Rapport présenté en séance plénière
Le 18 février 2016

COMMUNICATION CADRE RELATIVE AU « *BIG DATA* »

M. Maurice RONAI
Rapporteur

Avec le concours de :

Mme Noémie LICHON, juriste au service des affaires économiques

M. Franck BAUDOT, ingénieur au service de l'expertise technologique

M. Félicien VALLET, ingénieur au service de l'expertise technologique

SOMMAIRE

Résumé / synthèse de la position de la CNIL

La révolution des données massives (« *big data* ») cristallise un changement d'époque.

La CNIL a été créée à la fin des années 1970 pour protéger la vie privée des citoyens par rapport aux grands fichiers publics. C'était un univers simple, assez statique. Les mégadonnées dessinent un univers différent. Nous sommes passés des fichiers aux données. Les données sont partout. Elles sont produites par les individus ou les entreprises, utilisées par l'ensemble des acteurs publics et privés.

Le régulateur n'évolue plus dans le même univers. Il doit s'intéresser à l'usage qui est fait de ces données et non plus seulement à leur collecte. Il y a un foisonnement de données que l'on n'arrive pas toujours à contrôler et que l'individu a du mal à appréhender. On ne sait pas toujours *a priori* quelles vont être les finalités pertinentes, puisque c'est justement par ce croisement - parfois à l'aveugle - des données que naissent de nouvelles connaissances et qu'apparaissent de nouveaux services.

L'essor des mégadonnées donne lieu à un débat sur l'inadaptation des législations et notamment aux principes de finalité et de recueil du consentement. Un certain nombre d'acteurs utilisent ces interrogations pour tenter de déconstruire le modèle de régulation à l'européenne.

Votre rapporteur ne souscrit pas aux diagnostics tranchés sur la nécessité de revoir de fond en comble les principes qui sous-tendent le modèle de régulation européen, notamment les principes de finalité et de consentement.

- L'objet de ce rapport est de faire le point sur les solutions offertes par la loi et par le futur règlement pour encadrer cette nouvelle génération de traitements. Il revient sur les solutions qui ont déjà pu être exploitées par le passé.
- Il présente les dispositions du règlement relatives au « *big data* » et tire les conséquences des outils nouveaux dont disposera la Commission pour encadrer cette nouvelle génération de traitements (annexe 1).
- Il fait également le point sur les dispositions du projet de loi numérique (encore en discussion) qui renforcent les capacités d'action de la Commission.

L'essor du « *big data* » donne lieu à toute une série de réflexions qui tournent autour de la transparence, de l'évaluation et du contrôle des algorithmes. Ces problématiques ne sont pas nouvelles pour la Commission (articles 10 et 39 de notre loi). Le projet de loi numérique introduit des dispositions relatives à la transparence des algorithmes qui vont désormais coexister avec celles qui figurent dans la loi de 1978.

Un certain nombre de solutions techniques pourraient permettre de concilier l'essor du « *big data* » et le respect des libertés individuelles : certaines sont d'ores et déjà opérationnelles. D'autres technologies pourraient permettre d'exécuter des calculs sur des données personnelles et obtenir des résultats utiles sans pouvoir accéder ni voir les données des personnes.

Dans cette phase de transition, votre rapporteur estime qu'il convient d'explorer toutes les pistes qui permettront d'appréhender et d'encadrer cette phase nouvelle de l'univers numérique.

Il propose, à cet effet, une grille d'analyse selon deux critères principaux : l'origine des données et l'objectif poursuivi par le traitement. Cette grille d'analyse permet de différencier les traitements en fonction de leurs caractéristiques et de leurs enjeux, pour déterminer le

régime juridique qui leur est *a priori* applicable. Cette approche vise à permettre à la fois une ouverture et une diversification possible des usages des données, allant de pair avec un renforcement des droits et des moyens de contrôle des personnes pour les traitements ayant un impact sur elles.

N.B Dans un souci de lisibilité, les principaux développements du présent rapport sont récapitulés dans des encadrés disponibles dans le corps du texte.

Cliquez ici pour taper du texte.

I. PRESENTATION DES ELEMENTS DE DEFINITION ET DE CONTEXTE

Le terme « *big data* » cristallise une mutation profonde dans la circulation, le traitement et l'économie des données : la « *mise en données du monde* » (*datafication*).

Cette « mise en données du monde » se situe au croisement de trois évolutions qui se complètent les unes les autres :

- Un changement d'échelle dans les volumes de données collectées, lié notamment à la chute des coûts et à la multiplication des capteurs et des objets connectés[1] ;
- L'apparition de technologies qui permettent de traiter des masses de données, de les stocker, de les analyser de façon toujours plus précise et à des coûts de stockage et de traitement réduits ;
- des processus et techniques d'analyse de plus en plus performants permettant de passer de l'ère de l'observation à celle de la prévision et de l'anticipation.

Le « *big data* » est aussi associé à l'émergence d'une « science des masses de données » (*DataScience*[2]).

Il entraîne avec lui de nouvelles approches, comme les « stratégies fondées sur la donnée » (*data-driven strategies*), la mise en place de nouvelles organisations en vue d'assurer une « gouvernance des données »[3] et la désignation de « *Chief data officers* » dans les entreprises, au niveau des états[4] ou dans les villes (Paris et Lyon).

Il cristallise le débat sur la transparence des algorithmes.

Il recouvre, enfin, une grande diversité de projets.

Aussi, dans un premier temps, semble-t-il nécessaire de définir à quels traitements l'on fait référence lorsque l'on parle de « *big data* » et quels sont les caractéristiques communes de ces différents projets.

Au-delà de ces éléments de définition, et face à l'ampleur de ce phénomène, il semble également nécessaire d'avoir une appréhension globale de cette thématique, sans se limiter aux réflexions d'ordre juridique, afin de prendre la mesure des attentes économiques que suscite le développement du « *big data* » et les implications politiques qu'il engendre.

---

[1] Les capteurs représentent un marché de plus de 64 milliards de dollars (Frost & Sullivan). Les capteurs de base historiques (flux, de pression, température) ne représentent plus que 32 % du marché, contre près de 60 % pour les capteurs d'application, capteurs de vitesse, biocapteurs ou de vibration par exemple. Quant aux capteurs émergents, de plus en plus spécialisés, leur part atteint déjà 9 % à près de 6 milliards de dollars. S'agissant des objets connectés (machines, terminaux et appareils connectés), leur nombre serait de 15 milliards, contre 4 milliards en 2010. À l'horizon 2020, le nombre de ces nouvelles générations d'objets connectés pourrait atteindre 80 milliards. L'Internet des choses serait composé à hauteur de 85 % d'objets connectés, pour 11 % de terminaux communicants (smartphones, tablettes, boxes ou téléviseurs) et pour 4 % seulement de machines (M2M).

[2] Les *datasciences* utilisent l'ensemble des méthodes des statistiques et du « *machine learning* » (apprentissage automatique ou profond), la régression linéaire, la régression logistique, les arbres de décisions, les forêts aléatoires, les algorithmes de segmentation et l'ensemble des méthodes de visualisation de données pour concevoir de nouvelles applications. Administrateur général des données Rapport au Premier ministre sur la gouvernance de la donnée 2015.

[3] Une organisation globale des données permettant d'en assurer la qualité, la fraîcheur, l'interopérabilité, la disponibilité dans des formats techniques en facilitant l'utilisation rapide et la meilleure circulation possible afin que l'organisation et chacun de ses agents puisse en tirer parti

[4] La fonction d'Administrateur général des données a été créée en septembre 2014 par le Premier ministre. En Grande Bretagne, le premier CDO a été désigné en mars 2015.

A. Une notion renvoyant à une diversité de projets

Le concept et la pratique du « *big data* » est apparu aux **États-Unis**. Google, Yahoo, Amazon et Facebook (mais aussi Apache) ont pris une part essentielle dans l'émergence du « *big data* » entre les années 2000 et 2006[5]. Les technologies alors existantes, comme les bases de données relationnelles, se révélant incapables de gérer les quantités de plus en plus considérables de données que ces sociétés amassaient (requêtes sur les moteurs, ciblage publicitaire, données d'usage, etc.), ces sociétés ont été amenées à développer leurs propres technologies de stockage et de traitement de ces données. Les premiers projets majeurs ont également été mis en place aux États-Unis.

Les technologies et les démarches « *big data* » devraient concerner, à terme, tous les domaines.

Des approches de ce type sont d'ores et déjà mises en œuvre dans des secteurs aussi divers que ceux du marketing, du e-commerce, de la finance, de l'assurance, de la santé, de la recherche, des ressources humaines, des transports, du régalien ou encore de l'environnement, de l'humanitaire[6] ou des industries culturelles.

Tous ces projets n'impliquent pas nécessairement des traitements de données personnelles, le concept de « *big data* » étant beaucoup plus large et pouvant concerner, par exemple, des projets dans les domaines scientifique et environnemental[7]. Toutefois, en pratique, de nombreuses applications concrètes du « *big data* » touchent directement ou indirectement à des activités ou des comportements humains.

S'agissant des traitements soumis à la loi **«** informatique et libertés », les développements liés au « *big data* » sont aujourd'hui majoritairement valorisés dans le **domaine du e-commerce et du marketing**, pour des usages assez classiques de l'analyse des données, permettant de mieux connaître les attentes et les comportements des consommateurs.

Les grands éditeurs de logiciels investissent dans les technologies « *big data* ». Ils veillent cependant à préserver les revenus qu'ils tirent des solutions plus traditionnelles de bases de données. Ils ont tendance à privilégier l'évolution en utilisant beaucoup de mémoire vive pour gagner en performance, sans remettre en cause leur architecture. Cette « remise à niveau » des bases de données fonctionne mais s'avère onéreuse.

Concrètement, il peut s'agir par exemple pour une marque de surveiller ce qui se dit sur elle en analysant des grandes quantités de messages et commentaires laissés sur les médias sociaux. La publicité en ligne mobilise également ces techniques pour affiner le ciblage comportemental en fonction du profil de l'utilisateur, avec des adaptations en temps réel du

---

[5] Dans les années 2000, Google, en sa qualité de moteur de recherche, affirmait son leadership en proposant un service incomparable à celui de ses concurrents. En 2003, Google publie un premier papier sur le Google File System, et révèle ainsi les premiers secrets de son succès. En 2004, le fonctionnement de MapReduce est découvert, et l'année suivante, deux employés de Yahoo (Doug Cutting et Michael Cafarella), inspirés par les travaux de Google, créent Nutch Search Engine, qui deviendra Hadoop. C'est la naissance du « *big data* ». En 2006, Yahoo lègue le projet à Apache, qui reste depuis le cœur névralgique d'Hadoop.

[6] Voir, par exemple, le programme UN Global Pulse « Big data for development », permettant d'évaluer en temps réel le degré de réussite d'une action humanitaire et de l'adapter, l'améliorer ou la recadrer en fonction de cette évaluation.
http://www.unglobalpulse.org/projects/BigDataforDevelopment

[7] Les applications « *big data* » sont nombreuses dans le domaine scientifique et environnemental : géologie, météorologie, par l'intermédiaire de capteurs permettant de surveiller et de prévoir le déclenchement de phénomènes naturels. Par exemple, au Japon, la société Futjitsu a équipé une exploitation agricole d'outils permettant d'optimiser la fertilisation des terres arables. Des capteurs prennent en temps réel des mesures sur la nature des terrains (PH, humidité, etc.) ces données sont ensuite utilisées pour prévoir le meilleur moment où les fertiliser.

message affiché, et le recours à des techniques de « *re-targeting* » ou reciblage publicitaire[8] pour améliorer l'efficacité d'une campagne. La société française Criteo est ainsi citée en exemple dans le Guide du « *big data* » 2014-2015 pour sa nouvelle approche de la vente autour des concepts de RTB – *real time biding* – de la géolocalisation, des cookies, du suivi du parcours client, de l'individualisation et de l'optimisation du CRM (« *customer relationship management* »).

Certaines entreprises vont plus loin en intégrant la logique « *big data* » à leur mode de fonctionnement et leur organisation logistique. Par exemple, début 2014, la société Amazon annonçait non seulement être en mesure de connaître le prochain acte d'achat de chacun de ses clients mais, surtout, la société indiquait qu'elle avait transformé sa chaîne de logistique pour tenir compte des résultats de ces analyses, préparant les commandes pour l'expédition avant même que celles-ci aient été passées par les clients[9].

Le **secteur automobile** prend le virage du « *big data* », avec l'installation de capteurs dans les nouveaux véhicules. Le secteur de l'assurance suit aussi cette tendance, la société Axa aurait par exemple débloqué un budget de 800 millions d'euros sur trois ans pour se lancer dans le domaine[10]. Les offres « *Pay as you drive/Pay how you drive* » émergent sur le marché, permettant aux compagnies d'assurance d'obtenir, en temps réel, des informations réelles sur la conduite de leurs assurés. L'offre de PHYD d'Axa, examinée par la Commission dans sa phase expérimentale, est désormais commercialisée par Direct Assurance auprès d'un échantillon de consommateurs. Elle reprend les recommandations formulées par la Commission. Des recherches sont également menées en matière d'accidentologie, par exemple par le laboratoire commun à Renault et PSA Peugeot Citroën[11].

Le « *big data* » se développe également dans les **secteurs de la santé, de la recherche et de la génétique**. Un exemple emblématique du « *big data* » est le projet « *Google Flu Trends* », traitement mis en œuvre par la société Google et reposant sur des algorithmes à visée prédictive. L'objectif est de parvenir à la modélisation et à la prédiction de la propagation de pandémies de grippe et de dengue à partir de l'analyse des requêtes de millions d'internautes. Lancé en 2008, Google a mis fin en août dernier au projet sous sa forme initiale, pour communiquer les estimations non plus au grand public mais à des institutions spécialisées partenaires[12]. Google vient également de passer un accord avec la Food and Drug Administration aux États-Unis pour repérer les effets secondaires inconnus des médicaments, l'idée étant que les requêtes anonymisées des utilisateurs du moteur de recherche permettent notamment de repérer des effets secondaires qui apparaissent tardivement après le début du traitement et qui sont parfois sous-estimés par les dispositifs actuels de pharmacovigilance[13]. Des sociétés, comme OpenHealth Company, se sont également spécialisées dans l'application de ces nouvelles technologies au domaine de la santé pour anticiper et gérer les crises sanitaires, améliorer l'efficience des politiques et systèmes de santé, etc.

---

[8] Lorsqu'un internaute clique sur un produit, consulte sa fiche sur un site de vente en ligne mais ne l'achète pas, ces techniques permettent de lui proposer ultérieurement ce même produit dans des bannières publicitaires intégrées aux prochains sites qu'il consultera.

[9] Guide du Big data 2014-2015 http://www.bigdataparis.com/guide/2014-2015/BD14-15_Guide_BD_14136.pdf

[10] « Les enjeux de la révolution big data », http://lexpansion.lexpress.fr/high-tech/sondage-exclusif-les-enjeux-de-la-revolution-big-data_1700110.html, juillet-aout 2015.

