réglementaires ou programmatiques, avec la conviction que les défis énergétiques et environnementaux sont tels qu'ils ne peuvent pas se satisfaire de simples solutions normatives ou techniques. Les enjeux environnementaux sollicitent le monde de l'architecture dans son ensemble. Celui-ci est désormais confronté à de nouveaux défis. Il est en effet invité à aller encore plus loin dans la sobriété énergétique et la décarbonation, tout en rendant nos bâtiments et nos villes résiliants face au changement climatique.

À partir de la fin des années 1990, l'intégration des questions énergétiques et, plus généralement, des questions environnementales dans la construction commençait à peine à sortir de la marginalité. J'ai créé mon agence d'architecture à cette époque, avec la volonté de m'impliquer activement dans la transition écologique alors émergente et la conviction qu'elle ne peut se résumer à de simples dispositions techniques ou normatives architecturalement neutres. La démarche de mon agence, depuis sa création, est ainsi portée par l'idée que les défis environnementaux sont tels qu'ils demandent des réponses innovantes et créatives, et sont ainsi de nature à susciter une renouvellement architectural profond.

Les quatre exemples de programmes immobiliers frugaux présentés ci-après sont quelques-uns des jalons de cette démarche qui est par nature évolutive.

L'opération du passage Fréquel

La construction de 17 logements du passage Fréquel (20e arrondissement de Paris), réalisée pour le bailleur social SIEMP, est la première opération parisienne de construction de logements collectifs passifs. Nous avons réalisé ce bâtiment à la suite d'un concours que nous avons gagné en 2006, avec un projet dont les ambitions énergétiques, basées sur le standard allemand Passivhaus, allaient bien au-delà des prescriptions du programme considéré.

Photo © Pascal Gontier

Le bâtiment passif du passage Fréquel.

Ce projet est situé dans un tissu urbain dense, sur un terrain essentiellement orienté au nord et dont la façade sud-est en vis-à-vis proche avec des bâtiments élevés. Cette situation urbaine peu favorable avait conduit le maître d'ouvrage à demander que le bâtiment se contente de répondre aux prescriptions du label THPE (Très Haute Performance Énergétique) de l'époque.

En proposant de réaliser un bâtiment passif, j'avais souhaité démontrer que l'objectif d'une consommation de chauffage au mètre carré de 15 kW/h était atteignable, et ce quel que soit le contexte urbain. J'avais également souhaité aller à l'encontre de la doxa bioclimatique en vigueur à l'époque en France, qui voulait qu'un bâtiment passif soit nécessairement orienté au sud et qu'il se devait d'être hyper compact, doté de noyaux de distribution aveugles et de fenêtres chichement dimensionnées.

Nous avons donc fractionné le programme en deux bâtiments distincts, et avons ainsi créé une courette « parisienne » afin de ménager les jours de souffrance du bâtiment voisin et d'apporter un maximum de lumière dans les logements et les parties communes. Ce parti pris nous a permis de créer des logements qui sont tous multi-orientés, ainsi que des studios traversants. La cage d'escalier et les paliers d'étage bénéficient de lumière naturelle ainsi que de vues sur l'extérieur. Enfin, les fenêtres, toutes à triple vitrage, sont réparties entre les différentes façades et ont une surface de 30 % supérieure à celle qui était demandée dans le programme.

Photo © Stephan Lucas

Vue intérieure du bâtiment passif du passage Fréquel.

Pour arriver à combiner performances énergétiques, multi-orientation des logements et générosité de l'éclairage naturel, nous avons mis en œuvre l'ensemble de la panoplie technique des bâtiments passifs : forte isolation par l'extérieur, chasse impitoyable aux ponts thermiques, étanchéité maîtrisée, loggias et coursives portées par une structure autonome. Ainsi, l'expression du bâtiment passe plus par ses détails architectoniques que par une gestuelle formelle.

Ce projet nous a permis de montrer que les bâtiments passifs pouvaient s'accommoder de situations urbaines denses et complexes et aussi qu'une ville constituée de bâtiments énergétiquement performants ne ressemblait pas nécessairement à une ville homogène et héliotropique.

Sa réalisation nous a aussi permis de constater que les consommations de ventilation, du fait du double flux, étaient supérieures, en énergie primaire, aux consommations de chauffage. Postérieurement à l'achèvement de ce programme, nous avons continué et nous continuons encore aujourd'hui à réaliser des bâtiments passifs, à l'instar des 41 logements Passivhaus que nous avons livrés à la ville de Gonesse en 2012. Mais nous travaillons également sur des modèles de ventilation alternatifs, qui permettent de réduire les consommations du couple chauffage-ventilation.

Par ailleurs, le bilan carbone que nous avions réalisé dans le cadre du programme Fréquel nous a fait prendre conscience du fait que les émissions de carbone durant la phase de construction représentaient l'équivalent des émissions produites durant soixante années d'exploitation du bâtiment. Les bâtiments conçus par la suite l'ont été avec le souci de minimiser l'énergie grise et de réduire l'impact carbone des matériaux mis en œuvre.

Le bâtiment Duée-Pixérécourt

Le bâtiment Duée-Pixérécourt situé dans le 20e arrondissement de Paris, qui a été réalisé pour la RIVP et livré en 2013, s'inscrit dans un environnement bâti fortement hétérogène. Il se caractérise par une juxtaposition d'événements architecturaux contrastés, construits à différentes époques, et qui présentent une grande diversité d'échelles, de formes et de matériaux. La parcelle se présente sous la forme d'une longue bande mono-orientée qui longe un passage étroit et pentu.

Cette réalisation se compose de trois entités indépendantes, réalisées en structure bois, disposées le long du passage et organisées autour de trois cours ouvertes. Cette configuration, qui reprend un type d'organisation relativement courant dans les rues avoisinantes, permet d'offrir une certaine intimité aux occupants des logements, et surtout de leur assurer un bon niveau de confort lumineux ainsi que des orientations offrant des vues multiples malgré l'étroitesse du passage. L'ensoleillement des logements a par ailleurs été optimisé grâce à des fenêtres à triple vitrage, dont les dimensions sont encore plus généreuses que celles des fenêtres du bâtiment du passage Fréquel.

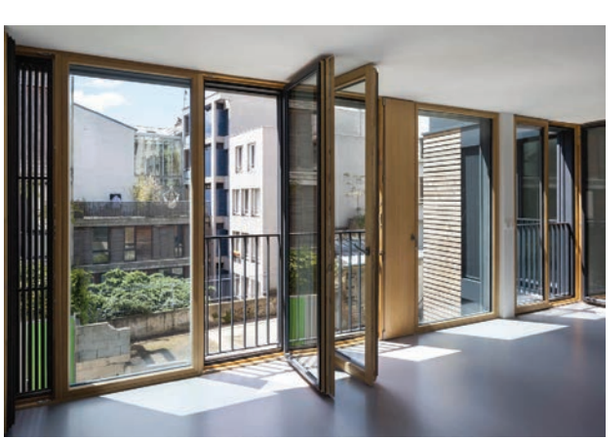


Photo © Hervé Abbadie

Vue intérieure du bâtiment Duée-Pixérécourt.

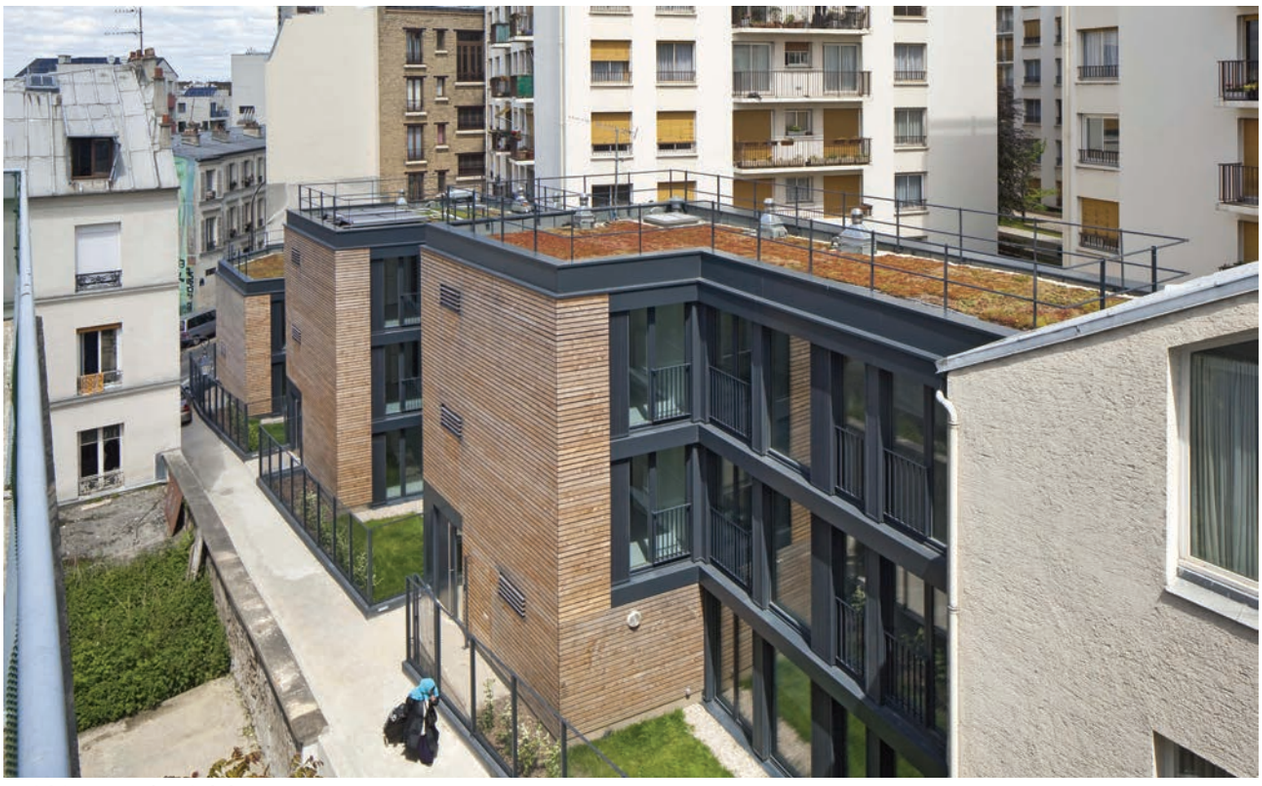


Photo © Hervé Abbadie

Le bâtiment Duée-Pixérécourt.