[11] http://www.journaldunet.com/solutions/analytics/big-data-chez-psa-et-renault.shtml

[12] http://www.tdg.ch/sante/sante/Google-Flu-Trends-vaincu-par-ses-poussees-de-fievre/story/28554893
http://googleresearch.blogspot.fr/2015/08/the-next-chapter-for-flu-trends.html

[13] Administrateur général des données, « Rapport au Premier ministre sur la gouvernance de la donnée 2015. Les données au service de la transformation de l'action publique », décembre 2015.

Pour citer un autre exemple concret mis en œuvre dans ce secteur, au Canada, des chercheurs essaient de localiser les infections chez les bébés prématurés avant même que les symptômes n'apparaissent, en exploitant plus de 1 000 données par seconde (pouls, tension, respiration, niveau d'oxygène dans le sang, etc.)[14]. Ils ont ainsi réussi à établir des corrélations entre des dérèglements mineurs et des maux beaucoup plus sérieux. Cette technique vise à permettre aux médecins d'intervenir en amont pour sauver des vies, en partant du principe que, lorsque la vie de nourrissons est en jeu, « il est plus utile d'anticiper ce qui pourrait se produire que de savoir pourquoi ». Avec le développement des objets connectés, on assiste également à l'émergence d'une nouvelle approche des questions de santé.

La mise en œuvre de ces technologies dans les domaines de la **gestion des villes** (*Smart Cities*) est encore balbutiante. Les premières applications portent sur la prévision de trafic[15]. Le traitement Flux vision de la société Orange permet ainsi de convertir en temps réel des millions d'informations techniques provenant du réseau mobile en indicateurs statistiques, pour analyser la fréquentation de zones géographiques et les déplacements de population, qui peuvent être utilisés dans le domaine du tourisme, de l'aménagement du territoire, du trafic routier ou du commerce. L'offre repose sur un procédé d'anonymisation développé par Orange en concertation avec la CNIL, qui supprime toute possibilité d'identifier les clients de l'opérateur, grâce notamment au taux de collision choisi entre les identifiants hachés. La Ville de New York a identifié les immeubles à risque d'incendie pour aider les pompiers dans leur action de prévention (passant ainsi en quelques semaines, de 10 % de contrôles positifs à 78 % de contrôles positifs). Un projet a également permis d'améliorer l'offre de transports dans la ville d'Abidjan. Dans le cadre du programme « *Data4Development* », Orange a mis à disposition de chercheurs des données anonymisées concernant la localisation de ses abonnés en Côte d'Ivoire et au Sénégal. Ces données ont notamment été utilisées pour améliorer le réseau de transport de la ville d'Abidjan, en augmentant le nombre de lignes et en améliorant les parcours et les temps d'attente. Les données avaient été utilisées au préalable pour déterminer les flux origine-destination au sein de la ville et converties en parcours au niveau du réseau de transport existant.

Les projets « *big data* » se développent également pour la maîtrise des consommations énergétiques, comme les offres Cofely services du groupe Engie[16], ou le déploiement de projets de « bâtiments intelligents » dans lesquels toutes les données relatives à l'appartement et aux habitudes de ses résidents seraient analysées pour leur proposer des services à valeur ajoutée (pilotage de la consommation énergétique, proposition de produits et services en fonction des habitudes de consommation et de vie, gestion du parking, etc.).

---

[14] http://www.monde-diplomatique.fr/2013/07/CUKIER/49318

[15] Un système de gestion du trafic a ainsi permis à la ville de Stockholm de réduire de 20% les embouteillages, de 12% les émissions polluantes et d'augmenter le recours aux transports publics. Le système repose sur une taxe affectée aux véhicules qui traversent la ville aux heures de pointe. Les véhicules sont identifiés automatiquement par des systèmes mêlant laser et caméras, le montant de la taxe est fonction de l'heure de passage. Ce dispositif repose sur la manipulation de masses de données diverses (vidéos, OCR, capteurs) analysées en temps réel.

[16] Cofely Services a développé il y a deux ans l'offre VERTUOZ, qui permet à ses clients gestionnaires de parcs immobiliers de monitorer leurs consommations en temps réel et d'amener la « *Business Intelligence* » jusque chez ses clients. 3 niveaux d'utilisation de la donnée : 1°) système de suivi de la consommation d'énergie pur : connaître ses consommations réelles au temps-T et en garder une trace ; 2°) intégrer des fonctionnalités de « *Business intelligence* » pour mieux comprendre et exploiter ces données ; 3°) approche « *big data* » en intégrant des données externes, des volumes de données importants et un traitement en temps quasi réel. Le concept est qu'il est certes intéressant pour le client de savoir combien il consomme, mais cette donnée prend d'autant plus de sens lorsqu'il est possible de la comparer de manière fine avec les consommations de structures similaires à celle du client. Pour Cofely services, le « *big data* » est également un moyen de mieux connaître son client et donc de mieux le conseiller (un seul client pouvant représenter plus de 1000 sites, chaque site ayant ses caractéristiques de consommation propres). Permet aussi l'identification de typologies de clients, qui permettent à la société de comprendre de manière beaucoup plus fine ses clients. L'entreprise va également plus loin en lançant une société commerciale, DEEPKI, développant une cartographie permettant de détecter les gisements d'économies d'énergie grâce aux données existantes du secteur privé, des ministères, des collectivités, et aussi à l'aide des données ouvertes. Ces informations sont structurées, segmentées en classes de bâtiments à l'aide d'algorithmes, puis traitées avec un moteur d'inférence.

Le « *big data* » trouve à s'appliquer dans la **lutte contre la fraude**. La société Data&Data consulting a ainsi lancé l'outil *Brand Watchdog*, qui permet de lutter contre la contrefaçon notamment dans les domaines pharmaceutique et du luxe. Le principe est que cet outil va « *screener* » internet et les réseaux sociaux à la recherche d'objets contrefaits, en utilisant des algorithmes spécifiques adaptés à chaque secteur et à chaque typologie de produits. Pour le luxe, le premier constat était que les techniques traditionnelles qui consistent à comparer un certain nombre de points (comme la photographie et le prix) ne suffisaient pas à identifier efficacement les sources frauduleuses, alors que l'analyse de la source du site web, son service client, sa licence, etc., est souvent révélatrice. En moyenne, et pour chaque source identifiée, plus de 300 points de mesure sont analysés, à partir d'une douzaine de médias sociaux les plus fréquentés et de l'ensemble du web. L'idée est d'identifier les réseaux, de tracer la cartographie et de voir les nœuds avec les sites cachés derrière une filière de distribution d'articles contrefaits. L'action à suivre est ensuite définie au cas par cas, selon la stratégie de la marque, qui peut soit souhaiter dénoncer l'hébergeur, avertir les autorités, ou simplement notifier le propriétaire du site.

La Commission a par ailleurs rendu un avis (délibération n° 2014-045 du 30 janvier 2014) sur un projet d'arrêté portant création par la direction générale des finances publiques d'un traitement automatisé de lutte contre la fraude fiscale dénommé « Ciblage de la Fraude et Valorisation des requêtes ». Ce traitement, réalisé à titre expérimental, était basé sur des techniques dites de « *datamining* » (exploration de données) et alimenté par le croisement d'une dizaine de bases de données. Il visait à une modélisation des comportements frauduleux afin de mener des actions de prévention, de recherche, de constatation ou de poursuite d'infractions pénales ainsi que des opérations de constatation ou de poursuite de manquements fiscaux.

Le « *big data* », enfin, est au cœur de l'économie des plateformes collaboratives.
Qu'il s'agisse de mobilité (Uber, Blablacar) ou d'hébergement de courte durée (Airbnb, Bedicasa), ces sociétés n'existeraient pas sans les données, tant la place qu'elles occupent est centrale dans leurs modèles d'affaires. Les données fournissent d'abord l'ingrédient indispensable aux échanges : la confiance, par l'analyse des transactions et la notation réciproque des intervenants. Les données sont également utilisées pour améliorer en permanence le service : Uber est capable de prédire les zones où la demande sera la plus forte à un instant T, et donc d'encourager les chauffeurs à s'y rendre ou de modifier les prix en fonction de l'offre et de la demande. Airbnb analyse en permanence les recherches et l'historique d'un client, et sait quel bien il faut lui présenter en premier lieu pour répondre à ses goûts. Il peut aussi aider les hôtes à fixer le meilleur tarif de location (tout au moins celui qui maximise le revenu de la plateforme).

B. Eléments de définition et caractéristiques communes

Plus encore qu'une technologie, le « *big data* » peut être défini comme un « programme », un « régime d'action » (« *data-to-action* ») : recueillir des monceaux de données, les croiser, mesurer quotidiennement, voire en temps réel, les effets d'une décision, tester en permanence l'efficacité d'une action et la modifier dès qu'elle ne semble plus produire les effets désirés.

La délimitation du territoire des « mégadonnées » ou des données massives[17] reste incertaine.

[17] Journal officiel du 22/08/2014 ou http://www.culture.fr/franceterme.

Si la ligne de partage entre « *big data* » et traitements plus traditionnels ne va pas de soi, il est possible d'identifier des caractéristiques communes à ces projets, et qui pourraient permettre à la CNIL de déterminer lorsqu'elle est en présence d'un traitement « *big data* ».

A cet égard, votre rapporteur considère qu'il est opportun d'adopter une conception large de ce qu'est le « *big data* », dans la mesure où cette conception est celle retenue par la majorité des acteurs et qu'elle permet de répondre à leurs attentes.

Une définition axée sur le développement de nouveaux moyens technologiques de traitement des données

La **définition française** du terme « *big data* » a été adoptée par la Commission générale de terminologie et de néologie le 22 août 2014. La traduction officielle est « mégadonnées » et la définition retenue est la suivante : « **données structurées ou non dont le très grand volume requiert des outils d'analyse adaptés** ».

Concrètement, le « *big data* » est un terme qualifiant des traitements mettant en œuvre des volumes importants de données.

Ces données peuvent être statiques, de l'ordre du pétaoctet, soit 1 million de milliards d'octets, ou dynamiques, en temps réel, par exemple plusieurs milliards d'événements par jour.

A l'ère du « *cloud computing* », et de l'accès de tous à des capacités de calcul sans précédents, cette définition rend compte de l'accroissement du volume de données numériques produites depuis les années 2010, conduisant à ce que nous stockions annuellement quelques mille milliards de gigaoctets d'informations.

La définition de la Commission générale de terminologie et de néologie rend ainsi compte du développement de nouvelles technologies et outils d'analyse qui permettent aujourd'hui de traiter des volumes conséquents de données, qu'elles soient ou non structurées[18].

Cette définition rejoint la « règle des 3V ». Depuis l'étude du cabinet Gartner, réalisée en 2001, il est effectivement communément admis que le « *big data* » répond à **trois caractéristiques connues sous le nom de règle des 3V : Volume, Vitesse ou Vélocité et Variété.** La notion de « *big data* » renvoie ainsi à l'idée de traiter des Volumes de données considérablement supérieurs à ceux traités auparavant, à une Vitesse incomparable, en intégrant une Variété de données beaucoup plus riche. L'enjeu principal de ces traitements est la création de valeur. Aussi, ces 3V sont généralement augmentés par d'autres V, tels que Valeur, Véracité, ou encore Visibilité, etc.[19].

L'apparition de la notion de « *big data* » est ainsi intrinsèquement liée à la composante technologique, les innovations technologiques ayant profondément modifié la manière dont les entreprises et les individus produisent, transmettent, stockent et analysent des données. Cette notion recouvre ainsi une combinaison de progrès technologiques, d'innovations d'usage voire d'évolutions sociales et culturelles concernant le partage d'informations. Les données sont omniprésentes et ont acquis une importance centrale, modifiant la manière dont les entreprises appréhendent, considèrent et utilisent les données, éléments clés de l'augmentation de la connaissance et de la richesse.

[18] Des traitements « *big data* » ne peuvent être mis en œuvre à partir d'outils traditionnels. De nouvelles technologies se sont ainsi développées, telles que Hadoop ou MapReduce.

[19] Guide du Big data 2014-2015 http://www.bigdataparis.com/guide/2014-2015/BD14-15_Guide_BD_14136.pdf. Voir également le site internet data-business.fr, à l'adresse suivante http://www.data-business.fr/big-data-definition-enjeux-etudes-cas/.

S'agissant tout d'abord du volume, l'émergence du « *big data* » est en premier lieu liée à une augmentation de la masse des données, phénomène appelé « *datafication* » par les spécialistes : il est estimé que **90 % des données récoltées depuis le début de l'humanité ont été créées durant les deux dernières années**[20].

L'essor des *smartphones* et des objets connectés contribue activement à ce « déluge » de données. À la fin de l'année 2015, le nombre de *smartphones* dans le monde était estimé à environ 2 milliards d'unités[21]. S'agissant des objets connectés, 15 milliards sont déjà sur le marché et des estimations prévoient qu'il y en aurait 75 milliards en 2020[22]. En effet, d'après des estimations, chacun d'entre nous posséderait en moyenne 8 objets connectés à titre personnel en 2018 et, en 2020, nous en aurions déjà 10[23]. La multiplication des objets connectés contribue ainsi à la création de quantités gigantesques de données personnelles, considérées comme le véritable or noir du XXIe siècle.

Ainsi, les entreprises disposent d'un volume considérable de données, 70 % des données étant créées par les individus mais 80 % d'entre elles étant stockées et gérées par les entreprises. Aujourd'hui, on parle du stockage de zettaoctets[24] de données alors qu'il y a à peine 10 ans on parlait de mégaoctets, stockés sur des disquettes. Cette « *datafication* » est ainsi couplée à une **révolution dans le stockage**, avec la démocratisation du « *cloud computing* ». Des évolutions considérables ont également eu lieu en matière de **capacités de traitement** des données en des temps réduits. C'est le V de Vitesse. Des données qui étaient auparavant traitées en plusieurs jours parfois peuvent à présent être traitées en quelques heures voire minutes. L'impact économique et le gain d'efficacité sont évidents.

Enfin, grâce aux progrès des techniques d'analyse et des outils de visualisation de données, les **données traitées** dans le cadre du « *big data* » **n'ont plus besoin d'être structurées** selon des critères communs. Le « *big data* » permet de traiter tout type de données, dans sa forme originelle : images, sons, vidéos, commentaires de blogs, logs, simples clics sur un site internet, données issues de capteurs, données de géolocalisation, etc.

Cette variété se matérialise également par des traitements ayant recours à des sources de données diverses : sources internes[25] ou externes[26] à l'organisme, provenant en partie de « nouvelles » sources de données[27].

---

[20] Selon les estimations de la société IBM. Guide du Big data 2014-2015 http://www.bigdataparis.com/guide/2014-2015/BD14-15_Guide_BD_14136.pdf et http://www-01.ibm.com/software/fr/data/bigdata/

[21] http://www.cbnews.fr/digital/pres-de-2-milliards-de-smartphone-dans-le-monde-fin-2015-a1016742.

[22] http://www.lemonde.fr/idees/article/2015/06/17/faisons-du-big-data-une-chance-pour-l-europe_4655737_3232.html.

[23] Guide du Big data 2014-2015 http://www.bigdataparis.com/guide/2014-2015/BD14-15_Guide_BD_14136.pdf.

[24] Correspond à 10^{21} octets.

[25] Les données internes à l'organisme peuvent par exemple être ses bases de données, les courriels, les documents, tous les historiques de processus métiers (logs), et tout autre type de données structurées, semi-structurées ou non-structurées que l'organisme produit et stocke. Dans certains cas, ces données sont utilisées dans le cadre des traitements « *big data* » pour d'autres finalités que celle initialement prévue (par exemple pour l'aménagement d'un territoire via les métadonnées d'appels d'un opérateur mobile : traitement Flux vision d'Orange).

[26] Les données externes peuvent être des bases de données externes (publiques, fournisseurs de données, partenaires commerciaux, etc.) ou des contenus échangés sur les réseaux sociaux ou publiés en ligne (par exemple pour analyser les sentiments des usagers d'un service à travers les commentaires postés sur les réseaux sociaux ou pour modéliser l'évolution des tendances boursières à partir des messages des réseaux sociaux).

[27] Qu'il s'agisse de données internes ou externes, on observe également que de « nouvelles » sources de données viennent alimenter le « *big data* », comme les données issues de la navigation Internet, collectées par exemple via les *cookies* lors d'une visite sur le site internet de l'organisme ou les données collectées dans le cadre des services proposés à la personne (transaction bancaire, email…). Figurent également parmi ces nouvelles sources de données les données, géolocalisées ou non, transmises par les *smartphones*, notamment par le biais des applications, et les objets connectés, tels que les compteurs communicants, les boitiers numériques de métrologie des automobiles, et les objets liés au « *quantified self* ».

D'autres V sont à présent associés au « *big data* », renvoyant à des notions de **Valeur,** de **Véracité**, etc. Ils traduisent le besoin de disposer de données fiables et pertinentes, qui permettent de donner suffisamment de sens et d'intérêt économique aux analyses réalisées. Ces notions renvoient, non pas à la seule innovation technologique qui a permis d'améliorer des capacités de traitement, mais au **véritable enjeu du « *big data* » : la création de valeur** par le biais d'une nouvelle approche de la donnée.

Une définition liée à une nouvelle approche de la donnée

Il peut sembler légitime de se demander si le « *big data* » est un effet de mode, une révolution ou un nouveau paradigme. Le terme « *big data* » est incontestablement un « *buzz word* », de nombreuses sociétés souhaitant mettre en avant le fait qu'elles font du « *big data* », que leurs projets correspondent véritablement, ou non, aux définitions et critères identifiés par la doctrine. Toutefois, au-delà de cette volonté d'affichage d'une approche moderne du traitement des données par les sociétés concernées, il semble indéniable que l'émergence du « *big data* » constitue un phénomène en soi, impliquant une approche nouvelle du traitement des données.

Si l'on associe à Internet la notion de révolution numérique, c'est le terme de paradigme qui est plus volontiers rattaché au « *big data* ». Le « *big data* » nous ferait entrer dans une **nouvelle ère cognitive**, l'approche « *big data* » n'impliquant donc pas seulement des évolutions technologiques mais une démarche nouvelle, visant à faire des données un mode de décision, un actif stratégique et une façon de créer de la valeur. Il s'agirait d'un véritable renversement de paradigme d'organisation, dans lequel l'organisme serait guidé par les données[28].

James Gray[29] a émis l'hypothèse que, sous l'impulsion du « *big data* », la science serait en train de changer de paradigme. Jusqu'à présent fondée sur une capacité d'abstraction et de généralisation, récemment aidée des ordinateurs, permettant l'élaboration de modèles mathématiques capables d'interpréter les données, **ce serait dorénavant les données elles-mêmes qui constitueraient le modèle**. Si nous disposons en permanence de données de la réalité, il n'est plus nécessaire de construire un modèle de cette réalité.

S'il semble prématuré de qualifier un phénomène si récent, on ne peut que constater que le « *big data* » s'inscrit dans la continuité de transformations de la société poussées, dictées par les technologies, qu'elles prennent la forme d'évolution, de mutation ou de rupture.

En tout état de cause, par l'analyse de quantités immenses de données, les traitements « *big data* » ouvrent de nouveaux horizons à la prise de décision, puisqu'ils permettraient de trouver des corrélations et informations qui n'étaient pas anticipées à l'origine par ceux qui les avaient initiés. Cet apport fondamental de cette technologie est la « sérendipité »[30].

Le « *big data* » promettrait ainsi une amélioration de tous les processus décisionnels, **voire une capacité d'anticipation par l'élaboration de modèles prédictifs**. Deux raisons principales sont avancées pour soutenir cette hypothèse.

---

[28] http://www.data-business.fr/big-data-definition-enjeux-etudes-cas/.

[29] L'évolution de la science a connu trois paradigmes : le premier, apparu il y a plusieurs milliers d'années, a permis l'avènement de la science empirique, fondée sur les représentations anthropocentriques. Le second a trait à la science analytique, reposant sur la capacité d'abstraction et de modélisation, dont sont issues les théories de Newton à Einstein. Le troisième paradigme nait de l'adossement de l'informatique à la science. http://research.microsoft.com/en-us/collaboration/fourthparadigm/

[30] Protection des données personnelles et Big Data : inconciliables, vraiment ?, F. Coupez, Avril 2015 (http://www.silicon.fr/protection-donnees-personnelles-big-data-inconciliables-114312.html).

La première est que les processus décisionnels actuels ont été élaborés à partir d'une sélection d'informations intelligemment collectée. L'idée qui prévalait est qu'on peut inférer une tendance à partir d'un échantillon. Alors que cette approche, qui serait par essence limitée, ne pourrait jamais rendre compte de l'ensemble des informations, le « *big data* » permettrait de s'affranchir de cette limitation.

La seconde raison tiendrait au fait que le « *big data* » ne chercherait pas à interpréter des données. Il se contenterait d'établir des corrélations (*si A alors B*), sans expliquer la causalité des événements (*Pourquoi A ?*). On passerait d'une logique déductive à une logique par inférence. Il s'agirait de dépasser la relation de cause à effet pour se concentrer sur la caractérisation du résultat. En d'autres termes, on chercherait des corrélations sans pour autant déterminer les facteurs de corrélation et on agirait à partir de l'observation empirique d'un phénomène[31]. Cette nouvelle approche suscite des inquiétudes, les données collectées pouvant être affectées de biais et les corrélations pouvant ne correspondre à aucun lien de causalité. **Antoinette Rouvroy** (Université de Namur) alerte ainsi sur le **risque de « gouvernementalité algorithmique »** dû au renoncement de la compréhension des causes et à une approche fondée sur la rationalité au profit d'une prise de décision fondée sur une « pseudo-objectivité machinique ».

Et les données d'aujourd'hui seraient ainsi quasi-systématiquement collectées car elles procureraient demain un avantage concurrentiel dans l'élaboration des modèles, notamment prédictifs. Il s'agirait dès lors d'exploiter des jeux de données plus importants et dont la valeur informationnelle serait plus ténue. Le « *big data* » permettrait d'exploiter des données moins signifiantes - pauvres en information - dont les coûts de traitements rendaient auparavant leur utilisation non pertinente. On devrait dès lors assister à une multiplication des occasions d'interconnexion ou de rapprochement de données, pour pouvoir en exploiter la valeur.

Toutefois, exploiter le potentiel du « *big data* » nécessite de disposer d'un savoir-faire dans plusieurs domaines ardus : les mathématiques (l'algèbre des ensembles notamment), les statistiques et les probabilités, les systèmes d'information, et enfin les techniques algorithmiques, avec son corollaire, la programmation. A ces domaines génériques, il faut ajouter la connaissance du métier ciblé. « Faire parler » les données est donc une affaire d'experts.

Se développe ainsi la profession de « *datascientist* », terme désignant les experts qui ont non seulement la capacité d'analyser les données et de développer du code informatique mais qui disposent surtout des compétences pour faire émerger de cette analyse de nouveaux usages. Ce terme désigne ainsi des personnes qui analysent des données non pas pour produire des rapports ou des statistiques mais afin d'améliorer le produit ou le service de l'organisation pour laquelle ils travaillent. Les grands acteurs du numérique développent ainsi de nouveaux usages des méthodes statistiques et des algorithmes d'apprentissage statistique, permettant par exemple à un réseau social d'utiliser les données des utilisateurs pour prédire qui sont ses amis ou ses contacts professionnels, à un service de vidéo à la demande de prédire les films que l'utilisateur aime ou à un site de vente en ligne de prédire les produits qu'un client est susceptible d'acheter[32].

[31] Dans l'exemple cité précédemment, les chercheurs canadiens essaient de localiser les infections chez les bébés prématurés avant même que les symptômes n'apparaissent, en exploitant plus de 1 000 données par seconde (pouls, tension, respiration, niveau d'oxygène dans le sang…). Ils ont ainsi réussi à établir des corrélations entre des dérèglements mineurs et des maux beaucoup plus sérieux. Cette technique vise à permettre aux médecins d'intervenir en amont pour sauver des vies, en partant du principe que, lorsque la vie de nourrissons est en jeu, « il est plus utile d'anticiper ce qui pourrait se produire que de savoir pourquoi ».

[32] Administrateur général des données, « Rapport au Premier ministre sur la gouvernance de la donnée 2015. Les données au service de la transformation de l'action publique », décembre 2015.

C. Mise en perspective avec le contexte économique

Les analystes prévoient un taux de croissance annuel moyen mondial de 31,7 % du marché du « *big data* » (technologie et services) pendant la période 2011-2016. Ce marché dépasserait les 50 milliards de dollars d'ici 2018, avec un taux de croissance annuel moyen de 40 %[33]. En 2020, les données personnelles permettraient une création de valeur représentant 8 % du PIB européen[34].

Les technologies du « *big data* » sont largement l'affaire de jeunes entreprises. Les grands éditeurs de logiciels d'Informatique décisionnelle investissent dans ces technologies : ils veillent cependant à préserver les revenus qu'ils tirent des solutions plus traditionnelles de bases de données. Ils ont tendance à privilégier la remise à niveau de leurs solutions pour gagner en performance, sans remettre en cause leur architecture.

A l'échelle européenne, **l'adoption du « *big data* » par les entreprises françaises et européennes reste encore limitée**[35]**, mais devrait progresser**[36].

En France, selon un sondage réalisé durant l'été 2015 par Opinion Way auprès de 500 dirigeants et managers d'entreprises en France, seuls 18 % des dirigeants français recourent au « *big data* » dans leurs entreprises. 68 % considèrent ainsi que les entreprises de leur secteur sont en retard dans ce domaine. Les freins identifiés par les chefs d'entreprise sont variés, et ne concernent pas nécessairement la protection des données personnelles[37]. Pour 86 % des personnes interrogées, la notion de « *big data* » reste floue, et 34 % considèrent qu'il s'agit d'un effet de mode, qui n'apporte rien de nouveau[38].

Pour les entreprises qui se sont engagées dans la démarche « *big data* », les domaines dans lesquels les traitements sont mis en œuvre sont : les processus internes (83 %) ; la connaissance du client (78 %) ; la connaissance du marché (57 %) ; les nouvelles offres de produits ou services (57 %) ; la diminution des coûts (52 %) et l'innovation (26 %).

François BOURDONCLE, copilote du plan « *big data* » initié par le gouvernement en 2013 (Nouvelle France industrielle), estimait en novembre 2014, qu'il faudrait une quinzaine d'années pour que le « *big data* » gagne l'ensemble des secteurs de l'économie[39]. Il **y aurait, selon lui, urgence pour les grands groupes français**, pour ne pas être distancés par les

---

[33] http://www.zdnet.fr/actualites/l-avenir-du-big-data-nouveaux-concepts-et-forte-croissance-due-aux-metiers-39824090.htm

[34] http://lte.ma/enjeux-du-big-data

[35] « Les enjeux de la révolution big data », http://lexpansion.lexpress.fr/high-tech/sondage-exclusif-les-enjeux-de-la-revolution-big-data_1700110.html, juillet-aout 2015. Guide du Big data 2014-2015 http://www.bigdataparis.com/guide/2014-2015/BD14-15_Guide_BD_14136.pdf.

[36] Selon le cabinet IDC, le marché du « *big data* » en France s'inscrit dans les tendances de l'Europe de l'Ouest. A cette échelle, la croissance annuelle du marché (infrastructures, logiciels et services) serait de 24,6% entre 2014 et 2018. En France, en 2014, le marché « *big data* » était estimé à 285 millions d'euros. Il devrait atteindre 652 millions en 2018. http://www.zdnet.fr/actualites/big-data-le-marche-francais-devient-mature-39823546.htm

[37] Parmi les freins identifiés par les chefs d'entreprise, 50% estiment qu'ils sont liés à la complexité des infrastructures/matériels/logiciels ; 42% à l'absence de compétences internes ; 41% aux difficultés à collecter les données ; 39% à des moyens financiers insuffisants ; 39% à l'absence de maturité de l'entreprise ; 37% à la difficulté à gérer la confidentialité des données et 33% à l'absence de bénéfices identifiés.

[38] Il est intéressant de noter que la perception n'était pas exactement la même fin 2013, une étude réalisée par EMC indiquant que 74% des entreprises étaient convaincues de l'intérêt du « *big data* » pour leur activité, mais 41% d'entre elles n'avaient encore engagé aucune dépense sur le sujet, en raison notamment de la faible prévisibilité du retour économique de ces investissements. http://www.bigdataparis.com/guide/2014-2015/BD14-15_Guide_BD_14136.pdf. L'étude réalisée par le cabinet Markess auprès de 190 décideurs informatique et métiers en France montrent également un engouement plus fort pour le « *big data* », 57% des décideurs plaçant le « *big data* » dans leurs 3 principaux enjeux de gestion de l'information et 41% prévoyant d'exploiter leurs données non structurées d'ici à fin 2017. http://www.zdnet.fr/actualites/big-data-le-marche-francais-devient-mature-39823546.htm

[39] « Il y a urgence pour les grands groupes à expérimenter le « big data », Le Monde, 06/11/2014.

géants de l'Internet et par leurs concurrents étrangers, **à s'investir pleinement dans le « *big data* »**, en expérimentant et en s'associant à des start-up innovantes[40].

Cette dynamique en faveur du « *big data* » est générale et les pays émergents suivent également la tendance. Selon une étude menée par la société MarketsandMarkets, ces pays devraient rattraper leur retard en la matière dans les années à venir et représenter une part importante de la croissance du marché[41].

D. Des politiques publiques

Le **gouvernement américain** s'est activement engagé dans le domaine du « *big data* »[42]. Dès 2012, le gouvernement Obama annonçait la mise à disposition de 200 millions de dollars pour un fonds de recherche sur la thématique du « *big data* ». L'investissement du gouvernement américain ne s'est pas amoindri depuis, de lourds investissements étant réalisés notamment dans les domaines de la formation, de la recherche, de la sécurité nationale, de la santé ou encore des services publics[43]. L'organisme MeriTalk a publié en juin 2013 une étude estimant que le « *big data* » permettrait à l'Etat américain de réaliser 14 % d'économies, soit 500 milliards de dollars.