Pour aller au-delà des ambitions énergétiques du standard Passivhaus, le dispositif de ventilation des logements a été dédoublé. Le bâtiment est en effet doté d'une ventilation mécanique double flux à récupération d'énergie, qui n'est utilisée qu'en période de chauffage. Hors de la saison de chauffe, cette ventilation est coupée et remplacée par une ventilation naturelle qui est assurée par une simple ouverture de volets d'air disposés en façade et dissimulés derrière des grilles persiennes. L'ouverture de ces volets s'effectue manuellement lorsque la ventilation mécanique double flux est arrêtée (un arrêt qui, dans chaque appartement, est indiqué par un voyant lumineux situé dans la cuisine). Des conduits sont disposés dans les pièces humides pour permettre une extraction naturelle de l'air obtenue grâce à une simple ouverture des volets évoqués ci-dessus.

Le dédoublement du dispositif de ventilation a permis de diviser par deux les consommations électriques afférentes et d'assurer le renouvellement de l'air dans les logements durant la moitié de l'année.

Le bâtiment Max Weber

Le bâtiment Max Weber est implanté dans l'enceinte de l'Université de Paris Nanterre, vaste campus dont les différents bâtiments en béton et en métal sont autant de témoignages de l'architecture universitaire française qui a prévalu dans les années 1960.

Le programme établi par l'Université de Paris Nanterre comprenait le regroupement en un même lieu de ses différents laboratoires de recherche en sciences sociales et humaines. À partir de ce programme architectural, assimilable à celui d'un bâtiment de bureaux, la maîtrise d'ouvrage demandait la construction d'un bâtiment de prestige capable de valoriser l'image de ses laboratoires de sciences, de favoriser les échanges entre les chercheurs et de donner à la recherche qu'elle mène une identité forte et une attractivité auprès des chercheurs étrangers.

L'ambition environnementale forte qui a présidé à la conception du bâtiment nous a conduits à réinterroger en profondeur la nature même des espaces de travail offerts aux chercheurs et à proposer des pistes architecturales nouvelles. Les immeubles de bureaux sont en effet trop souvent des produits ultra-standardisés, qui ne parviennent à trouver leur identité que dans une surenchère formelle au niveau des façades et de la décoration. Le bâtiment Max Weber a été conçu de façon à éviter, dans une logique environnementale clairement affichée, une telle banalisation des espaces de travail. Il s'agit en effet d'un bâtiment totalement atypique, entièrement construit en bois, passif et doté d'une ventilation naturelle qui fonctionne en hiver comme en été.

Les faux plafonds et la climatisation ont été bannis des bureaux, ainsi que la ventilation mécanique contrôlée. Contrairement aux usages répandus, la structure en bois visible marque fortement les espaces intérieurs, leur donnant un caractère singulier et chaleureux.

Les couloirs et les cages d'escalier sont éclairés par la lumière naturelle et bénéficient de vues sur l'extérieur.

Les réseaux ont fait l'objet pour leur intégration d'une organisation spatiale spécifique tenant compte de la nature de chacun d'eux. Ainsi, les plafonds en bois

Photo © Hervé Abbadie

Le bâtiment Max Weber situé dans l'enceinte de l'Université de Paris Nanterre.

massif s'interrompent régulièrement pour ménager des cavités destinées à accueillir les différents réseaux électriques et les luminaires.

Si au niveau de son enveloppe il est de type « passif », le bâtiment s'écarte toutefois fortement de ce modèle par son système de ventilation naturelle assistée et contrôlée, ce qui permet d'éviter les consommations habituelles liées à la ventilation mécanique double flux.

Ce dispositif de ventilation, qui a fait l'objet d'études et de recherches très poussées dès la phase de son esquisse, constitue une première pour un immeuble de bureaux en France. Il se manifeste en toiture par vingt-cinq cheminées sculpturales en aluminium de trois mètres quatre-vingts de haut.

Photo © Schnepp Renou

Les cheminées sculpturales installées sur le toit du bâtiment Weber.

Le projet de Pessac

La construction de cinquante-six logements collectifs à Pessac (département de la Gironde) est un projet expérimental initié en 2015. Il a été conçu entièrement à partir d'un cahier des charges précis que nous avons établi afin de répondre à une ambition environnementale particulièrement forte, portée par le bailleur social Aquitanis.

Nous avons ainsi proposé de réaliser un ensemble de bâtiments pouvant compter jusqu'à dix étages, avec une structure entièrement en bois, y compris celle des cages d'escalier et d'ascenseur, utilisant comme isolant la paille et qui est doté d'un dispositif de ventilation naturelle assistée et contrôlée. Il s'agissait d'un programme représentant un objectif ambitieux que nous n'étions pas totalement sûrs de pouvoir atteindre.

Nous avons donc élaboré le projet selon une méthodologie particulière adaptée à ces enjeux. Habituellement, les études de formes et de fonctions des bâtiments sont réalisées avant les études techniques. L'architecte dessine, puis l'ingénieur calcule. Ici, nous avons revu totalement les processus de conception. La forme du projet a résulté d'une conception paramétrique comprenant des études multi-critères de la volumétrie des bâtiments, de leur structure, de l'organisation spatiale et des systèmes aérauliques.

Nous avons ainsi élaboré plusieurs propositions volumétriques à partir desquelles différentes simulations ont été réalisées dans le but de tester leur capacité à répondre aux exigences du cahier des charges que nous nous étions fixé.

Photo © Artefactory

L'ensemble de bâtiments de logements collectifs envisagé à Pessac.

Il est vite apparu que pour des raisons d'insertion du projet dans le site, la meilleure solution consistait à créer un ensemble composé de deux bâtiments de hauteurs différentes : quatre niveaux pour le premier et dix niveaux pour le second.

À partir de ce principe volumétrique, nous avons réalisé différentes versions du projet que nous avons comparées afin d'être en mesure de répondre à deux questions :

• Est-il possible d'assurer un renouvellement de l'air au moyen d'une ventilation naturelle assistée qui soit capable de répondre aux besoins des logements en la matière sur la totalité de l'année, avec un ensemble bâti comportant un épannelage, dont le niveau le plus bas correspond à R + 5 et le niveau le plus haut à R + 9 ?
• Existe-t-il une volumétrie permettant d'optimiser cette ventilation et d'éviter le recours à l'assistance mécanique, et ce quelle que soit la saison et quelles que soient la direction et la force du vent ?

Une version du projet s'est révélée plus performante que les autres, tant du point de vue de l'insertion urbaine que de celui de l'efficacité de la ventilation. Dans cette version, les deux bâtiments étaient de forme hexagonale et comportaient des toitures plissées à quatre pentes. Le caractère aérodynamique de cette forme était particulièrement bien adapté au site, car il permettait de réduire l'impact visuel des deux bâtiments par rapport aux constructions voisines. Il permettait par ailleurs de réduire de façon significative les effets de pression négative sur les façades du bâtiment de plus faible hauteur en cas de vent défavorable. Dans cette version, les simulations aérauliques ont permis de démontrer qu'il était possible d'assurer le renouvellement de l'air par une ventilation naturelle sans recours à une assistance mécanique.

Ces études nous ont permis de démontrer que l'efficacité d'un système de gestion relevant d'une logique environnementale n'est pas qu'une simple affaire de dispositifs techniques, mais qu'elle relève avant tout de la justesse de la conception architecturale du projet, laquelle est nourrie par une réflexion technique.

Le projet considéré n'a malheureusement pas pu être réalisé. Mais la conception envisagée pour les bâtiments en question nous a permis de développer de nouvelles manières de construire qui nous servent pour nos projets actuels mais aussi futurs.

Conclusion

Nous pouvons aujourd'hui considérer que si les consommations de chauffage dans les bâtiments neufs peuvent encore être améliorées, elles ne constituent plus aujourd'hui le principal enjeu énergétique. La succession des réglementations thermiques et énergétiques a en effet permis d'améliorer considérablement le niveau général de la production du bâti. La nouvelle réglementation énergétique (RE 2020) témoigne de cette évolution puisqu'elle porte désormais sur des problématiques beaucoup plus vastes liées à la décarbonation générale du bâtiment.

Les améliorations sont donc à chercher au niveau des autres postes de consommation, et notamment des postes de ventilation et de climatisation. La réalisation de bâtiments capables de s'adapter aux évolutions climatiques sans recourir ou que faiblement à l'air conditionné est ainsi devenue un sujet central qui est encore insuffisamment pris en compte.

Par ailleurs, la réduction des émissions de CO_2 liées à la phase de construction et à la transformation des bâtiments est devenue, elle aussi, un sujet central. Si l'utilisation du bois et des matériaux bios et géosourcés apporte une première partie de la réponse, d'autres évolutions sont à espérer. De ce point de vue, la réalisation de bâtiments évolutifs et réversibles est devenue un enjeu essentiel.

Data protection in a global economy

Éditorial

Marie-Laure Denis, State Councillor, President of the CNIL

Introduction
What kind of world do we want? How far can and should we go?

Dr Laure Tabouy, PhD, Neuroscientist and ethicist, Ethics and Epistemology team, CESP-INSERM U1018, Espace éthique APHP, Université de Paris-Saclay

1 – Data use: private companies and academic research, together for a more responsible future of data management

In a world that is both connected and in tension, what are the challenges and approaches to ensure the value and protection of data?

Jérôme Andres, Secure Information and Communication Systems, Thales Group

Digital revolution has induced massive economical transformations, since the last thirty years. Data exchange, thanks to the Internet has been cornerstone to this revolution, enhanced by their valuation in all domains: science, health, society, politics, arts, entertainment… Information and knowledge they produce become actual assets, to protect for their own value, as well as for consequences they can cause, directly or indirectly. Data emergence is not neutral, like other aspects of Digital and any innovation: as much a cure as a poison, they can lead to both risks and opportunities. Cybersecurity is a practice and an industrial sector that can allow gauging and marshalling such routes. In this context, political regulation and balances are needed, at both national and international.

Targeting, through data, the breaks in the healthcare pathways: an opportunity for patients and innovators

Marco Fiorini, Director of the "Artificial Intelligence and Cancer" project of the strategic contract for the healthcare industries and technologies, **Stéphanie Kervestin**, General Delegate of Ariis (Alliance for Research and Innovation in the Healthcare Industry), and **Virginie Lasserre**, Director of External Affairs at Janssen France, co-leader of the "AI and Care Pathways" axis of the strategic contract for the healthcare industries and technologies sector

In this article, we put forward the idea that the pharmaceutical, medical device design, diagnostics and digital health industries have distinct histories and technologies.

In contrast, the production of health data by each of these industries creates a common breeding ground for innovation. This breeding ground is particularly fertile when it focuses on the description of patients' health care pathways, which cover the prevention, diagnosis, treatment and follow-up phases. This "patient perspective" is the one served by all these sectors and, beyond the private sector, the one that is the subject of public policy.

We describe here how, within a strategic contract for the health industries and technologies, the ability of these industries to work with the State to develop a common semantic for the description of health care pathways makes it possible to consider a method for quantifying the disruptions that occur within these pathways, which are as much a loss of opportunity for the patients.

The aim here is to objectively assess their importance in order to prioritize them with regard to the technologies available in France.

Towards a right of ownership of personal data

Alain Bensoussan, Attorney at Law, Lexing Alain Bensoussan Avocats

The challenges linked to "the permanent emergence of new technologies and the omnipresence of personal data processing in all areas of life", as recently underlined by the president of the CNIL, Mrs. Marie-Laure Denis, place data more than ever at the center of all attention, and with it, the question of the ownership of personal data, their patrimonialization and the right of everyone to monetize their own information.

What rights for data in a data-centric economy?

Bertrand Warusfel, Professor at University of Paris 8, member of the Paris Bar (FWPA)

In an economy whose innovations and productivity are largely based on the production, exchange and processing of digitized information, data is acquiring increasing economic, social and political value. But the corollary of this statement is that, like any value, "data" is subject to extreme competition and provokes both litigation and a demand for European, or even international, regulation.

We are therefore seeing the gradual emergence of a data law, which we would like to summarize here. This legal framework remains rather heterogeneous and very partial. But the very important (and no doubt quite disturbing) effects of the exploitation and algorithmic processing of data in the years to come should serve to accelerate the structuring of this law, provided that clear political choices are made to specify the essential values that a digitized economy must respect.

Data-driven and AI in healthcare: bringing the "Human Guarantee" MedTech and HealthTech ecosystem to life!

David Gruson, Director of the Luminess Health Program and co-founder of Ethik-IA

PariSanté Campus is the new major ecosystem of digital health in France and internationally. Its priority orientations include the development of a strong Healthtech and Medtech sector, focused on data-driven management and artificial intelligence (AI) in healthcare. The attractiveness of France and Europe will depend not only on the deployment of innovative technologies and methodologies in these areas, but also on a strong commitment to positive ethical regulation centered on the new principle of the "Human Guarantee" of AI.

People in the Sun: light up your data

Charles Huot, People in the Sun Company

It is nowadays a truism to say that data is one of the essential drivers of the economy. Digital is present at all levels of society and one of its physical expression is the growing increase (+50% per year) of data production and storage capacities. Data is born and grows over and over again; it never dies; and the more we consume it, the richer it becomes. Unlike our natural resources, data is an inexhaustible resource that irrigates the world's communication networks under water, on land, in the air and in space. We should perhaps replace the verb "to irrigate" with the verb "to flood" or "to submerge" because the volume of data is so overwhelming our information systems. We can even talk about digital bulimia.

At People in the Sun, we accompany organizations in their reflections on the use of data. The questions are multiple and cover technological, economic and legal fields, but also ethical ones. Each of these fields deals with a particular aspect of data.

Valorization of human health research and data protection in the digital age

Frédérique Lesaulnier, Doctor of Law, Data Protection Officer of the Brain Institute

Human health research is undergoing a digital revolution due to the huge amounts of data available, which are collected in multiple environments, and the possibility of extracting knowledge and correlations from them thanks to technologies that increase storage and processing capacities. At a time of open science grand challenges, the GDPR is leading to a refocusing of the data protection organization on the organizations that process it. Compliance with the regulation must be integrated into a global approach to data governance and requires the involvement of the people who generate the data. We will present here a few avenues of reflection and action developed by research actors for an ethical and responsible use of personal health data.

International research and data protection

Gaëlle Bujan, Data Protection Officer, CNRS

Research is evolving on a global scale in terms of its foundations, its values and its networks of experts. In the digital age, the acceleration of exchanges and research perspectives accompany the dissemination of knowledge, the limits of which are constantly being pushed back. At the same time, the risks are new and multiplied; they are associated with the imperatives of integrity, ethics and respect for people. Regulations are adapting to this movement: they aim to make the players more responsible, to contribute to society's confidence in research and are becoming increasingly protective of rights and individuals.

In this evolving and seemingly restrictive environment, research organizations have put in place policies that promote a culture of data protection and the achievement of their objectives on an international scale. International research and data protection regulation are not mutually exclusive.

Aligning access to microbiome data and privacy considerations for better solutions for health and wellbeing of society and environments

Frederik Coppens, VIB-UGent, ELIXIR Belgium, Gent, Belgium, **Lene Lange**, BioEconomy Research & Advisory, Copenhagen, Denmark, and **Kathleen D'Hondt**, Department Economy, Science and Innovation, Flemish Government, Brussels, Belgium

There is a growing body of evidence that underpins the importance of microbiomes in biology. Understanding the functioning of microbiomes and their interaction with the environments will allow to develop novel interventions to support human, animal, and plant health as well as the environment. The potential that microbiomes can have to prevent the onset of non-communicable diseases is huge. This can only be developed when studying the impact of lifestyle, nutrition and environment in the context of the genetic content. As human microbiomes have been shown to be stable over time and can allow to identify the 'carrier' of the microbiome, access to microbiome data has been questioned in the light of privacy protection and the General Data Protection Regulation. In this paper we discuss the potential of microbiomes in different areas and how microbiome data may be shared to support the concept of doing good.

2 – The data economy: what uses for data at the heart of collaborations, partnerships or collaborative platforms. What data to share? Who should have access to it?

Epistemological issues in data science

Jean-Gabriel Ganascia, Specialist in artificial intelligence

After recalling the singularity of "masses of data", which is not only due to their volume, but also to their evolutivity and variability, we will show that both their accumulation and their exploitation have proved necessary for the major players of the Web and that this is due to three reasons linked to the specificity of digital industries. We will then reflect on data science and on the opposition between, on the one hand, those who claim that correlations are now sufficient and, on the other hand, those who still insist on the use of models and on the key epistemological function they play in the scientific process. We will conclude on the current lack of a mathematical theoretical framework for data science, while evoking the old theories, those that existed in the 1990s, and opening up to progress in this direction.