La Commission européenne, dans un communiqué du 2 juillet 2014[44], encourageait les pouvoirs publics à saisir au plus tôt le potentiel que représentent les données massives.

En France, une impulsion politique en faveur du développement du « *big data* » a été initiée avec le plan « *big data* » pour la Nouvelle France industrielle[45]. Le marché est estimé à 9 Md€ en France en 2020, et à un potentiel de 137 000 emplois crées ou consolidés. La

---

[40] Selon François BOURDONCLE, le contexte économique a changé et il faut en tenir compte. Le financement des start-up de l'Internet est aujourd'hui beaucoup plus conséquent et le développement de ces start-up géantes passe par la maîtrise, en amont, de la production et, en aval, de la relation client essentiellement grâce au « *big data* ». Cela conduit à terme à la création de « monopoles naturels », tels que le montrent aujourd'hui les exemples de Google, Facebook ou Apple.

[41] http://www.marketsandmarkets.com/PressReleases/big-data.asp

[42] Guide du Big data 2014-2015 http://www.bigdataparis.com/guide/2014-2015/BD14-15_Guide_BD_14136.pdf *(consulté le 16/09/15)*. Des études big data auraient par ailleurs été utilisées, durant la campagne présidentielle, pour permettre la réélection du Président (Voir l'émission Spécial investigation *www.youtube.com/watch?v=XBbzvxfzKgI*).

[43] A titre d'exemple, le département de la Défense compte une dizaine de projets big data et bénéficie de plus de 250 millions de dollars d'investissement annuel. Le programme ADAMS, entre autres, a pour objectif de repérer des comportements anormaux, des changements d'attitude préoccupants chez un soldat ou un citoyen américain. Le Projet Mind's Eye a pour objet d'améliorer les performances de reconnaissance vidéo et d'analyse automatisée. http://www.bigdataparis.com/guide/2014-2015/BD14-15_Guide_BD_14136.pdf

[44] http://europa.eu/rapid/press-release_IP-14-769_fr.htm.

[45] http://www.economie.gouv.fr/files/files/PDF/nouvelle-france-industrielle-sept-2014.pdf. Le plan Big data pour la Nouvelle France industrielle, porté par François BOURDONCLE et Paul HERMELIN, et dont la feuille de route a été validée en juillet 2014, est la première réelle impulsion économique émanant de l'Etat visant à développer l'écosystème « *big data* » en France. Des grandes entreprises ont été choisies pour participer aux travaux du plan : Orange, La Poste, GDF Suez, Alstom, AXA, la Société générale, Covéa (le groupe MMA, GMF et MAAF), etc. Les actions du plan se déclinent en trois axes :

- Développement de l'écosystème « *big data* » en France : formation de « data scientists » ; accès des start-up aux données et aux infrastructures ; soutien au financement et à l'accélération des start-up ; observatoire des usages.
- Lancement d'initiatives sectorielles sur le « *big data* » : favoriser la diffusion des technologies du « *big data* » dans le secteur privé ; moderniser l'action publique grâce à l'utilisation du « *big data* ».
- Evolutions de la réglementation.

Sur ce dernier point, il était prévu, au cours du second semestre 2014 d' « adapter le contexte réglementaire et législatif pour permettre le développement d'une filière Big Data », pour notamment:

1. « permettre des certifications de processus industriels afin de favoriser l'utilisation du Big data »,
2. « étudier l'opportunité d'adapter le contexte réglementaire et législatif ».

seconde phase de la Nouvelle France Industrielle a été lancée le 18 mai 2015 (projet Industrie du futur)[46].

Le rapport Lauvergeon[47], rendu le 11 octobre 2013, dont l'objet était de placer la France à la pointe de l'innovation, identifiait le « *big data* » comme priorité à l'horizon 2030 et comme secteur clé pour l'avenir économique de la France[48]. La Commission 2030 préconisait notamment un « droit à l'expérimentation », qui viserait à permettre aux acteurs du numérique de rester compétitifs en suivant le rythme des évolutions technologiques de leur secteur, sans être freinés par des instances et des contraintes réglementaires. Il s'agirait de trouver un juste équilibre entre libre innovation et protection des données personnelles.

Deux rapports récents, du CGEIET[49] et de l'Administrateur général des données (AGD)[50], plaident également pour l'adoption d'une nouvelle approche de la donnée au sein de l'administration, favorisant sa valorisation pour le développement d'usages innovants.

II. PRESENTATION DES PROBLEMATIQUES ET DES ENJEUX « INFORMATIQUE ET LIBERTES » SOULEVES PAR LE « *BIG DATA* »

Votre rapporteur est favorable à une recherche d'équilibre dans le développement du « *big data* », permettant de concilier innovation dans l'usage des données et respect des principes fondamentaux de la protection des données personnelles, ce respect devant rester un impératif. Or, le « *big data* » est parfois présenté comme difficilement conciliable avec la protection des données personnelles.

Comme indiqué précédemment, le « *big data* » n'implique pas nécessairement de traitement de données personnelles. Cependant, de nombreuses applications concrètes touchent directement ou indirectement à des activités ou des comportements humains et impliquent le traitement de données personnelles.

Le « *big data* » offre la possibilité d'une **connaissance plus fine de populations ciblées** et, le cas échéant, la **construction de modèles prédictifs de comportements** - voire de prise de décision – grâce, comme nous l'avons vu :

---

[46] http://www.economie.gouv.fr/files/files/PDF/industrie-du-futur_dp.pdf. Cette seconde phase capitalise sur le travail accompli au cours des 18 derniers mois par les équipes des 34 plans industriels. Le projet Industrie du futur est forgé autour de 9 solutions industrielles pour 9 marchés prioritaires, dont ceux de l'économie des données, des objets intelligents, de la confiance numérique, des villes durables ou encore de la médecine du futur.

[47] http://www.elysee.fr/assets/pdf/Rapport-de-la-commission-Innovation-2030.pdf

[48] La commission 2030 préconisait 5 leviers d'actions : l'open data comme accélérateur d'innovation, la valorisation des données publiques (au sein des administrations), la mise à disposition des ressources technologiques au sein d'un centre à destination des start-up, l'aide à l'export et le « droit à l'expérimentation ».

S'agissant de l'open data, le gouvernement français soutient également son développement. Il s'agit d'un des champs de bataille d'Axelle LEMAIRE : « *toute donnée publique doit être ouverte par défaut. Et s'il y a fermeture, il faut qu'elle soit expliquée, justifiée et réversible* ». L'idée qui sous-tend le développement de l'open data est que l'ouverture des données permet la création de valeur. Une information détenue par un acteur d'un certain secteur peut, si elle est partagée, permettre à un autre acteur de développer une innovation, un service, une analyse. La France est loin d'être en retard sur le sujet de l'open data. L'action d'Etalab, l'engagement de l'Etat et la collaboration de nombreux grands groupes ont permis d'ouvrir un grand nombre de données et de créer diverses start-up et services innovants. Actuellement, 21 345 jeux de données sont disponibles sur la plate-forme gouvernementale dédiée à l'open data « data.gouv.fr » (consultation début janvier 2016).

[49] En juillet 2015, le Conseil général de l'Economie, de l'Industrie, de l'Energie et des Technologies (CGEIET) a remis au Ministre de l'Economie, de l'Industrie et du Numérique, à la Secrétaire d'Etat chargée de la réforme de l'Etat et de la simplification et à la Secrétaire d'Etat chargée du Numérique un rapport sur « Les meilleures pratiques du big data et de l'analytique dans l'administration ».

[50] Administrateur général des données, « Rapport au Premier ministre sur la gouvernance de la donnée 2015. Les données au service de la transformation de l'action publique », décembre 2015.

- au traitement de masse de données structurées comme non structurées, et issues de multiples sources dont le web social et les objets connectés ;
- à des algorithmes d'analyse sophistiqués.

De nombreuses voix se font entendre pour estimer que le « *big data* » **remet en cause les principes cardinaux de la protection des données personnelles.**

Il convient dès lors de s'interroger sur la portée de ces affirmations et sur les pistes de réflexion permettant de relativiser et dépasser cette présomption de contradiction.

A. Vie privée et « *big data* » : le débat américain

Deux rapports remis au Président Obama en 2014 s'attachent à mettre en lumière les répercussions du « *big data* » pour la vie privée.

Le premier, rédigé par John Podesta[51], s'il met l'accent principalement sur l'apport du « *big data* » aux problématiques de santé, d'éducation ou d'urbanisation, pointe aussi toute une série de risques : discrimination en termes de prix, de services ou d'opportunités, absence de contrôle des internautes sur les données utilisées à des fins de ciblage commercial ou publicitaire[52], « persistance des données »[53], risques associés à l'utilisation du « *big data* » pour la médecine prédictive[54]. Il émet, en outre, de sérieux doutes sur les méthodes d'anonymisation[55].

« Dans un monde où le coût de stockage de données a chuté et ou l'innovation future reste imprévisible, la logique de maximisation de la collecte de données est forte ». Les tendances à la collecte massive, à leur réutilisation et leur combinaison, les limites de l'anonymisation *« pourraient nous conduire à réexaminer de près la doctrine du « notice and consent » qui a été un pilier central de notre approche de la protection de la vie privée depuis plus de quatre décennies. Dans un contexte technologique de sur-collecte structurelle, dans laquelle la ré-identification est de plus en plus puissante que l'anonymisation, l'accent sur le contrôle de la collecte et sur la conservation des données personnelles, bien qu'important, pourrait ne plus être suffisant pour protéger la vie privée ».*

Ce point de vue est exprimé de manière nettement moins nuancée dans le rapport "Big Data & Privacy : A Technological Perspective" préparé par le Comité des conseillers en Science et Technologie (Pcast) du Président.

---

[51] "Big Data : Seizing Opportunities, Preserving Values"

[52] Aucune solution satisfaisante n'a encore été trouvée pour donner la possibilité à l'internaute que ses données ne soient pas utilisées à des fins publicitaires. Les projets de mise en place de "Do Not track System" ou "Opt-out" se sont révélés jusqu'à présent inefficaces.

[53] La multiplication et l'ubiquité des données rendent difficiles leur suppression : les données personnelles, qui sont par défaut confidentielles, restent ainsi enregistrées sur internet parfois à l'insu de l'utilisateur.

[54] Et notamment le fait que l'information collectée sur un individu, s'étend également aux personnes ayant potentiellement des gènes similaires (enfants ou famille proche). Le rapport fait état d'un cadre réglementaire peut-être mal adapté pour répondre aux développements futurs tout en favorisant les innovations Le rapport met, à cette occasion, en garde à propos d'un encadrement trop strict de la collecte des données dans le domaine de la recherche médicale.

[55] « Une autre réalité de gros volumes de données est qu'une fois que les données sont collectées, il peut être très difficile de garder l'anonymat. Bien qu'il existe des efforts de recherche prometteurs en cours pour masquer des informations permettant d'identifier les personnes au sein de grands ensembles de données, des efforts encore plus avancés sont à l'œuvre pour ré-identifier les données apparemment «anonymes». L'investissement collectif pour fusionner les données est plusieurs fois supérieur à l'investissement dans les technologies qui permettront d'améliorer la vie privée ».

Selon ce rapport, les réglementations doivent refléter ce qui est et ce qui n'est pas technologiquement réalisable. Il serait quasi impossible de faire appliquer une réglementation de la collecte à grande échelle. Une réglementation de la collecte, en outre, de données pourrait avoir des conséquences négatives, sur la croissance économique par exemple. La régulation devrait porter sur l'utilisation des données plutôt que sur leur collecte. Afin de pallier le manque de technologies appropriées, les experts du PCAST préconisent d'augmenter l'effort de R&D liée à la protection de la vie privée. Souhaitant faire des États-Unis le pays leader dans ce domaine, ils avancent le chiffre de 80 millions de dollars pour alimenter les fonds de recherche alloués à ce domaine au niveau des agences fédérales.

B. Enjeux « informatique et libertés » et « *big data* »

Ces objections rencontrent aussi un écho en Europe : les **principes de finalité et de pertinence** y sont présentés comme des freins à l'innovation et au déploiement du « *big data* ».

Selon cette analyse, la logique qui sous-tendrait le « *big data* » impliquerait qu'un maximum de données soit recueilli en amont et les usages qui pourraient en être faits seraient définis seulement ensuite. Une optique « *big data* » remettrait ainsi en cause le principe selon lequel seules peuvent être collectées et traitées les données strictement nécessaires à la poursuite de finalités déterminées, explicites et légitimes - ce qui suppose donc qu'elles aient été préalablement définies.

Dans cette même optique, le « *big data* » remettrait en question les principes liés à la **conservation des données personnelles** puisqu'il favoriserait le stockage sans finalité immédiate de données, misant sur une capacité de réutilisation pour des finalités inconnues, ce qui induit une tension avec l'idée d'une conservation des données corrélée à la finalité.

Le « *big data* » poserait également la question de la « gouvernementalité algorithmique », l'objectif du « *big data* » pouvant être d'établir des modèles prédictifs en détectant les corrélations, et ceci dans le but d'anticiper des situations voire de prendre des décisions collectives ou individuelles par le biais des analyses statistiques. Se poserait alors la question de sa compatibilité avec **l'article 10** de la loi[56], qui interdit toute décision, produisant des effets juridiques à l'égard d'une personne, prise sur le seul fondement d'un traitement automatisé destiné à définir le profil de l'intéressé ou à évaluer certains aspects de sa personnalité.

---

[56] Le futur règlement européen relatif à la protection des données personnelles contient également ce principe, tout en prévoyant des exceptions. Article 20 « *1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.*
1a. Paragraph 1 shall not apply if the decision :

a) is necessary for entering into, or performance of, a contract between the data subject and a data controller; or

b) is authorized by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
c) is based on the data subject's explicit consent
1b In cases referred to in paragraph 1a (a) and (c) the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision
3.Decisions referred to in paragraph 1a shall not be based on special categories of personal data referred to in Article 9(1), unless points (a) or (g) of Article 9(2) apply and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place »
Definition "profiling means any form of automated processing of personal data consisting of using those data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements". Voir également le considérant 58.

De façon connexe, se poserait également la **question de l'information des personnes** concernées par ces traitements, notamment pour l'utilisation des données issues de l'Internet, informations certes publiquement accessibles en ligne mais qui peuvent concerner des personnes identifiées, personnes qui bénéficient de **droits** au regard de la loi du 6 janvier 1978 modifiée.

En outre, un des enjeux principaux du « *big data* » serait lié aux **possibilités de ré-identification des personnes**. En effet, outre le fait que les outils du « *big data* » peuvent porter sur des données personnelles, ils peuvent également conduire à ce que des données anonymes à l'origine, par recoupement avec d'autres données, permettent de déduire plus d'informations sur les personnes, voire de les identifier ou les ré-identifier[57].

Dans un article publié en juin dernier, **Yann Padova**, ancien Secrétaire général de la CNIL, et **Axel Voss**, Député européen allemand, « Shadow rapporteur » du projet de règlement européen relatif à la protection des données personnelles, appelaient les autorités européennes à engager une « révolution copernicienne » pour répondre à la révolution technologique introduite par le « *big data* »[58]. Selon leur analyse, **les principes de finalité, d'information des personnes et de consentement**, tels qu'ils sont actuellement interprétés, **seraient difficilement conciliables avec le « *big data* »** et ne permettraient pas de libérer l'innovation et de favoriser la croissance numérique. Aussi, **ils plaident pour** une conciliation entre la protection des données et le principe d'innovation, en adoptant **une approche fondée sur l'analyse de risques** et l'analyse *in concreto* et non *a priori*.

Ils considèrent que « *nous ne devons plus nous fonder exclusivement sur la question de la finalité initiale de la collecte des données, mais nous concentrer sur l'usage qui en est fait – ou qui en sera fait – et sur ses conséquences pour les personnes. Nous ne devons plus nous satisfaire du seul formalisme du consentement de la personne, mais mettre en place les instruments qui la protègent concrètement contre les risques de décisions prises sur le fondement de l'analyse de ses données ».*

Ce point de vue est exprimé, en France, par **Gilles Babinet**, l'actuel *Digital Champion* de la France auprès de la Commission européenne[59] : « *La **CNIL** remplit très bien son rôle mais **le rôle qui lui a été dévolu ne correspond sans doute plus à l'esprit du temps.** L'encadrement de la protection de nos données personnelles devrait probablement a minima faire l'objet d'un débat au Parlement et plus probablement au sein de la société civile elle-même »*[60].

François Bourdoncle, pilote du Plan « *big data* », plaide également pour une réglementation orientée sur la **finalité de « l'usage actuel » plutôt que sur la finalité de la « collecte initiale »**[61].

---

[57] Par exemple, une étude réalisée par des chercheurs du MIT, à partir d'une base de données d'horodatage des antennes relais auxquelles le téléphone s'est connecté, a mis en défaut l'anonymat présupposé de ces données. L'étude a ainsi démontré qu'il suffisait de connaître quatre points de localisation d'une personne pour isoler un individu avec une très forte probabilité, de près de 90% : dès lors, connaître la position d'un individu à quatre instants devient suffisant pour le rapprocher de son identifiant dans la base anonymisée et lui réassocier l'ensemble de ses déplacements mesurés. Voir l'article publié dans le journal Nature le 25 mars 2013.

[58] http://www.lemonde.fr/idees/article/2015/06/17/faisons-du-big-data-une-chance-pour-l-europe_4655737_3232.html

[59] C'est-à-dire la personne responsable des enjeux de l'économie numérique et chargée de promouvoir les avantages d'une société numérique en France.

[60] http://www.ladn.eu/actualites/big-questions-big-data,article,26293.html

[61] François Bourdoncle « Ne manquons pas la révolution industrielle du Big data ! », Statistique et société, Vol. 2, n° 4, décembre 2014.

La feuille de route[62] du plan « *big data* » envisageait d'ailleurs d' « *adapter le contexte réglementaire et législatif pour permettre le développement d'une filière Big Data* », pour notamment:

- « *permettre des certifications de processus industriels afin de favoriser l'utilisation du Big data* »,
- « *étudier l'opportunité d'adapter le contexte réglementaire et législatif* ».

Dans le rapport qui jetait les bases du Concours mondial d'innovation 2030, Anne Lauvergeon plaidait, à propos du « *big data* », une des sept ambitions de concours, pour un droit à l'expérimentation. « *L'approche traditionnelle (réglementation et administration de contrôle) est mal adaptée aux constantes du temps des usages qui se développent grâce à ces technologies. Un droit à l'expérimentation doit être reconnu, et encadré par un « observatoire des données ». Il importe en effet de ne pas légiférer sur ce thème de manière générique. L'usage des données est sectoriel et demande une approche au cas par cas. Cette méthode pourrait être progressivement élargie à l'échelle européenne de manière, dans la mesure du possible, à construire une réglementation commune au niveau européen* »[63].

Synthèse/propositions :

Votre rapporteur ne sous-estime nullement les défis présentés par le « *big data* » : il estime que ces défis appellent une approche ouverte des principes fondamentaux de la protection des données personnelles ainsi qu'une recherche d'innovation de la part des acteurs pour les appliquer, en se basant sur la réalité des traitements mis en œuvre et sur les solutions offertes par la loi, solutions qui ont déjà pu être exploitées par le passé et qui sont reprises dans le futur règlement européen.

En tout état de cause, les droits des personnes et les principes clés de la législation ne sauraient être écartés, l'impératif de transparence étant notamment fondamental pour développer la confiance des personnes et contribuer à l'essor de l'économie numérique.

C. Les possibilités offertes par la loi pour un déploiement des traitements « *big data* » en conformité avec les principes fondamentaux de la protection des données personnelles

A titre liminaire, **votre rapporteur souhaite souligner que différents traitements présentés aujourd'hui comme étant du « *big data* » ne sont pas totalement nouveaux pour la CNIL**.

Elle a déjà eu à connaître d'applications en infocentre[64] ou de « *datamining* » reposant sur l'exploitation statistique de base de données internes, visant à avoir une meilleure connaissance de catégories de populations (comme les demandeurs d'emploi ou les assurés

[62] Cf. http://proxy-pubminefi.diffusion.finances.gouv.fr/pub/document/18/17721.pdf#page=9

[63] La Commission pense possible, par une approche sectorielle et par type d'usage, de définir une législation et une réglementation pertinente. Il importera de prendre le temps d'observer le développement des nouveaux usages avant de légiférer. L'exemple de la relation de confiance entre les banques et les usagers prouve qu'il est possible d'avoir une approche gagnant-gagnant dans le domaine de la gestion des données personnelles, mais certains systèmes comme le profilage des utilisateurs pour la publicité devront sans doute être gérés de manière spécifique.
De même, il est indispensable d'imposer une étude d'impact économique avant toute législation sur ce sujet, afin de préserver l'équilibre souhaitable entre innovation, compétitivité et respect de la vie privée.

[64] « Base de données qui agrège des données provenant de sources différentes (plusieurs autres bases de données par exemple, ou encore des données provenant de plusieurs programmes différents) ». http://www.cil.cnrs.fr/CIL/spip.php?article1688

sociaux), à déterminer des profils de personnes vulnérables (par exemple dans le domaine sanitaire et social), ou encore à détecter des comportements à risques ou anormaux (comme en matière de lutte contre la fraude ou dans le cadre du crédit *scoring*)[65]. Le « *big data* » apporte cependant une dimension nouvelle à ces objectifs de connaissance et de profilage, du fait de l'explosion du volume des données et de la mise à disposition d'outils d'analyse toujours plus puissants qui contribuent à la généralisation de cette démarche.

Dans le passé, la CNIL, se prononçant sur ses applications, a su trouver des solutions permettant une application adaptée des principes de protection des données personnelles, résidant principalement dans les solutions d'anonymisation (anonymisation des identités, restriction sur certaines requêtes, interdiction de certains croisements de données), les mesures de sécurité (traçabilité des accès, nombre d'utilisateurs limité), et les catégories de familles d'usages jugées compatibles. Autant de pistes de réflexion qui doivent être étudiées pour répondre aux enjeux que présente l'essor du « *big data* ».

En concertation avec les différents acteurs concernés, la CNIL pourrait poursuivre sa recherche de solutions d'accompagnement aux projets « *big data* », innovantes tout en étant conformes aux principes fondamentaux. Cet objectif de recherche d'innovation pour une application effective et protectrice des principes relatifs à la protection des données personnelles est partagé par les différentes autorités de protection des données, étant entendu que le respect des principes doit rester une garantie incontournable[66]. Comme l'indique l'EDPS dans son rapport de novembre 2015, la question n'est pas de savoir s'il faut appliquer les principes « informatique et libertés » au « *big data* » mais comment les appliquer de manière innovante mais protectrice.

Aussi, s'agissant tout d'abord de la **présomption d'incompatibilité entre le « *big data* » et les principes fondamentaux**, **il convient d'écarter cette affirmation** en analysant les possibilités offertes par la loi et les solutions identifiées précédemment par la doctrine.

S'agissant du principe de finalité, l'observation empirique des exemples de traitements actuellement mis en œuvre permet de relativiser cette contradiction, au moins pour les nombreux projets « *big data* » menés par des acteurs dans leur secteur d'activités. Les projets « *big data* » mis en œuvre, par exemple, par des professionnels de la banque ou de l'assurance sont normalement sous-tendus par la poursuite de **finalités s'inscrivant dans le cadre de leurs activités**. De même, la plupart des organismes qui se lancent dans une démarche « *big data* » ont réfléchi en amont à l'objectif poursuivi par ces projets - la lutte contre la fraude, la connaissance des pratiques de consommation d'un segment de population, la détermination de tendances et de corrélations en matière sanitaire, etc. - et aux catégories de bases de données disponibles. Dès lors, l'approche n'est peut être pas si éloignée de celle prévue par la loi et de la conception de la CNIL du principe de finalité. En effet, il convient de rappeler que **la Commission n'a pas une approche rigide et stricte du principe de finalité**, mais plutôt souple, tout en veillant effectivement à ce que les données soient collectées pour des finalités déterminées, explicites et légitimes[67]. Cela est notamment illustré par les finalités identifiées dans les normes simplifiées. Par exemple, les normes simplifiées n° 48 et n° 56 couvrent tous les champs liés à la gestion des clients et des

[65] Cf. par exemple les avis rendus sur la segmentation comportementale dans le domaine bancaire (1993), sur des systèmes d'aide à la décision dans le domaine social (ex SIAM et SNIIRAM pour l'assurance maladie), aide à la sélection et au contrôle fiscal des particuliers (SIRIUS), outils statistiques d'aide à la connaissance des demandeurs d'emploi SIAD, etc.

[66] Cf. par exemple « Working paper on big data and privacy », Groupe de Berlin, mai 2014; « Big data and data protection », ICO, juillet 2014; « Statement of the WP29 on the impact of the development of big data on the protection of the individuals with regard to the processing of their personal data in the EU", Groupe de l'Article 29, septembre 2014; « Meeting the challenges of big data », EDPS, novembre 2015.

[67] Il convient de rappeler que le détournement de finalité est sanctionné par l'article 226-21 du Code pénal d'une peine de cinq ans d'emprisonnement et de 300 000 euros d'amende.

prospects, et se déclinent en de nombreuses sous finalités[68], à l'exception de celles soumises à autorisation préalable. Il est également intéressant de noter que l'article 25.I.5° de la loi, relatif aux interconnexions, fait référence à la notion de finalité « principale » d'un fichier et que la notion de famille de finalités a pu être utilisée dans le domaine de l'assurance maladie. Concernant le principe de famille de finalités, le considérant 25 du futur règlement européen, sur la notion de consentement, précise que « *le consentement donné devrait valoir pour toutes les activités de traitement ayant la même finalité. Lorsque le traitement a plusieurs finalités, un consentement devrait être donné pour l'ensemble des finalités du traitement* ».

S'agissant de la **possibilité de réutiliser les données pour des finalités autres** que celles pour lesquelles elles ont été initialement collectées, **l'article 6.2° de la loi**[69] prévoit de telles possibilités, sous certaines conditions, pour les finalités jugées compatibles et pour celles pour lesquelles il existe une présomption de compatibilité, à savoir un traitement ultérieur des données à des fins statistiques ou à des fins de recherche scientifique ou historique (sous réserve que ce traitement ne soit pas utilisé pour prendre de décisions à l'égard des personnes concernées). **Cet article, qui trouve son pendant dans le futur règlement européen[70], offre des possibilités très intéressantes pour le « *big data* »**, qui seront détaillées dans la dernière partie du présent rapport.

S'agissant de la problématique liée à la durée de conservation des données, il convient de souligner que la question ne se pose pas automatiquement pour tous les traitements « *big data* ». En effet, les données utilisées peuvent être collectées spécifiquement pour la mise en œuvre du projet « *big data* », peuvent être des données disponibles en open data ou anonymisées ou encore être des données dont dispose le responsable de traitement dans le cadre de sa relation contractuelle qui le lie aux personnes concernées[71].

Lorsque cette problématique se pose, l'article 36 de la loi du 6 janvier 1978 modifiée prévoit, sous certaines conditions, des possibilités de conservation des données au-delà de la durée nécessaire à la réalisation de la finalité initiale[72].

---

[68] Par exemple, pour la NS 48, la finalité générale de gestion des clients et des prospects se décline en sous-finalités : effectuer des opérations relatives à la gestion des clients (contrats, livraisons, factures, etc.), effectuer des opérations relatives à la prospection commerciale (sélection des personnes, opérations techniques, etc.), élaborer des statistiques commerciales, gérer les impayés, céder, louer ou échanger des fichiers, organiser des jeux concours, etc.

[69] Selon l'article 6.2 de la loi, les données « *sont collectées pour des finalités déterminées, explicites et légitimes et ne sont pas traitées ultérieurement de manière incompatible avec ces finalités. Toutefois, un traitement ultérieur de données à des fins statistiques ou à des fins de recherche scientifique ou historique est considéré comme compatible avec les finalités initiales de la collecte des données, s'il est réalisé dans le respect des principes et des procédures prévus au présent chapitre, au chapitre IV et à la section 1 du chapitre V ainsi qu'aux chapitres IX et X et s'il n'est pas utilisé pour prendre des décisions à l'égard des personnes concernées* ». Il est intéressant de noter que, plutôt que d'imposer une véritable exigence de compatibilité, le législateur a choisi une double négation en interdisant l'incompatibilité, ce qui offre plus de souplesse quant aux finalités poursuivies. Le fait que le traitement de données s'effectue pour une finalité différente de celle d'origine ne signifie pas nécessairement que les finalités sont incompatibles.

[70] Selon l'article 5.1(b) « *Personal data must be : (…) (b) collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes; further processing of personal data for archiving purposes in the public interest, or scientific and historical research purposes or statistical purposes shall, in accordance with Article 83(1), not be considered incompatible with the initial purposes; ("purpose limitation)"*. Voir également le considérant 40.

[71] Les durées de conservation des données peuvent être importantes puisque les données peuvent être conservées le temps de la relation contractuelle/commerciale. Voir, par exemple, la NS 48.

[72] Selon l'article 36 de la loi, « *Les données à caractère personnel ne peuvent être conservées au-delà de la durée prévue au 5° de l'article 6 qu'en vue d'être traitées à des fins historiques, statistiques ou scientifiques ; le choix des données ainsi conservées est opéré dans les conditions prévues à l'article L. 212-4 du code du patrimoine. (…)*
Il peut être procédé à un traitement ayant des finalités autres que celles mentionnées au premier alinéa :
- soit avec l'accord exprès de la personne concernée ;
- soit avec l'autorisation de la Commission nationale de l'informatique et des libertés ;
- soit dans les conditions prévues au 8° du II et au IV de l'article 8 s'agissant de données mentionnées au I de ce même article ».