Encryption, or the contribution of cryptology to the security of data storage, transmission and processing

Louis Goubin, Professor at the University of Versailles Saint-Quentin-en-Yvelines – Université Paris-Saclay, director of the "Cryptology and Information Security" research group at the LMV laboratory (UMR CNRS 8100)

The development of techniques for the storage, transmission and processing of digital data creates an increasingly acute need for securing this data. Cryptology, which is often called the science of secrecy, provides solid answers, often with mathematical proofs, to the question of data confidentiality, and thereby to the protection of privacy. We propose here an excursion through the problem of data encryption, from basic cryptographic principles to more complex applications requiring the ability to perform calculations on encrypted data. New perspectives are opening up thanks to recent techniques, one of the challenges being to articulate security and regulatory compliance, when it comes, for example, to supply chains in industry or platforms implementing a sometimes worrying analysis of the users' personal data.

Helping researchers better manage their research data: the data librarian profession

Laetitia Bracco, Curator of libraries at the University of Lorraine

Research data are progressively considered as scientific productions of their own, in all disciplines. In the context of Open Science, more and more requirements are made by funders and by public policies in general to better produce, structure, preserve and open these data. Access to the data underlying publications is also increasingly requested by scientific journals, as part of their commitment to integrity and transparency. Scientific and technical information staff, in university libraries as well as in other structures, are thus required to support and train researchers in all these issues. To carry out this mission, it is necessary to increase the skills and develop new activities for those who hold the emerging position of data librarian.

The Health Data Hub, a lever for the valorisation of health data

Stéphanie Combes, Director of the Health Data Hub (HDH)

By seeking to mobilize the full potential of health data to improve the health system, the Health Data Hub is at the crest of the innovation wave.

Its objective is to enable and simplify access to health databases (mainly the Système national des données de santé (SNDS), the French national health insurance database) in order to reuse them for research purposes.

Faced with the many issues raised, the platform has taken the decision to include its ecosystem and citizens in its project, as well as to open up to an international perspective: to be able to align the Health Data Hub with all the actors involved and future prospects.

Data science in health and data protection of Metavers

Adel Mebarki, Co-founder and CEO of Kap Code

The use of social networks in healthcare is constantly growing withinin an increasingly connected society. These platforms have become real tools in patients' pathways. From searching for information to building communities, social networks are becoming a permanent part of the "digital care pathway" for patients and healthcare stakeholders. This spontaneous generation of real-life data is the subject of multiple research projects for public health purposes. However, with the advent of the Metavers, several ethical and societal issues arise around the use of these sensitive data.

The challenges of Metaverse in terms of personal data protection

Thomas Fauré, President and founder of Whaller

The metaverse announced by Mark Zuckerberg will be a universe beyond the one we know in which "you will be able to do almost anything you imagine: work, learn, play, buy, create (...)". Cut off from the physical world, the "metanaut" will be totally immersed in a new digital world. With no limits, the metaverse will have major anthropological consequences. At the same time, the amount of personal data produced will be exponential, as will the difficulties of protecting it. Everything the user does can be exploited for or against him. "Augmented data", biometric and behavioral, will make it possible to offer even more targeted advertising than today. GAFAMs are already filing patents for their collection

and potential exploitation. In this new world, where current regulations will quickly become outdated, it is important to think about the future protection of each person's fundamental rights.

3 – Humanities and data: the challenges of changing societies through data. Trust in the digital age

A decade and a half of OECD action on data governance policy-making

Elettra Ronchi, Senior Policy Consultant on Data Governance and Digital Health WHO/Europe; former Head of the Data Governance and Privacy Unit in the Division for Digital Economy Policy at the OECD, and **Christian Reimsbach-Kounatze,** Information Economist and Policy Analyst in the Division for Digital Economy Policy at the OECD

The OECD has long recognized the need to better understand how to reconcile the risks and benefits of data access and sharing to help governments reap the benefits of data-driven innovation. To guide policy-making, the OECD has produced over the last decade and a half a significant body of analytical work and legal instruments setting out principles and best practices to address sector-or domain-specific challenges in the governance of data. These Recommendations include: the Recommendation concerning Access to Research Data from Public Funding ; the Recommendation for Enhanced Access and More Effective Use of Public Sector Information ; and the Recommendation on Health Data Governance . In what appears to be the latest strong demonstration of its commitment to the issue, the OECD Council adopted in 2021, the Recommendation on Enhancing Access to and Sharing of Data (EASD Recommendation). Differently from the preceding ones, the EASD Recommendation provides an overarching set of principles and policy guidance to help governments reconcile potential risks and benefits and unlock the re-use of all types of data across and within sectors, jurisdictions, organisations, and communities. The aim of this paper is to put in context this significant body of work and set out the main policy issues addressed by these OECD Recommendations.

The CNIL faces the challenges of building a trusted digital society

Étienne Maury, CNIL

For the CNIL, ensuring trust in the digital age responds to legal and regulatory issues, but also ethical, technological, economic and societal ones. Its action and its role are evolving and face multiple challenges, which are not only national but also European and international, given the geography of the globalized digital economy. The apprehension of these challenges constitutes the framework in which the CNIL places its strategy to respond to current and future matters. In addition to the need to guarantee the effectiveness of the fundamental right to the protection of personal data, there are also issues related and intrinsic to the evolution of the digital ecosystem, whether in terms of support and innovation but also legislative and regulatory developments at European level.

Data altruism: how can data be used to serve the general interest?

Éric Salobir, President of the executive committee of the Human Technology Foundation and founder of OPTIC

The volume of data continues to grow, yet it is largely underutilized. This paradox is a major obstacle to initiatives that benefit the general interest. Data altruism, an innovation in data sharing theorized by the European Commission, could be a solution to this problem by removing the mistrust that hinders such sharing. In the report "Data altruism: a European initiative, data for the public good", the Human Technology Foundation and the Sopra Steria Next Exploratory set out to show how the still theoretical concept of Data altruism can become a reality. In this article, we detail the major proposals of this report and show that it is possible to build a system that greatly facilitates the provision of data to help initiatives working in favor of the general interest.

Has data become the primary issue in cybercrime?

Éric Freyssinet, General officer of the Gendarmerie

Data is at the heart of cybersecurity issues, as the main target of cybercriminals, as a tool for the same, but also as a tool for information systems defenders. This concerns both personal data and all sensitive data of organizations. And the risks are not only related to the famous ransomware, even if it is the most dynamic threat.

The challenge of data for cyber defense

Didier Danet, Lecturer (HDR) at the University of Rennes 1

The digital space is a field of conflict, where the control of data is an issue of growing importance. Indeed, contrary to a widespread vision, cyber defense is not only interested in the protection of interconnected information systems (the container), but also in the protection of information content, of which data is the raw material. However, if the control of data is becoming a central issue in cyber defense, it is rendered almost illusory, notably because of the "datafication" of the world, the digital revolution at work, and new social practices that weaken the capacity of states to control the production and circulation of data. In this article, we suggest exploring two avenues of reflection: the first aims to better protect the stocks of data in the possession of civilian and military institutions, and the second to make actors responsible for generating data flows.

4 – What does the data say about humans?

Accountability at the heart of data protection: what data says about the human being

Dr Laure Tabouy, PhD, Neuroscientist and ethicist, Ethics and Epistemology team, CESP- INSERM U1018, APHP Ethics Area, University of Paris-Saclay

The acceleration of innovations makes it essential to reflect on the societal, ethical and legal issues related to the use of data, in particular on the notion of responsibility. The design of interdisciplinary safeguards and evaluation and monitoring systems, as well as the definition of governance adapted to the sociological, ethical and legal values of different countries, are currently emerging worldwide. It is around the need to agree on the notion of social responsibility that, for example, the neuroethics called for by the OECD Council through its recommendation n°0457 of 2019 on responsible innovation in neurotechnologies is being built. In the reflection on the notion of responsibility, the philosopher can bring an important light on this question. It is therefore by summoning Hans Jonas and Hannah Arendt, but also by using research ethics and neuroethics as well as existing laws and recommendations, that this work around social responsibility concerning data has taken shape.

Concluding Remarks

Dr Laure Tabouy, PhD, Neuroscientist and ethicist, Ethics and Epistemology team, CESP- INSERM U1018, APHP Ethics Area, University of Paris-Saclay

Miscellany

Experiences with passive or positive energy buildings

Pascal Gontier, Professor at the École nationale supérieure d'architecture de Nantes and member of the Académie d'architecture

From the end of the 1990s onwards, the subject of energy savings has been gaining ground in the architectural debate, and requirement levels are gradually evolving. Those requirements which were focused on heat consumption at first are now extending to energy consumption as a whole and are taking carbon consumption into account.

My architecture practice has been environmentally committed since its very beginning, and this commitment has always been taking us forward, by anticipating changes when possible and going beyond regulatory or programmatic requirements. We are convinced that the twofold challenge that is energy consumption and environmental issues cannot be fulfilled through simple normative and technical solutions only.

Issue Editor : **Dr Laure Tabouy, PhD**

Ont contribué à ce numéro


D.R.