De manière générale, des possibilités de réutilisation des données par un responsable de traitement sont également offertes par le biais de **l'anonymisation des données**. Des solutions ont ainsi été trouvées par le passé, comme pour le traitement Flux vision de la société Orange, qui a été évoqué précédemment[73].

Votre rapporteur considère toutefois que les possibilités offertes par l'anonymisation des données ne doivent pas être surestimées compte tenu de l'exigence requise pour qualifier des données d'anonymes, nécessitant le recours à différentes techniques qui conduisent à un appauvrissement des données exploitées, et en raison des risques futurs de ré-identification des personnes, qui sont difficiles à anticiper. Il note toutefois que des pistes de réflexion pourraient être étudiées, dans le sillage du « caveat » défini par la CNIL, la CADA, la DILA et Etalab à l'occasion de l'ouverture des bases de données de jurisprudence. Ce « caveat » précise le cadre juridique applicable à la réutilisation des jeux de données et rappelle notamment aux réutilisateurs que, dès lors qu'un jeu de données a fait l'objet d'une anonymisation totale ou partielle, la réutilisation, notamment dans le cadre de croisements de données, ne peut avoir ni pour effet ni pour objet de ré-identifier les personnes[74].

S'agissant du **croisement de données** contenues dans des bases issues de sources très diverses, il convient de rappeler que différents cas de figure peuvent être distingués. Si la ou les sociétés souhaitent interconnecter différents jeux de données, il est a priori possible de le faire librement lorsque les jeux de données sont anonymisés, sous réserve de s'assurer que cet appariement ne conduit pas à une ré-identification des personnes. Si les jeux de données ont vocation à être interconnectés préalablement à tout processus d'anonymisation, et que les traitements concernés ont des finalités principales différentes, le ou les traitements devront être autorisés par la CNIL conformément à l'article 25-I-5° de la loi[75].

L'application de cette garantie offerte par la loi aux projets « *big data* » concernés permettrait à la CNIL d'être saisie des projets qui peuvent sembler problématiques dans la mesure où ils recourent à des bases de données ayant des sources et des finalités très différentes.

S'agissant du **principe de pertinence des données (article 6.3° de la loi)**, **la Commission apprécie cette pertinence au cas par cas** et cette appréciation ne se fait pas nécessairement de manière stricte. Par exemple, en matière de *scoring*, l'évaluation prend en compte la combinaison des critères pour évaluer la pertinence du traitement, et contrôle les garanties associées. Par exemple, dans le cadre d'un traitement de *scoring* mis en place par la société Bouygues Telecom en matière de prévention contre la fraude et les impayés,

---

La CNIL a eu recours aux dispositions de l'article 36.3° pour autoriser des traitements à des fins généalogiques (délibérations n°2013-105 du 25 avril 2013, n°2012-416 du 29 novembre 2012 et n°2011-383 du 24 novembre 2011) et pour autoriser les ministères et conseils départementaux responsables de services d'archives publics à mettre en œuvre des traitements de données personnelles ayant pour finalité de numériser, indexer et diffuser sur internet les « registres matricules » des soldats ayant participé à la Première Guerre mondiale (délibération n°2013-281 du 10 octobre 2013).

[73] Ce service permet de convertir en temps réel des millions d'informations techniques provenant du réseau mobile en indicateurs statistiques, pour analyser la fréquentation de zones géographiques et les déplacements de population, qui peuvent être utilisés dans le domaine du tourisme, de l'aménagement du territoire, du trafic routier ou du commerce. L'offre repose sur un procédé d'anonymisation développé par Orange en concertation avec la CNIL, qui supprime toute possibilité d'identifier les clients de l'opérateur, grâce notamment au taux de collision choisi entre les identifiants hachés.

[74] http://rip.journal-officiel.gouv.fr/index.php/content/download/620/3209/file/CAVEAT_RIP_2015_09_10.pdf

[75] Ce régime juridique était notamment rappelé en mai dernier à la société Data Publica, qui pilote le projet XData (courrier SA151054 du 7 mai 2015). L'article 25.I.5° de la loi prévoit que sont mis en œuvre après autorisation de la CNIL « *Les traitements automatisés ayant pour objet :*
- l'interconnexion de fichiers relevant d'une ou de plusieurs personnes morales gérant un service public et dont les finalités correspondent à des intérêts publics différents ;
- l'interconnexion de fichiers relevant d'autres personnes et dont les finalités principales sont différentes ».
A contrario, en présence de finalités principales identiques, le traitement d'interconnexion relève d'une simple déclaration normale. Ce cas de figure ne conduisant finalement qu'à un enrichissement.

autorisé en 2012, le rapport précisait que compte tenu de la complexité de l'algorithme, il est difficile de déterminer que l'ensemble des critères retenus est totalement pertinent. Toutefois, dans la délibération d'autorisation, la Commission a considéré qu'à partir du moment où ce traitement ne reste qu'une aide à la décision, et qu'aucun critère ne peut à lui seul être un facteur d'exclusion, les données sont pertinentes, adéquates et non excessives. La possibilité d'adopter au cas par cas une approche similaire en matière de « *big data* », lorsque la finalité du traitement le permet, pourrait être étudiée, tout en veillant à ne pas dénaturer ou vider de son sens le principe de pertinence, mais en se montrant accueillant pour permettre le traitement de données dont la pertinence n'est pas nécessairement évidente.

S'agissant du principe posé à l'article 6.4° de la loi, selon lequel les données doivent être exactes, complètes et, si nécessaire, mises à jour, votre rapporteur considère qu'il conserve toute sa pertinence dans un contexte « *big data* », et qu'il permet de satisfaire l'un des V qui caractérise cette approche, le V de Véracité.

Votre rapporteur considère ainsi que les vrais enjeux du « *big data* » en matière de protection des données personnelles ne sont pas nécessairement la remise en question de principes tels que celui de finalité, ou la réutilisation des données pour d'autres finalités, lorsqu'elles sont jugées compatibles.

L'enjeu semble davantage résider dans l'utilisation qui peut être faite du « *big data* » pour cibler les personnes et prendre des décisions à leur égard. Comme nous l'avons vu, de nombreux traitements « *big data* » ne sont pas utilisés à cette fin. Toutefois, lorsque tel est le cas, le recours au « *big data* » et aux algorithmes prédictifs comporte un véritable risque de prise de décision automatisée que les individus ne comprennent pas et sur lesquelles ils n'ont aucun contrôle.

En l'absence de transparence pour les personnes concernées, **ces traitements « *big data* » pourraient conduire à un processus décisionnel totalement opaque**, **comportant à la fois des risques d'imprécision, de mauvaise inférence, de discrimination voire d'exclusion pour les intéressés**. Comme le souligne l'autorité de protection des données norvégienne dans son rapport sur le « *big data* », il existe un risque de « dictature des données », les personnes n'étant plus jugées sur leurs actions mais sur la base de toutes les données les concernant qui indiquent quelles pourraient être leurs actions probables. Les bénéfices attendus de ces traitements basés sur les statistiques prédictifs pourraient générer une confiance excessive en leurs capacités. Les applications « *big data* » peuvent en effet identifier des corrélations fallacieuses, dans des cas où il n'y a pas de relation directe de cause à effet entre deux phénomènes présentant pourtant une corrélation forte entre eux. Il existe alors un risque de déduire de ces traitements des conclusions imprécises voire, lorsqu'ils sont appliqués à un niveau individuel, des conclusions potentiellement injustes ou discriminatoires. Cette approche pourrait également renforcer et conforter l'existence de stéréotypes.

Dans son étude annuelle 2014, « Le numérique et les droits fondamentaux », le Conseil d'Etat souligne également que lorsque les usages du « *big data* » visent les personnes en tant que telles, les algorithmes présentent trois sources de risques pour l'exercice des libertés : « *l'enfermement de l'internaute dans une « personnalisation » dont il n'est pas maître ; la confiance abusive dans les résultats d'algorithmes perçus comme objectifs et infaillibles ; de nouveaux problèmes d'équité du fait de l'exploitation toujours plus fine des données personnelles* ». Le Conseil d'Etat en déduit que « *lorsque les usages du Big Data visent les personnes en tant que telles, par exemple pour établir un profil prédictif de leurs caractéristiques (solvabilité, dangerosité…), la pleine application des principes fondamentaux de la protection des données personnelles est (…) requise. Si les principes*

conservent leur pertinence, les instruments de la protection des données doivent en revanche être adaptés et renouvelés ».

Ces inquiétudes et risques liés au « *big data* » sont loin de concerner l'ensemble des traitements actuellement mis en œuvre sous cette appellation. **Cela milite donc pour une approche différenciée des traitements, en fonction notamment de leur objectif et des enjeux qu'il engendre**.

Votre rapporteur vous propose une typologie permettant de différencier les traitements « *big data* » en fonction de leurs enjeux, pour déterminer le régime juridique qui leur est applicable en fonction de ceux-ci. Cette approche vise à permettre à la fois une ouverture et une diversification possible des usages des données, allant de pair avec un renforcement des droits et des moyens de contrôle des personnes pour les traitements ayant un impact sur elles.

Synthèse/propositions :

- La CNIL doit se donner comme priorité de rechercher, avec les acteurs concernés, des solutions d'accompagnement aux projets « *big data* » qui peuvent être innovantes mais protectrices et conformes aux principes fondamentaux de protection des données personnelles.

- La présomption d'incompatibilité entre le « *big data* » et les principes fondamentaux doit être écartée sur la base des possibilités offertes par la loi et par la doctrine de la CNIL (conception du principe de finalité, possibilité de réutilisation des données pour des finalités compatibles, possibilités offertes par l'anonymisation des données, etc.).

- Le véritable enjeu est l'utilisation du « *big data* » qui peut être faite pour cibler les personnes et prendre des décisions à leur égard. Il existe un risque de processus décisionnel automatisé et opaque pour les personnes concernées, comportant à la fois des risques d'imprécision, de mauvaise inférence, de discrimination voire d'exclusion pour les intéressés.

- Ce constat milite pour une approche différenciée des traitements « *big data* » en fonction de leurs caractéristiques et de leurs enjeux. Cette approche vise à permettre à la fois une ouverture et une diversification possible des usages des données, allant de pair avec un renforcement des droits et des moyens de contrôle des personnes pour les traitements ayant un impact sur elles.

III. PISTES DE REFLEXION ET POSITIONNEMENT POUR UNE APPROCHE DIFFERENCIEE DES TRAITEMENTS « *BIG DATA* »

Au vu des développements précédents, votre rapporteur propose une approche différenciée des traitements « *big data* » en fonction de leurs caractéristiques et de leurs enjeux. Cette approche permettrait de déterminer le régime juridique applicable aux différents traitements. Elle permettrait également d'explorer les possibilités d'ouverture et les points sur lesquels il serait opportun de trouver des solutions innovantes, tout en mettant l'accent sur les traitements pour lesquels des garanties fortes devraient être apportées, notamment en termes de transparence envers les personnes et de moyens de contrôle.

Les propositions présentées ci-dessous sont des pistes d'analyse et de réflexion, ainsi que des ébauches de solutions. Elles n'ont pas vocation à être exhaustives ou gravées dans le

marbre et votre rapporteur souhaite les soumettre à la Commission pour savoir si elle partage ces analyses.

A. Identification des critères de différenciation des traitements permettant d'élaborer une typologie

À partir d'exemples de traitements qui sont actuellement présentés comme étant des traitements « *big data* », il est possible de dégager différents critères permettant de distinguer et de catégoriser ces traitements.

Ces exemples permettent de constater que certains traitements sont qualifiés par les acteurs comme étant du « *big data* » alors qu'ils ne présentent pas le degré de complexité qui est souvent associé *a priori* au « *big data* », et ne soulèvent pas tous les enjeux qui sont généralement associés à ces traitements. Il s'agit d'ailleurs souvent de traitements sur lesquels la CNIL s'est déjà prononcée et qu'elle est donc en mesure d'appréhender de manière effective. Cette analyse permet de conclure que tous les traitements « *big data* » ne présentent pas les mêmes problématiques en termes de régulation et de « compatibilité » avec la loi du 6 janvier 1978 modifiée.

La typologie proposée, qui résulte de deux critères de différenciation, l'origine des données et l'objectif du traitement, devrait permettre de qualifier les traitements et de les étudier en conséquence, en déterminant notamment une présomption de base légale du traitement. Chaque critère est détaillé de manière distincte dans le présent rapport, mais ils devraient être croisés pour étudier les traitements, comme l'illustre le tableau récapitulatif (cf. annexe 2).

Cette différenciation des traitements selon ces facteurs permet aussi de mettre en exergue les deux enjeux pour la CNIL en termes de régulation et les deux points de vigilance pour les responsables de traitement :

- quelles sont les possibilités d'ouverture pour la réutilisation des données pour d'autres finalités ?
- quelles sont les conditions pour un traitement loyal et licite des données ?

Ces deux points feront l'objet de développements détaillés dans un second temps.

<u>Premier critère de différenciation : d'où proviennent les données ?</u>

Un **premier facteur de différenciation** est lié à **l'origine des données qui sont exploitées** pour la mise en œuvre du traitement. Cette origine peut soulever des problématiques en termes d'information et d'exercice des droits des personnes et peut engendrer des différences quant au régime juridique applicable au traitement.

Cas n°1 : le responsable de traitement met en œuvre son traitement à partir des données dont il dispose, sans avoir recours à des sources externes de données. Ces données peuvent collectées pour la mise en œuvre du traitement « *big data* » (cas n°1.1) ou être des données dont le responsable de traitement dispose déjà et qui font l'objet d'un traitement ultérieur « *big data* » (cas n°1.2).

Plusieurs exemples de traitements peuvent être cités, qui soulèvent diverses problématiques appelant différentes solutions/pistes de réflexion.

Tout d'abord, comme nous l'avons vu, différents traitements qui sont aujourd'hui présentés comme étant du « *big data* » ne sont pas totalement nouveaux pour la CNIL, qui a déjà eu à connaître **d'applications en infocentre ou de « *datamining* »** reposant sur l'exploitation

statistique de base de données internes, visant à avoir une meilleure connaissance de catégories de populations, à déterminer des profils de personnes vulnérables, ou encore à détecter des comportements à risques ou anormaux[76].

Ces types de traitement ne présentent pas « d'incompatibilité » avec la loi « informatique et libertés », à partir du moment où ils sont mis en œuvre en respectant les droits et libertés des personnes concernées et qu'ils ne sont pas contraire à **l'article 10 de la loi**[77]. À cet égard, lorsque ces traitements peuvent engendrer des conséquences négatives pour les personnes concernées, leur encadrement est effectué par le biais de la demande d'autorisation, conformément aux dispositions de **l'article 25-I-4° de la loi**[78]. Lorsqu'il s'agit d'interconnexion entre des fichiers ayant des finalités différentes, une demande d'autorisation doit être présentée en application de **l'article 25-I-5° de la loi**.