Jérôme ANDRES est directeur de la Politique, en charge de l'IA des produits et de la robotisation, pour la branche Systèmes d'information et communication sécurisés du groupe Thales.
Ancien élève de l'École polytechnique et de l'École Télécom Paris, il intervient depuis vingt-cinq ans dans l'industrie, fort d'une double connaissance en matière de télécommunications et de secteurs souverains.
Il a occupé différentes fonctions dans le groupe Alcatel, puis Alcatel-Lucent : ingénierie, standardisation, gestion de produits et de lignes de produits, soutien avant-vente et après-vente, dans un contexte à forte composante internationale dans le domaine des réseaux mobiles et de l'Internet.
Il rejoint le groupe Thales en 2013 pour jouer un rôle au sein de la direction des Offres et du business development pour de grands systèmes d'information en France et en Europe, en lien avec la sécurité civile, la justice, la cybersécurité et la protection de la vie privée. En 2018, il renoue avec des fonctions de politique produit concernant les données critiques en grand volume (*Big Data*) dans un contexte de souveraineté, puis pour coordonner la politique produit de Thales dans les communications et les systèmes d'information sécurisés. Il travaille aujourd'hui plus spécifiquement sur l'introduction de l'IA dans ces produits, les enjeux ESG (Environment, Social & Governance) ainsi que sur l'introduction de la robotisation et les impacts de celle-ci dans le champ de bataille aéroterrestre.
Il porte une attention majeure aux implications géopolitiques, sociales et industrielles des nouvelles technologies, ainsi qu'un regard exigeant et lucide sur la place de la souveraineté au niveau national et européen.


D.R.

Alain BENSOUSSAN est avocat à la cour d'appel de Paris. Il est un précurseur du droit des technologies avancées et un expert reconnu du droit des données à caractère personnel, dont il a accompagné l'émergence dès 1978. Il a fait de l'élaboration de concepts nouveaux l'une de ses marques de fabrique : domicile virtuel, droits de l'homme numérique, vie privée résiduelle, etc. En 2012, après avoir créé Lexing®, premier réseau international à fédérer des avocats en droit du numérique et des technologies avancées, il lance au sein de son cabinet un département sur le droit de l'intelligence artificielle et des technologies robotiques, y voyant « la reconnaissance par le droit d'une mutation technologique au moins aussi importante que l'ont été l'informatique et les réseaux sociaux au XX^{e} siècle ».
Il est président fondateur de l'Association des Data Protection Officers (ADPO).


D.R.

Laetitia BRACCO est conservatrice des bibliothèques au sein de la mission Appui recherche de la direction de la Documentation, à l'Université de Lorraine. Elle y occupe le poste de *data librarian*. À ce titre, elle coordonne l'accompagnement des chercheurs, ingénieurs et doctorants de l'établissement sur la question des données de la recherche et fait partie de l'équipe de bibliométrie. Elle coordonne également la communication Science ouverte de l'Université de Lorraine. Au niveau national, elle anime le groupe de travail Science ouverte – Données de Couperin et pilote le projet de Baromètre français de la Science ouverte relatif aux données de la recherche et codes logiciels.


D.R.

Gaëlle BUJAN, économiste de formation, a été chargée de la politique régionale de recherche au conseil régional de Bretagne de 1994 à 2001.

À l'Institut de recherche pour le développement, elle est administratrice du Centre de recherche Île-de-France et est également administratrice de l'IRD à Paris et adjointe au Secrétaire général de cet organisme.
En 2007, elle est adjointe au délégué régional de la délégation Île-de-France Ouest et Nord du CNRS.
En janvier 2012, elle est nommée déléguée régionale de la délégation Alsace du CNRS, puis, en novembre 2014, elle devient déléguée régionale de la délégation Aquitaine du CNRS.
Depuis avril 2018, elle est la déléguée à la protection des données du CNRS.


D.R.

Stéphanie COMBES
est directrice du Health Data Hub (HDH), un groupement d'intérêt public visant à faciliter l'accès aux données de santé pour la recherche et l'innovation. Responsable du lancement du Health Data Hub, après avoir été rapporteure de la mission de préfiguration de ce groupement et cheffe de projet au *Lab* Santé de la Drees, elle a accompagné toutes les étapes de construction du HDH, jusqu'à l'entrée en service de ce dernier, en décembre 2019.
Diplômée de l'École polytechnique, de l'ENSAE et de la Paris School of Economics, elle intègre en 2010 la direction générale du Trésor, s'intéressant aux questions liées aux politiques énergétiques. Elle coordonne ensuite une équipe chargée de la production de prévisions à court terme relatives à l'évolution du PIB en France, avant de rejoindre, en 2014, l'Insee, où elle est en charge de la création de l'activité *Big Data*, qui préfigure la création du laboratoire d'innovation de l'Insee.


D.R.

Frederik COPPENS
is Head of Node for ELIXIR Belgium and is IT manager at the VIB-UGent Center for Plant Systems Biology. For more than a decade, he has focussed on providing infrastructure and services for data in life sciences. Frederik Coppens is heading a multidisciplinary team focussing on FAIR data and reproducible data analysis. The team is involved in leading roles in many European projects, contributing to the development of the vision of the European Open Science Cloud. Frederik Coppens is co-leading RDMkit, the data management toolkit for bioscientists and data stewards developed by ELIXIR, and Work-flowHub, the ELIXIR registry for computational workflows. Frederik Coppens is member of the Galaxy Executive Board and ELIXIR Belgium hosts a Belgian Galaxy instance. The team contributes to the further development of the Galaxy Research Environment, with a focus on facilitating access to and sharing of data and provenance of workflows. Frederik Coppens was appointed as Belgian delegate in the Strategy Working Group on Data, Computing and Digital Research Infrastructures for ESFRI, 1 Million Genomes WG5 (ICT), the Flemish Supercomputer Center User Council, and the Flemish Open Science Board. More recently, access to and sharing of human (genomic) data has become for him a priority, contributing to the establishment of a biobank and associated digital ecosystem in Belgium to link health and research data, aligned with the developments in Europe.


D.R.

Didier DANET
est maître de conférences (HDR) de l'Université de Rennes 1. Il est détaché à l'Académie militaire de Saint-Cyr Coëtquidan. Il est membre du groupement GEODE (Géopolitique de la datasphère), qui est dirigé par Frederick Douzet, Université Paris 8. Ses recherches portent sur les mutations à l'œuvre dans les conflits contemporains, en particulier l'autonomisation croissante des systèmes d'armes et la question des systèmes d'armes létaux autonomes (SALA), ainsi que la conflictualité dans l'espace numérique et la gestion des crises cyber. Parmi ses publications récentes, il a co-dirigé, avec Stéphane Taillat et Amaël Cattaruzza, *Cyberdéfense, politique de l'espace numérique*, aux Éditions Armand Colin, Collection U. Il a également produit plusieurs articles portant sur la question de la souveraineté numérique, en particulier avec Alix Desforges (publication dans la revue *Hérodote*, les *Annales des Mines*...).
Directeur du Mastère spécialisé en cyberdéfense, il intervient dans le cadre de la formation initiale et continue au travers d'enseignements de gestion appliquée au domaine militaire : gestion des crises en cyberdéfense, intelligence économique, analyse économique de la Défense…


D.R.

Marie-Laure DENIS
est Conseiller d'État. Elle est présidente de la CNIL depuis février 2019 (nommée par décret du Président de la République pour un mandat de cinq ans).
Elle est diplômée de l'Institut d'études politiques de Paris (1988) et ancienne élève de l'École nationale d'administration (promotion « Condorcet »).
Elle a été auditrice (de 1992 à 1995), puis maître des requêtes (de 1998 à 2002) au Conseil d'État. En 2017, elle est nommée Conseiller d'État, rapporteur à la 6e chambre de la Section du contentieux et membre de la Section du rapport et des études.
Elle a été directrice adjointe du cabinet du maire de Paris de 1996 à 1998 et directrice du cabinet du ministre délégué à la Famille et directrice adjointe du cabinet du ministre de la Santé, de la Famille et des Personnes handicapées de 2002 à 2004.
Elle a été membre du CSA (Conseil supérieur de l'audiovisuel) de 2004 à 2016, puis membre de l'Arcep (Autorité de régulation des communications électroniques et des postes) de 2011 à 2016. Depuis 2017, elle était membre du CORDIS (Comité de règlement des différends et des sanctions) de la Commission de régulation de l'énergie (CRE).


D.R.

Kathleen D'HONDT was trained as a molecular cell biologist both in academia and in industry. She got a PhD at the University of Ghent (Belgium) and spent several years as a postdoc in the Ghent University and VIB, the Wageningen University (the Netherlands) and the Biozentrum in Basel (Switzerland). In 2006, she joined the Department of Economy, Science and Innovation of the Flemish government as a Policy Analyst and joined in 2013 the OECD as a senior policy analyst in the Working Party on Bio-, Nano- and Converging Tech (BNCT). Since 2016, she is back at the Department of Economy, Science and Innovation of the Flemish government. Kathleen D'Hondt is the author of a number of science policy papers emphasising on the potential of microbiomes to address challenges related to health, nutrition and environment. Kathleen D'Hondt coordinates the SAPHIRe (Securing the Adoption of Personalised Health in Regions), a coordination action under Horizon 2020, and the Vanguard Initiative pilot on Smart Health a smart specialisation partnership on Personalised Medicine. Kathleen D'Hondt is also a member of MicrobiomeSupport and in this context ensuring the interaction and communication with the International Bioeconomy Forum and its Working Group on Food Systems' Microbiomes.