Comme cela a été évoqué précédemment, le responsable de traitement peut aussi décider de recourir à de « nouvelles » sources de données pour mettre en œuvre un traitement « *big data* », comme cela est proposé par des **compagnies d'assurance** pour de nouvelles offres ("***Pay as you drive*"/"*pay how you drive***") reposant sur l'analyse des données collectées dans le véhicule par le biais de capteurs ou d'applications mobiles, créés par l'assureur ou non. L'objectif est de permettre au conducteur de payer son assurance moins chère en fonction des distances parcourues ou du comportement du conducteur au volant.

Compte tenu des enjeux que représentent ces dispositifs pour la protection des données personnelles et de la vie privée, leur développement est encadré par des garanties : les personnes doivent disposer d'**informations claires** sur ces dispositifs, donner leur **consentement** et avoir la possibilité de retirer ce consentement à tout moment, en désactivant le dispositif. Les obligations des assurances au regard de la loi sont donc notamment de recueillir le consentement des personnes avant toute collecte de données de géolocalisation et de ne pas utiliser ces données pour la mise en place de fichiers d'infractions au code de la route.

Enfin, le responsable de traitement peut décider de réutiliser des données qui sont déjà en sa possession, pour d'autres finalités. C'est le cas notamment de la société Orange, pour le projet **Flux vision**. La solution réside ici notamment dans **l'anonymisation des données**[79], anonymisation qui peut permettre d'envisager des conditions d'information différentes des personnes et qui permet d'écarter la problématique de l'exercice des droits des personnes une fois les données anonymisées, celles-ci n'étant plus des données à caractère personnel.

Cas n° 2 : le responsable met en œuvre son traitement à partir de données issues en partie ou totalement de sources externes publiques

---

[76] Cf. par exemple les avis rendus sur la segmentation comportementale dans le domaine bancaire (1993), sur des systèmes d'aide à la décision dans le domaine social (ex SIAM et SNIIRAM pour l'assurance maladie), aide à la sélection et au contrôle fiscal des particuliers (SIRIUS), outil statistique d'aide à la connaissance des demandeurs d'emploi SIAD, etc.

[77] Un des principes fondamentaux de la loi informatique et libertés (article 10 de la loi) interdit toute décision, produisant des effets juridiques à l'égard d'une personne, prise sur le seul fondement d'un traitement automatisé destiné à définir le profil de l'intéressé ou à évaluer certains aspects de sa personnalité. Cet article trouve son pendant dans le futur règlement européen (article 20).

[78] « *Les traitements automatisés susceptibles, du fait de leur nature, de leur portée ou de leurs finalités, d'exclure des personnes du bénéfice d'un droit, d'une prestation ou d'un contrat en l'absence de toute disposition législative ou règlementaire* ». Ces traitements devraient au regard des dispositions du futur règlement relever des articles 33 et éventuellement 34 : réalisation d'une étude d'impact sur la vie privée et consultation préalable de l'autorité de protection des données.

[79] Ce service permet de convertir en temps réel des millions d'informations techniques provenant du réseau mobile en indicateurs statistiques, pour analyser la fréquentation de zones géographiques et les déplacements de population, qui peuvent être utilisés dans le domaine du tourisme, de l'aménagement du territoire, du trafic routier ou du commerce. Le processus d'anonymisation mis en œuvre a été développé en concertation avec la CNIL.

Un autre cas de figure concerne les traitements utilisant des données publiquement accessibles. Ici encore, différents types de traitements peuvent être cités, ne soulevant pas les mêmes problématiques.

En premier lieu, les données utilisées peuvent être des données anonymisées, comme les données préalablement anonymisées issues de l'open data. Auquel cas, aucune problématique particulière ne se pose quant à l'utilisation de ces données issues de sources externes publiques puisque ces données ne sont plus des données à caractère personnel.

En second lieu, les données utilisées peuvent être des données à caractère personnel qui sont publiquement accessibles sur l'Internet. Le responsable de traitement peut souhaiter exploiter ces données pour des finalités très différentes de celles pour lesquelles elles ont été diffusées ou initialement traitées (par exemple, pour une utilisation à des fins de recrutement, de prévention de la fraude ou encore de lutte contre le terrorisme).

Lorsque le responsable de traitement envisage une telle utilisation, il doit être particulièrement vigilant quant aux conditions à satisfaire pour un traitement loyal et licite des données. En effet, jusqu'à présent, la position de la CNIL sur cette question est celle de la décision Pages jaunes[80], selon laquelle la collecte massive, répétitive et indifférenciée de ces données sans en avertir les personnes concernées ne permet pas de considérer que les données ont été collectées de manière loyale et licite.

Toutefois, la CNIL a récemment accepté une exception à cette position, dans le cadre de la note d'arbitrage sur l'usage des réseaux sociaux par les assureurs pour la recherche des bénéficiaires de contrats d'assurance-vie non réclamés[81]. Dans le sillage de cette évolution, votre rapporteur considère qu'il convient de s'interroger dans le contexte du « *big data* » sur le maintien d'une position stricte pour l'ensemble de traitements alors que ceux-ci peuvent avoir des finalités très différentes. Sur cette question, il s'agit à nouveau d'identifier quels sont les enjeux et les risques pour les personnes concernées, quels sont les cas pour lesquels une ouverture et la recherche de nouvelles solutions semblent pertinentes et ceux pour lesquels, au contraire, les moyens de contrôle des personnes doivent être préservés, voire renforcés. A cet égard, des pistes de réflexion et des propositions seront présentées ultérieurement (Point C « Identification des conditions pour un traitement loyal et licite des données »).

Cas n°3 : plusieurs responsables de traitements souhaitent mettre en commun différentes bases de données.

Ce cas révèle la complexité des traitements « *big data* », lorsque différents responsables de traitements veulent mettre en commun leurs données pour mener de nouveaux projets. On peut citer à cet égard le cas du **projet XData** piloté par la société Data Publica. Ce projet, initié en 2013, réunit les sociétés EDF, La Poste, Orange et Veolia, les start-up Cinequant, Data Publica et Hurence et les instituts de recherche Inria et Télécom ParisTech. Ce projet vise à réunir différents jeux de données issus des sociétés membres, de l'open data et des

[80] Délibération de la formation restreinte de la CNIL n°2011-203 du 21 septembre 2011, décision du Conseil d'Etat du 12 mars 2014. La société Pages Jaunes proposait une fonctionnalité permettant d'ajouter aux résultats de recherche obtenus sur une personne déterminée des données personnelles collectées sur les réseaux sociaux. Il a été considéré que « *la circonstance que des profils personnels sont affichés publiquement sur Internet ne permet pas pour autant à un organisme tiers de procéder à une collecte massive, répétitive et indifférenciée de ces données sans en avertir les personnes concernées* ». En effet, l'information des personnes ne pouvait être considérée comme suffisante du fait de politiques de confidentialité de réseaux sociaux - mentionnant une possible indexation de leurs données par des moteurs de recherche - et la société n'était pas fondée à soutenir que l'information de ces personnes, dont elle avait les coordonnées, exigeait des efforts disproportionnés par rapport à l'intérêt de la démarche. Dès lors, il a été considéré que les données des personnes concernées n'avaient pas été collectées de manière loyale et licite.

[81] En lien avec les obligations légales résultant de la loi n° 2014-617 du 13 juin 2014 relative aux comptes bancaires inactifs et aux contrats d'assurance vie en déshérence. Cette recherche n'est toutefois pas automatisée et ne conduit pas à l'extraction des données publiées sur le web ou dans les réseaux sociaux en vue d'une réutilisation.

réseaux sociaux sur une plate-forme matérielle et logicielle, pour une mise à disposition des données sur le serveur pour un ou plusieurs partenaires du projet, les quatre finalités identifiées par le projet étant « la recherche fondamentale », « la recherche sur les méthodes d'anonymisation de jeux de données », « la recherche appliquée » et « l'exploration commerciale »[82].

Ce type de projets peut appeler différentes solutions/pistes de réflexion : si les parties souhaitent interconnecter différents jeux de données, elles peuvent le faire librement lorsque les jeux de données sont anonymisés, en s'assurant toutefois que cet appariement ne conduit pas à une ré-identification des personnes (et sous réserve que les parties aient effectué au préalable les formalités requises liées au traitement d'anonymisation). **La solution/piste de réflexion réside ici dans l'anonymisation des données et les responsables de traitements peuvent être orientés sur ce point vers l'avis du G29 du 10 avril 2014.**

Si les jeux de données ont vocation à être interconnectés avant tout processus d'anonymisation, ces traitements sont soumis à autorisation préalable de la CNIL conformément à **l'article 25-I-5° de la loi**. **La solution réside dans le contrôle de la CNIL a priori, qui aboutit à une autorisation ou un refus de mise en œuvre du traitement en fonction de sa conformité avec la loi « informatique et libertés ».**

Lorsque les responsables de traitements souhaitent mettre en commun des données issues de traitements ayant la même finalité, le traitement n'est pas soumis à autorisation préalable, l'appariement conduisant à un enrichissement[83]. Il faut alors veiller notamment au fait que les responsables de traitements soient bien identifiés comme destinataires des données, que les personnes aient été dûment informées et qu'elles soient en mesure d'exercer leurs droits. Selon la finalité, le consentement des intéressés peut être requis, comme nous allons à présent le voir en analysant la deuxième approche.

Deuxième critère de différenciation : que veut-on faire des données ?

Un second facteur de différenciation des traitements « *big data* » pourrait être l'objectif et la **finalité poursuivie par le traitement**. La **première catégorie** concerne les traitements mis en œuvre à des fins de **détection de tendances ou de corrélations** entre des données (c'est-à-dire à des fins de connaissance d'un phénomène, de recherche ou d'établissement de statistiques). La **seconde catégorie** concerne les traitements grâce auxquels le responsable de traitement s'intéresse plus directement à la personne et souhaite pouvoir la cibler (c'est-à-dire des **traitements à des fins individuelles, impliquant un ciblage, un retour individualisé vers la personne ou une prise de décision la concernant**). Cette distinction semble faire consensus. Elle est développée par le G29, dans son avis WP 206 « *Purpose limitation* », qui traite en partie du « *big data* », ainsi que par le Groupe de Berlin.

En l'absence d'autres fondements légaux visés à l'article 7 de la loi - notamment le respect d'une disposition légale incombant au responsable de traitement et prévoyant la mise en œuvre du traitement (par exemple pour les traitements régaliens), ou de l'existence de garanties particulières pour la mise en œuvre du traitement, en particulier le conditionnement de la mise en œuvre du traitement à l'autorisation préalable de la CNIL (par exemple pour

[82] Pour de plus amples informations, se référer au courrier adressé à la société DATA PUBLICA le 7 mai 2015 (Saisine n°15004921, courrier SA151054).

[83] C'est le cas, par exemple, des acteurs spécialisés dans la collecte et la revente de données, les « *data brokers* ». L'étude annuelle 2014 du Conseil d'Etat, « Le numérique et les droits fondamentaux », précise que le plus important d'entre eux affirme détenir des données sur 700 millions de personnes dans l'ensemble du monde. http://www.ladocumentationfrancaise.fr/var/storage/rapports-publics/144000541.pdf

les traitements d'exclusion)[84] ou à l'avis de la CNIL - il semble possible de déterminer que les **traitements visant à cibler plus précisément les personnes**, et à prendre le cas échéant des décisions à leur encontre, doivent être mis en œuvre avec le **consentement** des intéressés et dans le respect des conditions fixées à l'article 10 de la loi. La précision de ce ciblage ou l'effet attendu de la mise en œuvre du traitement pourrait justifier que l'on écarte l'intérêt légitime du responsable de traitement ou que l'on considère que celui-ci cède le pas face à l'impact sur les droits et libertés fondamentaux des personnes.

En revanche, pour les **traitements ayant des finalités de détection de tendances et de corrélations** entre les données, ils pourraient *a priori* bénéficier du fondement légal lié à la poursuite de **l'intérêt légitime du responsable de traitement**. Le consentement serait toutefois requis pour certains traitements, comme ceux pour lesquels des données sensibles sont collectées et traitées.

S'agissant de traitements actuellement mis en œuvre, on peut citer les **exemples** suivants :

Cas n° 1 : les traitements ayant pour objet la détection de tendances et de corrélations entre les données. Un projet mis en œuvre par l'École polytechnique de Milan et l'Université de Stanford a utilisé les données fournies par un opérateur téléphonique pour développer des modèles qui définissent le schéma de propagation de la schistosomiase au Sénégal[85].

Cas n° 2 : traitements visant à cibler plus précisément les personnes, et à prendre le cas échéant des décisions à leur encontre. De nombreux exemples de ces usages ont été cités dans le présent rapport, par exemple dans les domaines du marketing, de l'assurance ou de la détection des comportements à risque.

Votre rapporteur souhaite souligner que la classification *a priori* de certains traitements dans le cas n°1 ou le cas n° 2 n'est pas forcément aisée, comme le traitement permettant de définir les recommandations proposées aux utilisateurs sur le site internet Netflix. Le postulat peut être qu'en fonction des modalités de traitement des données, le traitement basculera nécessairement dans une case ou dans l'autre. Mais cette difficulté de classement du traitement *a priori* peut aussi révéler la nécessité d'envisager un troisième cas, à mi-chemin entre les deux autres. Cette grille d'analyse devra ainsi être confortée par la pratique, afin de déterminer s'il est nécessaire de l'affiner et de la complexifier, par exemple si l'existence d'une troisième catégorie pouvait avoir pour conséquence de modifier le régime juridique applicable au traitement. La Commission est actuellement saisie de projets « *big data* » qui pourront permettre une mise en application concrète de cette grille d'analyse.

[84] De nombreux traitements qui sont actuellement soumis à demande d'autorisation préalable seront, avec le futur règlement européen, subordonnés à la réalisation d'une étude d'impact sur la vie privée et, le cas échéant, consultation préalable de l'autorité de protection des données (articles 33 et 34). L'étude d'impact est notamment requise pour les cas suivants : « *a) a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the individual or similarly significantly affect the individual; b) processing on a large scale of special categories of data referred to in Article 9(1), or of data relating to criminal convictions and offences referred to in Article 9a; c) a systematic monitoring of publicly accessible area on a large scale"*

[85]http://www.orange.com/fr/Presse-et-medias/communiques-2015/communiques-2015/Orange-annonce-les-gagnants-de-son-challenge-Data-for-Development-au-Senegal-sur-fond-de-developpement-et-bien-etre-de-la-population.
"La schistosomiase est une infection parasitaire d'origine aquatique qui provoque des symptômes débilitants et qui affecte des millions de personnes. Nous démontrons qu'un modèle relativement simple permet de reproduire de façon fiable les caractéristiques régionales de la prévalence de la schistosomiase à travers le pays. Nous utilisons ce modèle pour étudier le rôle que tient la mobilité humaine dans les mécanismes de la maladie et pour analyser les stratégies d'intervention qui ont pour but de réduire son impact".