D.R.

Thomas FAURÉ est ingénieur centralien, entrepreneur et expert du numérique. Se mettant au code dès son adolescence, il reçoit sa première proposition d'embauche alors qu'il est encore au lycée. À sa sortie de l'École centrale, il entre comme ingénieur chez Sagem Sécurité, devenu Safran Morpho, le leader mondial des systèmes de reconnaissance biométrique d'empreintes digitales. En plus de ses missions, il conçoit une base de connaissances toujours opérationnelle au sein du groupe, et développe plusieurs applications collaboratives. Puis il rejoint Polyconseil, cabinet de conseil en stratégie d'innovation appartenant au groupe Bolloré. Il participe au déploiement d'Autolib, premier réseau de location de voitures électriques en France.

Cependant, mu par l'innovation, il souhaite créer un réseau social à la manière de Facebook, mais qui soit non intrusif et respectueux des données personnelles de l'internaute. Son principe est simple : chaque utilisateur possède un seul compte, avec lequel il a accès à plusieurs « sphères » : sa sphère professionnelle, ses sphères privées ou encore celles liées à ses engagements associatifs. C'est Whaller, la réunion de wh + all, c'est-à-dire « ensemble ». Seul, travaillant la nuit, il code. En 2013, grâce à Vincent Bolloré, qui est son premier investisseur, il crée Whaller.

Souvent présenté comme l'anti-Facebook, Whaller est un réseau social français précurseur, offrant le plus haut niveau de sécurité en ligne. Il est destiné aussi bien aux entreprises, aux collectivités et aux administrations, qu'aux particuliers. Il propose des outils de gestion de projets, des intranets et des espaces collaboratifs. Il est utilisé par près d'un million d'utilisateurs réguliers.

Whaller compte plus de 30 collaborateurs permanents et plus de 200 clients.

Thomas Fauré intervient régulièrement dans le débat public pour défendre la souveraineté numérique française, tant sous forme de tribunes que de participation à des conférences publiques ou à des auditions avec des élus ou des représentants des pouvoirs publics. Il a publié deux ouvrages, *Transmettez !* (aux Éditions Baudelaire, 2018) et *Après Facebook, rebâtir* (Les Nouvelles éditions de Passy, 2022).

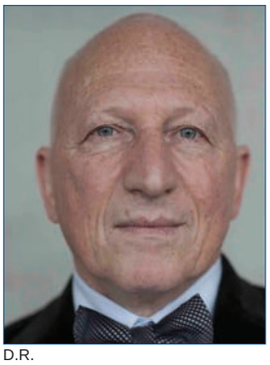

D.R.

Marco FIORINI dirige le projet « Intelligence artificielle et cancer » du contrat stratégique de filière des industries et technologies de santé. Il est membre du comité Innovation du Medef.

Avant de rejoindre ce projet, il a dirigé l'Alliance pour la recherche et l'innovation des industriels de la santé (ARIIS). Il a dirigé le consortium public de valorisation de l'Alliance nationale pour les sciences de la vie et de la santé (Aviesan). Marco Fiorini a débuté en tant que *data analyst* au Commissariat à l'énergie atomique et aux énergies alternatives (CEA) et y a été conseiller au sein de la direction de la Stratégie et des programmes.


D.R.

Éric FREYSSINET est officier général de gendarmerie, commandant en second de la gendarmerie dans le cyberespace, poste qu'il occupe après vingt-quatre ans de carrière dans différents postes à responsabilité relevant du champ de la lutte contre la cybercriminalité. Ingénieur de formation (École polytechnique, X92), il complète sa formation, en 2000, par l'obtention d'un mastère spécialisé en sécurité des systèmes d'information et des réseaux (Télécom ParisTech) et, en 2015, par la soutenance d'une thèse de doctorat en informatique sur la lutte contre les *botnets* (Université Paris 6).


D.R.

Jean-Gabriel GANASCIA est ingénieur et philosophe de formation initiale. Il s'est très tôt orienté vers l'informatique et l'intelligence artificielle. Il a soutenu, en 1983, à l'Université Paris-Sud (Orsay), une thèse de Doctorat sur les systèmes à base de connaissance, puis, en 1987, toujours à l'Université Paris-Sud, une thèse d'État sur l'apprentissage symbolique. Professeur à la faculté des sciences de Sorbonne Université depuis 1988, il poursuit ses recherches au LIP6, où il dirige l'équipe ACASA (Agents cognitifs et apprentissage symbolique automatique). Dans le passé, il a créé et dirigé le groupement d'intérêt scientifique « Sciences de la cognition » au CNRS et le labex OBVIL (Observatoire de la vie littéraire). Spécialiste d'intelligence artificielle (EurAI Fellow), d'apprentissage machine et de fouille de données, ses recherches actuelles portent sur le versant littéraire des humanités numériques, sur l'éthique computationnelle et sur l'éthique des technologies de l'information et de la communication.

Il est membre du CNPEN (Comité national pilote d'éthique du numérique), président du comité d'éthique de Pôle emploi, du comité d'orientation du CHEC (Cycle des hautes études de la culture) et de l'AFAS (Association française pour l'avancement des sciences). Enfin, il a présidé le comité d'éthique du CNRS de 2016 à 2021.

Au cours de sa carrière, il a publié plus de 500 articles dans des actes de conférences, dans des livres et des revues scientifiques. Il est aussi l'auteur d'une dizaine d'ouvrages destinés au grand public, dont les trois derniers sont : *Servitudes virtuelles*, aux Éditions du Seuil, Collection « Sciences ouvertes », 2022 ; *Le mythe de la singularité : faut-il craindre l'intelligence artificielle ?*, aux Éditions du Seuil, Collection « Sciences ouvertes », 2017 (cet ouvrage a obtenu le prix Roberval grand public) ; *Intelligence artificielle : vers une domination programmée ?*, Le Cavalier Bleu, Collection « Idées reçues », 2017.


© Alessandro Silvestri

Pascal GONTIER est diplômé de l'École nationale supérieure d'architecture de Versailles, Postgrade de l'EPFL (École polytechnique fédérale de Lausanne) en architecture et développement durable. Il est titulaire du Master européen en architecture et développement durable (École polytechnique fédérale de Lausanne).

Pascal Gontier est professeur à l'École nationale supérieure d'architecture de Nantes. Il est également membre titulaire à l'Académie d'architecture depuis 2016.

Il est l'auteur de nombreux articles sur l'architecture et l'écologie. Il est également l'auteur du livre *Home, habitat ouvert et sur mesure*, publié en 2018.

Il a créé son agence d'architecture et d'urbanisme en 1997. Sa démarche est marquée par un goût prononcé pour l'innovation et l'expérimentation, ainsi que par un engagement fort et des compétences reconnues dans le domaine de la transition écologique.

Il est membre du conseil d'administration de l'Association pour le développement des immeubles en bois (ADIVBOIS) et membre de l'Institut pour la conception environnementale du bâtiment (ICEB).

Son travail a été présenté dans le cadre de nombreuses expositions, notamment : Exposition Vivre Bois, Galerie au Roi, en 2019 ; Salon d'Automne, à Paris, 2016 ; Habitat III à Quito ; exposition à l'Alliance française d'un bâtiment en structure bois à Pessac, 2014 ; Architecture française en bois, Linz, Autriche, 2010 ; Habiter 10.09/09.10, exposition au Pavillon de l'Arsenal, Paris, 2010 ; Paris +2 degrés, Parc de Bercy, Paris, 2009-2010 ; Villes rêvées. Villes durables, Fondation EDF, Paris, 2009 ; Habiter 09.08/09.09, exposition au Pavillon de l'Arsenal, Paris, 2009.

Il a reçu différents prix pour ses réalisations : Grand prix d'architecture du Salon d'Automne, Vorarlberg Holzbau Price 2011 (Maison Gaïta – Maison passive et à énergie positive), catégorie bâtiment étranger ; le Prix AMO 2013 Spécial Fondation d'Entreprise SMA pour des logements sociaux à Paris pour la RIVP ; 1er prix Trophée Bois Île-de-France 2016 ; et 1er prix Bas Carbone des Green Building Awards 2016 France pour le bâtiment Max Weber – Université de Paris Nanterre.


D.R.

Louis GOUBIN est professeur de classe exceptionnelle à l'Université Versailles Saint-Quentin-en-Yvelines, au sein de l'Université Paris-Saclay, où il dirige le groupe de recherche « Cryptologie et sécurité de l'information » au laboratoire LMV (UMR CNRS 8100).

Ancien élève de l'École normale supérieure (Paris) et agrégé de mathématiques, il est titulaire d'un Doctorat en mathématiques pures de l'Université Paris-Sud (1995) et d'une Habilitation à diriger des recherches (HDR) de l'Université Paris-Diderot (2003).

Il a publié près de 90 articles de recherche, cités plus de 5 000 fois au total, sur la conception de nouveaux cryptosystèmes symétriques et asymétriques (post-quantiques), la cryptanalyse d'algorithmes et de protocoles théoriques et pratiques, la protection des implémentations logicielles ou matérielles contre les attaques physiques, et sur de nouvelles avancées dans le chiffrement complètement homomorphe (FHE) et la cryptographie en boîte blanche.

Il a également travaillé pendant huit ans dans l'industrie (Bull, Schlumberger, puis Gemalto, et aujourd'hui Thales). Il a déposé une vingtaine de brevets sur des

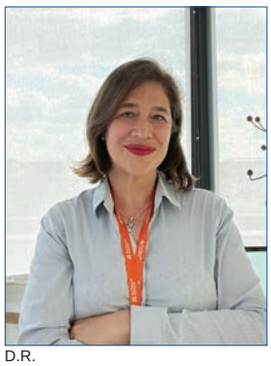
aspects appliqués de la cryptologie, a fondé un master en cybersécurité appelé SeCReTS (Sécurité des contenus, des réseaux, des télécommunications et des systèmes) et a été ou est le directeur de thèses de plusieurs étudiants (au nombre de 23 jusqu'à présent).


D.R.

David GRUSON est ancien élève de l'École nationale d'administration et de l'École des hautes études en santé publique. Il est directeur du programme Santé du groupe Luminess, spécialisé dans la transformation digitale. Il a occupé plusieurs postes à responsabilité dans les domaines des politiques publiques et de la santé. Il a, en particulier, été conseiller du Premier ministre chargé de la santé et de l'autonomie (2010-2012) et directeur général du Centre hospitalier universitaire de La Réunion (2012-2016). Il est professeur à la Chaire « Santé » de Sciences Po Paris. Il est le fondateur d'Ethik-IA, dont les propositions sur l'IA en santé ont inspiré le volet « Intelligence artificielle » de la révision de la loi de la bioéthique et le nouveau règlement européen sur l'IA.
Il est l'auteur de *S.A.R.R.A., une intelligence artificielle* et de *S.A.R.R.A., une conscience artificielle*, premiers polars bioéthiques sur l'IA en santé, parus respectivement en juin 2018 et en mars 2020 aux Éditions Beta Publisher. Leur *spin-off* « Tuer CAMUS – SARRA files » vient de paraître aux mêmes éditions : sous la forme d'une rencontre à travers le temps, y est décrite la rencontre entre Albert Camus et l'intelligence artificielle. Il a également écrit *La Machine, le Médecin et Moi*, ouvrage de référence sur le développement de l'IA en santé, paru en novembre 2018, aux Éditions de l'Observatoire et est co-auteur de *La Révolution du pilotage des données de santé*, un ouvrage paru en mai 2019, aux Éditions Les Études hospitalières.


D.R.

Charles HUOT est un entrepreneur de l'innovation et des *data,* justifiant de plus de trente années d'activité opérationnelle dans le management des environnements *data* et logiciels *high tech*.
Aujourd'hui, il dirige la société People in the Sun, qui accompagne les organisations dans leur parcours d'exploitation et de valorisation des données. C'est dans le cadre de l'une de ces missions qu'il est amené à collaborer à la création de l'entreprise HPC Data France pour le développement de *Big Data centers* en région parisienne. En 2020, il a co-fondé Hydria Data, à Montréal, qui œuvre au développement de *data centers* haute densité en Amérique du Nord ; il en a pris la présidence.
Il est également président du plus grand pôle de compétitivité français, Cap Digital, une association qui assure, dans les régions Île-de-France et Hauts-de-France, une animation entre les acteurs économiques et ceux de la recherche dans l'innovation digitale collaborative.
Avant de co-fonder TEMIS en juin 2000, racheté en 2015 par Expert System, Charles Huot a passé dix ans chez IBM en tant que directeur international des ventes pour les logiciels de Text Mining. Il a soutenu une thèse en analyse de données pour l'intelligence économique à l'Université de Marseille et au Centre européen de mathématiques appliquées d'IBM, obtenant ainsi un Doctorat.
Il est membre de l'Institut des sciences cognitives de l'Université du Québec à Montréal, chargé d'enseignement à l'École de guerre économique dans le domaine du *Big Data* et de l'intelligence artificielle et, depuis octobre 2020, membre du conseil d'administration de l'Institut géographique national.


D.R.

Stéphanie KERVESTIN-YATES est docteure en biologie de Sorbonne Université. Après trois ans à University of Massachusetts Medical School comme research associate, elle entre au CNRS en 2006, comme chargée de recherche, puis est nommée, en 2015, responsable de l'innovation et des partenariats industriels à l'Institut des sciences biologiques du CNRS. De 2019 à 2020, elle est chargée de mission à la délégation régionale à la recherche et à la technologie Île-de-France du ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, avant de revenir au CNRS en tant que responsable du service Partenariats et valorisation à la délégation régionale Paris Centre de l'organisme. Depuis mai 2022, elle a succédé à Marco Fiorini comme déléguée générale de l'Ariis (Alliance pour la recherche et l'innovation des industries de santé).


D.R.

Lene LANGE has been Research Director in both public and private (Novozymes) R&D, and has held full professorships at three Danish universities. She has experience from a portfolio of bioeconomy advisory roles, international, European, Nordic, and Danish, amongst others e.g. Vice-Chair of the Scientific Committee for BBI JU, a 3.7 billion Euro program and member of the Scientific Committee for EU CBE-JU. She has been Board Chair for CIMMYT, CGIAR, International Wheat & Maize Research Institute; and Program Chair for IRRI, CGIAR, International Rice Research Institute. Now, she having

her own company, BioEconomy, Research & Advisory. Current research portfolio includes EU-, Nordic- and Danish-funded projects. Research focus: 1. Discovery of new enzymes for converting biomass to higher value products. 2. Microbiome research, microbiome composition and function, of relevance for food and feed. 3. Inventor of new State-of-the-Art peptide-based functional enzyme protein annotation method (CUPP, Barrett & Lange, 2019). Lene Lange has a strong publication record: in all > 275 peer reviewed papers, books, monographs and several patents which together resulted in > 4.300 citations. From 2015-2022, Lene Lange published 64 peer-reviewed papers. Publication portfolio based on having own research group all the way through the career, alongside with upper management positions. Lene Lange has gained experience spanning the entire value chain, from R&D and patenting, to process and product development, as well as strategic marketing and policy framework, communication and dissemination. Lene Lange has publications in and experience from upgrading an entire spectrum of different types of biomass, green biomass (e.g. grass), yellow biomass (e.g. straw), blue biomass (e.g. macroalgae and fish-cut-offs), and red biomass (e.g. chicken feather or bristles); plus, most importantly upgrading of agro-industrial side-streams as well as municipality waste and the microbial biomass waste water sludge.


D.R.

Virginie LASSERRE est directrice des Affaires externes de Janssen France. Elle est diplômée en pharmacie de l'Université Paris-Sud et est également titulaire d'un master en Marketing et management de l'ESSEC. Elle débute sa carrière au sein du laboratoire Boehringer Mannheim en 1996 en qualité de chef de produit, avant de rejoindre GSK, où elle consolidera son expérience en stratégie commerciale. Virginie Lasserre intègre Janssen en 2006, où elle occupe plusieurs postes. Après une expérience marketing/vente confirmée dans plusieurs domaines, au niveau France et européen, elle intègre le département Accès au marché, en 2017, s'impliquant plus particulièrement sur de nombreux sujets relatifs à l'accès aux innovations, notamment en santé mentale, à la transformation numérique et aux enjeux sociétaux. Forte de son expérience, elle devient responsable des Affaires gouvernementales en 2019, puis directrice des Affaires externes en 2020. Également engagée dans des démarches partenariales auprès de l'écosystème français d'innovation, elle contribue activement à développer sur le territoire national la stratégie d'Open Innovation du groupe Johnson & Johnson.


D.R.

Frédérique LESAULNIER est docteure en droit, spécialiste des systèmes d'information et de la protection des données de santé. Elle a travaillé dix ans à la CNIL, où elle a coordonné le pôle Santé de la direction des Affaires juridiques, puis a rejoint l'Inserm en 2014, où elle a assuré les fonctions de déléguée à la protection des données depuis l'entrée en application du RGPD. Son implication dans la crise sanitaire aux côtés des chercheurs et de la CNIL a été récompensée par le prix Inserm de l'innovation 2020. Elle est depuis décembre 2021, la déléguée à la protection des données de l'Institut du cerveau, qui est spécialisé dans la recherche visant à comprendre le fonctionnement du cerveau sain et les causes et les mécanismes des maladies du système nerveux. Elle est également experte près du Conseil de l'Europe. Elle enseigne dans le cadre de troisièmes cycles universitaires (Master 2 Droit du numérique, parcours « Protection des données personnelles », Paris-Descartes…) et intervient auprès de professionnels. Elle est par ailleurs rédactrice en chef de la revue *Médecine & Droit*.


D.R.

Étienne MAURY a débuté sa carrière en 2007 à Bruxelles au sein du cabinet Weber Shandwick avant de rejoindre le Parlement européen en 2009, en tant qu'attaché parlementaire. De 2013 à 2017, il a été conseiller politique de groupe pour la Commission des libertés civiles, de la justice et des affaires intérieures (LIBE) du Parlement européen, suivant les évolutions législatives et politiques dans les domaines des droits fondamentaux, de la coopération policière et judiciaire, du droit pénal et de la sécurité intérieure. Il a notamment participé au processus législatif et aux négociations ayant conduit à l'adoption du Règlement général sur la protection des données (RGPD) et de la directive Police-Justice sur la protection des données. En 2017, Étienne Maury a rejoint la Commission nationale de l'informatique et des libertés (CNIL) en tant que juriste, pour exercer au sein du service des Affaires européennes et internationales, où il a notamment participé aux travaux du Comité européen de la protection des données (CEPD), prenant part aux activités de coopération réglementaire et à l'élaboration de lignes directrices sur les principales dispositions du RGPD, ainsi que sur les transferts internationaux et l'accès aux données par les autorités publiques de pays tiers. En 2021, il a été nommé conseiller juridique auprès de la Présidente et des Secrétaires généraux de la CNIL.

Il est diplômé de l'École des hautes études en sciences de l'information et de la communication (CELA – Sorbonne Université) et de l'Institut d'études politiques de Strasbourg (Sciences Po Strasbourg).


D.R.

Adel MEBARKI justifie d'une double compétence technique et managériale acquise au travers de ses diplômes d'ingénieur et liées à ses études en écoles de commerce. Spécialisé dans le management industriel, il a effectué son MBA à l'ESC La Rochelle ainsi qu'à l'EGSI La Rochelle, avant de se diriger vers le secteur des nouvelles technologies en santé.
En 2013, il rejoint la CRO Kappa Santé en tant que chef de projet Innovation et se spécialise dans la gestion et le développement de projets d'intelligence artificielle adaptés à la santé au sein du pôle Innovation. Il passe rapidement responsable Innovation et marketing du pôle, et est nommé directeur général adjoint et est co-fondateur de Kap Code lors de sa création en 2016. Aujourd'hui, il manage des équipes pluridisciplinaires présentant des profils allant d'ingénieurs, de pharmaciens ou de *data scientists* aux équipes marketing.
Il justifie aujourd'hui de plus de huit années d'expérience dans des projets liés au *Big Data* et à l'intelligence artificielle. Ses compétences regroupent le NLP (Natural Language Processing) à travers le développement de l'outil Detec't, le Computer Vision suite à la création de W'asm, les IoTs (Internet of Things) avec la production de Connect'inh et l'implémentation de projets *Big Data* complexes grâce à Presidio.
Ses travaux lui confèrent aujourd'hui un statut d'expert reconnu dans le monde de la santé digitale. Depuis plus de deux ans, il est co-porteur du diplôme universitaire Digital Health de Paris-Saclay. En parallèle, il met à disposition son expertise en intervenant dans diverses facultés de pharmacie. Il est également expert indépendant dans la revue de projets collaboratifs pour Cap Digital et est membre du conseil d'administration du Healthcare Data Institute.

D.R.

Christian REIMSBACH-KOUNATZE is an Information Economist/Policy Analyst at the OECD Directorate for Science, Technology and Innovation (STI). Christian Reimsbach-Kounatze has been working in STI on topics related to the Internet and digital economy since 2008. This includes in particular work on data-driven innovation and enhancing access to and sharing of data. He is now co-ordinating Phase III of OECD's Going Digital Horizontal Project on Data Governance for Growth and Well-Being (https://www.oecd.org/digital/going-digital-project/).

Before joining the OECD, Christian Reimsbach-Kounatze worked as a researcher at the Institute for Information and Market Engineering of the Karlsruhe Institute of Technology (KIT) and at SAP Research (Germany). Christian Reimsbach-Kounatze holds a Diploma in Information Science, Engineering and Management and in Economics, both granted by the KIT.


D.R.

Elettra RONCHI, PhD, MPP, is senior policy consultant in data governance, privacy and digital health. Since January 2022, she is consultant to the WHO/Europe Data and Digital Health Division of Country Health Policies and Systems. In her former capacity as Head of Unit in the Division for Digital Economy Policy at the Organisation for Economic Cooperation and Development (OECD), she has led work on data governance, privacy and risk management for over a decade. Work under her guidance has recently included the review of the OECD Privacy Guidelines, the development of the 2021 OECD Recommendation on Enhancing Access and Sharing of Data and the 2016 Council Recommendation on Health Data Governance. During her career as international policy analyst, Dr. Elettra Ronchi has worked extensively on evaluating the conditions for system innovation and digital transformation which she has examined in a variety of sectors with a particular focus on health care systems.
She has served as an expert on various advisory boards and panels, including most recently the Ethics Advisory Council of the International Covid-19 Data Alliance, the Working Group on Data Governance of the Global Partnership on Artificial Intelligence, and the Foundation Board of the Graduate Institute of International and Development Studies.
She started her policy career in 1993 as consultant for the United Nations Development Programme, before which she held academic research and teaching positions in the U.S. and France.


D.R.

Éric SALOBIR est le président du comité exécutif de la Human Technology Foundation et est le fondateur d'OPTIC, réseau international de recherche et d'action plaçant l'humain au cœur du développement des technologies.
OPTIC rassemble plusieurs milliers de chercheurs, d'entrepreneurs et de développeurs de technologies. Ce réseau réalise des projets de recherche en éthique pratique donnant lieu à la publication d'articles et de rapports (www.optictechnology.org). Il accompagne également les décideurs des politiques

publiques et les entrepreneurs dans leur évaluation de l'impact des technologies, à travers du coaching et des formations. À Paris, le réseau anime le Lab.222, un espace de réflexion dédié à l'innovation éthique.
Diplômé de l'École de commerce ISC Paris, il a travaillé à l'ambassade de France à Prague (section Économie et commerce) et au Crédit lyonnais (aujourd'hui, LCL), au sein du département Banque d'affaires, à Paris.
Il a rejoint l'Ordre des prêcheurs (Dominicains) en 2000. Il est prêtre, diplômé en théologie et en philosophie. Il a été rédacteur en chef de la radio ROC FM et membre du conseil d'administration de la Fédération française des radios chrétiennes. Il a également été responsable de la Web TV des émissions catholiques pour la chaîne de télévision française France 2.
Expert auprès du Saint-Siège, il conseille des dirigeants de grandes entreprises et des acteurs des politiques publiques sur les questions éthiques relatives aux technologies de rupture. Il est l'auteur de l'ouvrage *Dieu et la Silicon Valley*.


D.R.

Le **Dr Laure TABOUY** a une double casquette de neuroscientifique et d'éthicienne/neuroéthicienne, et est future PhD au carré (2e PhD en cours).

Elle est chercheuse, consultante, cheffe de projet et formatrice en éthique de la recherche et intégrité scientifique, sur des questions relatives aux enjeux éthiques, sociétaux et légaux des neurosciences, des neurotechnologies, du numérique et des données et de la bioéthique, ainsi qu'aux enjeux de la science ouverte sur le plan de la carrière des PhD.
Elle est docteure en neurosciences moléculaire et cellulaire, en génétique depuis le 19 décembre 2012 (Université de Paris) et justifie d'un master Éthique de la recherche de l'Université Paris-Saclay, obtenu en 2021.
Elle a eu l'audace de se lancer dans un deuxième doctorat en neuroéthique, qu'elle réalise depuis octobre 2021, au sein du CESP-U1018-INSERM de Paris-Saclay, dans l'équipe de Recherches en éthique et en épistémologie. Elle travaille en particulier sur les enjeux éthiques des neurosciences et des neurotechnologies.
Elle est également chercheuse et formatrice au sein du Laboratoire interdisciplinaire sur le Doctorat chez Adoc Talent Management, où elle assure le portage d'un projet traitant de l'influence de la science ouverte sur la formation doctorale et sur la poursuite de la carrière des docteurs à la lumière de l'éthique de la recherche.
Chercheuse engagée, elle est très active dans le mentorat de femmes scientifiques et de doctorantes, au sein de l'association Femmes et sciences. Par ailleurs, elle est membre de la Société française des neurosciences et de la FENS, ainsi que de la Société internationale de neuroéthique et du réseau CORTICO.
Elle intervient en tant que formatrice et consultante auprès d'un public de docteurs et de doctorants, d'ingénieurs, de commerciaux, de cadres, d'étudiants au sein de grandes écoles, de laboratoires de recherche et d'entreprises. Son intervention porte sur les questions éthiques, sociétales et juridiques des neurosciences, des neurotechnologies, de l'hybridation homme/numérique, du numérique et des données, de l'éthique et de l'intégrité de la recherche, de la science ouverte, des compétences et du développement de la carrière des docteurs, de la gestion des projets de doctorat...


D.R.

Bertrand WARUSFEL est professeur de droit à l'Université Paris 8, où il enseigne notamment le droit du numérique et de la propriété intellectuelle. Il est co-directeur du Master 2 Propriété industrielle et innovation en santé. Il est également vice-président de l'Association française de droit de la sécurité et de la Défense.
Il est co-auteur de l'ouvrage *Droit du numérique* (Wolters Kluwer) chez Lamy et l'un des contributeurs au nouveau Code du numérique (Lexis-Nexis, 2021). Il travaille en particulier sur les questions de protection des secrets et d'échange des données dans le contexte numérique.
Bertrand Warusfel est par ailleurs avocat au barreau de Paris (cabinet FWPA), spécialiste en droit de la propriété intellectuelle et des nouvelles technologies.