<u>Cas n° 1 Origine des données</u> : données dont dispose le RT	**<u>Cas n° 1 Finalité / objectif du traitement</u> : le traitement vise à détecter des tendances ou des corrélations entre des données (à des fins de connaissance d'un phénomène, de recherche ou d'établissement de statistiques)**	**<u>Cas n° 2 Finalité / objectif du traitement</u> : le traitement vise à cibler la personne (traitements à des fins individuelles, impliquant un retour individualisé vers la personne ou une prise de décision la concernant)**
<u>Cas n° 1.1</u> : collecte directe	**<u>Possibilité 1)</u> Elaboration de statistiques,** présomption de base légale : l'**intérêt légitime** du RT.	**<u>Possibilité 1)</u>** Traitement visant à une **connaissance fine des personnes** pour adaptation offre ou prise de décision. Présomption de nécessité du **consentement** comme base légale. **<u>Possibilité 2)</u>** Traitement pour détecter des comportements à risque. Besoin d'une **autorisation**, contrôle *a priori* de la CNIL (intérêt légitime peut être la base légale). **<u>Possibilité 3)</u> obligation légale**. **Traitements "régaliens". Avis de la CNIL.**
<u>Cas n° 1.2</u> : Réutilisation des données pour d'autres finalités	**<u>Possibilité 1)</u> Anonymisation** des données pour les exploiter. Présomption de base légale : l'**intérêt légitime** du RT. **<u>Possibilité 2)</u> Finalité ultérieure compatible**. Présomption de base légale : l'**intérêt légitime** du RT.	**<u>Possibilité 1)</u>** Réutilisation des données à des **fins de ciblage commerciale**. **Le test de compatibilité ne fonctionne a priori pas**. Présomption de nécessité du **consentement** comme base légale. **<u>Possibilité 2)</u>** Réutilisation des données pour **détecter les comportements à risque**. Besoin d'une **autorisation**, contrôle *a priori* de la CNIL (intérêt légitime peut être la base légale). **<u>Possibilité 3)</u> obligation légale. Traitements "régaliens". Avis de la CNIL.**
<u>Cas n° 2 Origine des données</u> : le RT met en	**<u>Possibilité 1)</u>** Données en **open data** ou anonymisées. **<u>Possibilité 2)</u>** Utilisation de **données personnelles issues**	**<u>Possibilité 1)</u>** Utilisation de **données personnelles issues de l'Internet pour ciblage**. Nécessité du **consentement** des personnes.

œuvre son traitement à partir de données issues en partie ou totalement de sources externes publiques (open data, Internet, etc.).	**de l'Internet**. **<u>Problématique particulière</u>** : Base légale peut être l'**intérêt légitime**, mais données personnelles dont les modalités de réutilisation pour des finalités jugées compatibles doit faire l'objet d'une **réflexion** de la part de la CNIL (**nouvelles modalités** pour respecter les droits des personnes).	**<u>Possibilité 2)</u> obligation légale**. **Traitements "régaliens". Avis de la CNIL.**
<u>Cas n° 3 Origine des données</u> : plusieurs RT mettent en commun différentes bases de données pour un projet commun	**<u>Possibilité 1)</u>** Les **bases de données ont été anonymisées** Présomption de base légale : l'**intérêt légitime.** **<u>Possibilité 2)</u>** Plusieurs RT souhaitent **mettre en commun des bases de données à des fins de recherche et utiliser des données personnelles disponibles sur l'Internet**. Possible de reconnaître comme base légale l'**intérêt légitime** (finalités compatibles au moins pour les trois premières). **Problématique particulière des données publiquement accessibles.** **Autorisation** si article 25-I-5° applicable. Contrôle *a priori* de la CNIL.	**<u>Possibilité 1)</u>** Mise en commun de bases de données à des **fins de profilage et de prospection commerciale**. Nécessité du **consentement** des personnes et d'**autorisation** de la CNIL **si** les traitements ont des **finalités principales distinctes**. **<u>Possibilité 2)</u>** Mise en commun de bases de données pour prendre des **décisions pouvant avoir des incidences négatives** sur les personnes. Besoin d'une **autorisation** (25-I-4° et/ou 25-I-5°), contrôle *a priori* de la CNIL (intérêt légitime peut être la base légale). **<u>Possibilité 3)</u> obligation légale**. **Traitements "régaliens"**. **Avis de la CNIL.**

Synthèse/proposition :

- Il est possible de différencier les traitements « *big data* » selon deux critères principaux : l'origine des données et l'objectif poursuivi par le traitement.

- S'agissant de l'origine des données, différents cas peuvent être distingués :
Cas n°1 le responsable de traitement met en œuvre son traitement à partir des données dont il dispose, sans avoir recours à des sources externes de données. Ces données peuvent être collectées pour la mise en œuvre du traitement « *big data* » (cas n°1.1) ou être des données dont le responsable de traitement dispose déjà et qui font l'objet d'un traitement ultérieur « *big data* » (cas n°1.2).
Cas n° 2 le responsable de traitement met en œuvre son traitement à partir de données issues en partie ou totalement de sources externes publiques.
Cas n°3 plusieurs responsables de traitements mettent en commun différentes bases de données.

- S'agissant de l'objectif et de la finalité poursuivis par le traitement, deux catégories sont identifiées : Cas n°1 les traitements ayant pour objet la détection de tendances et de corrélations entre les données.
Cas n° 2 les traitements visant à cibler la personne et à prendre, le cas échéant, des décisions à son encontre.

- Cette différenciation des traitements selon ces facteurs permet d'élaborer une typologie comportant le régime juridique applicable *a priori* au traitement : tableau en annexe 2.
Elle permet notamment de définir les traitements pouvant avoir pour base légale la poursuite de l'intérêt légitime du responsable de traitement et les traitements pour lesquels des garanties seront nécessaires (consentement, autorisation préalable, obligation légale, respect de l'article 10, transparence, etc.).

- Cette différenciation des traitements permet aussi de mettre en exergue les deux enjeux pour la CNIL en termes de régulation et les deux points de vigilance pour les responsables de traitement : quelles sont les possibilités d'ouverture pour la réutilisation des données pour d'autres finalités ? Quelles sont les conditions pour un traitement loyal et licite des données ? Ces deux points font l'objet des développements ci-après.

B. Identification des possibilités d'ouverture pour une réutilisation des données pour d'autres finalités

A titre liminaire, comme cela a déjà été indiqué, **il convient de rappeler que lorsque les données ont été anonymisées, il est possible de les réutiliser librement** pour un traitement « *big data* », ces données n'étant plus des données à caractère personnel.

S'agissant de la **réutilisation des données personnelles pour un traitement ultérieur, des possibilités sont offertes par les articles 6[86] et 36[87] de la loi** du 6 janvier 1978 modifiée qui peuvent être exploitées dans le contexte du « *big data* » et qui trouvent leur pendant dans le futur règlement européen[88].

[86] Selon l'article 6.2 de la loi, les données « *sont collectées pour des finalités déterminées, explicites et légitimes et ne sont pas traitées ultérieurement de manière incompatible avec ces finalités. Toutefois, un traitement ultérieur de données à des fins statistiques ou à des fins de recherche scientifique ou historique est considéré comme compatible avec les finalités initiales de la collecte des données, s'il est réalisé dans le respect des principes et des procédures prévus au présent chapitre, au chapitre IV et à la section 1 du chapitre V ainsi qu'aux chapitres IX et X et s'il n'est pas utilisé pour prendre des décisions à l'égard des personnes concernées* ».

[87] Selon l'article 36 de la loi, « *Les données à caractère personnel ne peuvent être conservées au-delà de la durée prévue au 5° de l'article 6 qu'en vue d'être traitées à des fins historiques, statistiques ou scientifiques ; le choix des données ainsi conservées est opéré dans les conditions prévues à l'article L. 212-4 du code du patrimoine. (…)*
Il peut être procédé à un traitement ayant des finalités autres que celles mentionnées au premier alinéa :
- soit avec l'accord exprès de la personne concernée ;
- soit avec l'autorisation de la Commission nationale de l'informatique et des libertés ;
- soit dans les conditions prévues au 8° du II et au IV de l'article 8 s'agissant de données mentionnées au I de ce même article ».

[88] Article 5.1(b) « *Personal data must be : (…) (b) collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes; further processing of personal data for archiving purposes in the public interest, or scientific and historical research purposes or statistical purposes shall, in accordance with Article 83(1), not be considered incompatible with the initial purposes; ("purpose limitation)"*

Article 5.1(e) « *Personal data must be : (…) (e) kept in form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the data will be processed solely for archiving purposes in the public interest, or scientific and historical research purposes or statistical purposes in accordance with Article 83(1) subject to implementation of the appropriate technical and organizational measures required by the Regulation in order to safeguard the rights and freedoms of the data subject ("storage limitation)".*

Ainsi, le principe de finalités déterminées n'exclut pas la possibilité de réutiliser les données pour d'autres finalités, à la condition que ces finalités soient compatibles[89]. Par le passé, la CNIL a déjà eu à se prononcer sur la compatibilité ou l'incompatibilité de certaines finalités, et certaines finalités (statistique, historique, scientifique) bénéficient d'une présomption de compatibilité cette compatibilité est confirmée par le règlement européen (considérant 40). En outre, des critères pour effectuer un test de comptabilité entre les finalités ont été dégagés par le G29, et sont repris dans le futur règlement européen.

Sur le principe de compatibilité des finalités : la doctrine actuelle de la CNIL

Il semble utile de déterminer quels types de finalités sont considérés par la CNIL comme étant compatibles ou incompatibles avec la finalité initiale, afin de tenter d'en dégager des règles applicables au « *big data* ».

Pour ce faire, une analyse a été réalisée en se référant aux normes simplifiées et aux délibérations adoptées par la CNIL (recommandations, autorisations, mises en demeure, avertissements), aux travaux du G29, d'autres autorités de protection des données et à la jurisprudence.

L'étude des normes simplifiées permet de constater que la CNIL a déjà raisonné autour de la **notion de famille de finalités**, une finalité très large se déclinant en différentes sous finalités (exemple de la norme simplifiée n° 48 précité). Toutefois, les normes simplifiées étant adoptées par secteur d'activités, et il n'est pas aisé de tirer de cette analyse des conclusions générales applicables au « *big data* », tout au moins au-delà de secteurs d'activités particuliers.

En matière d'interconnexion, il est intéressant de noter que les traitements autorisés relevaient, pour plusieurs d'entre eux, de la même base légale, à savoir l'exécution d'une **mission de service public** (article 7.3° de la loi)[90]. Cela pourrait être une piste de présomption de compatibilité à creuser, pour cette base légale.

L'analyse de la doctrine de la CNIL permet également de déterminer dans quels cas la Commission a constaté un **détournement de finalité**, ce qui permet ainsi de déduire quelles sont les finalités pour lesquelles il y a une présomption d'incompatibilité. Des détournements de finalités ont été constatés lorsque le traitement était utilisé pour mener des **actions pouvant porter préjudice aux personnes concernées ou avoir des incidences négatives à leur encontre** (par exemple le recouvrement de créance, la surveillance des salariés, etc.)[91].

Article 83.1 « *Processing of personal data for archiving purposes in the public interest, or scientific and historical research purposes or statistical purposes, shall be subject to in accordance with this Regulation appropriate safeguards for the rights and freedoms of the data subject. These safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure the respect of the principle of data minimisation. These measures may include pseudonymisation, as long as these purposes can be fulfilled by further processing of data which does not permit or not any longer permit the identification of data subjects the purposes shall be fulfilled in this manner* ». Les alinéas 2, 3 et 4 de cet article prévoient que des dérogations au respect des droits des personnes peuvent être prévues, sous certaines conditions, par les législations nationales ou européenne pour les traitements visés par cet article.

[89] Il est intéressant de noter que, plutôt que d'imposer une véritable exigence de compatibilité, le législateur a choisi une double négation en interdisant l'incompatibilité ce qui offre plus de souplesse quant aux finalités poursuivies. Le fait que le traitement de données s'effectue pour une finalité différente de celle d'origine ne signifie pas nécessairement que les finalités sont incompatibles. Cette compatibilité ou incompatibilité doit s'analyser au cas par cas.

[90] Voir, par exemple, les délibérations n° 2010-473 du 16 décembre 2010, n° 2010-474 du 16 décembre 2010, n° 2011-025 du 20 janvier 2011 et n° 2011-148 du 19 mai 2011.

[91] La délibération n° 2009-451 du 2 juillet 2009 met en demeure l'hôpital ayant communiqué à une société de recouvrement de créances les données relatives à la personne proche du débiteur, « personne à prévenir en cas de problème ». Ainsi, la finalité du traitement ultérieur, qui était le recouvrement de créances, était incompatible avec la finalité de la collecte de données qui visait la gestion de situations d'urgence au cours de l'hospitalisation d'un patient.

De même, une utilisation ultérieure de données à des fins de **prospection commerciale ou politique** est considérée comme étant un détournement de finalité. La CNIL a prononcé, dans sa délibération n° 2015-040 du 12 février 2015, un avertissement public à l'encontre du Théâtre National de Bretagne qui a utilisé les adresses électroniques des abonnés à des fins de communication politique, en réponse à un article publié par le quotidien régional à l'occasion des élections municipales. Ces données sont habituellement utilisées afin d'aborder la gestion de leur abonnement et leur adresser des informations culturelles sous la forme de newsletters électroniques. Par conséquent, une utilisation de ces données à des fins de communication politique est incompatible avec la finalité initialement déterminée. La CNIL a émis dans sa délibération n° 94-022 du 29 mars 1994 un avis défavorable au projet de la Caisse centrale de mutualité sociale agricole (CCMSA) d'utiliser le fichier des assurés des caisses départementales et pluri-départementales de mutualité sociale agricole à des fins publicitaires[92].

Sur les présomptions de compatibilité : traitement ultérieur de données à des fins statistiques ou à des fins de recherche scientifique ou historique

L'article 6.2° de la loi du 6 janvier 1978 modifiée prévoit la possibilité d'une réutilisation des données à des fins statistiques ou à des fins de recherche scientifique ou historique. Une telle réutilisation est jugée compatible avec les finalités initiales de la collecte lorsqu'elle est réalisée dans le respect des autres principes fixés par la loi, et sous réserve que le traitement ultérieur ne soit pas utilisé pour prendre des décisions à l'égard des personnes concernées[93]. Cette possibilité offerte par la loi est **particulièrement intéressante en matière de « *big data* »** puisque nombre des usages ne visent pas les personnes en tant que telles mais **l'exploitation statistique des données** les concernant.

Dans son étude annuelle 2014, le Conseil d'Etat précise à cet égard que « *le principe de finalités déterminées n'exclut pas la liberté de réutilisation statistique : dans le cadre juridique actuel, la finalité statistique est toujours présumé compatible avec la finalité initiale du traitement* ». Cette position est dans le sillage de la jurisprudence du Conseil d'Etat[94]. Des garanties de protection de la vie privée doivent toutefois être mises en place pour accompagner cette ouverture à une réutilisation statistique des données, visant notamment à garantir que ces traitements ne sont pas utilisés pour prendre des décisions à l'égard des

L'utilisation du numéro de sécurité sociale (NIR) d'enseignants, recrutés par une société pour le compte d'un groupe de soutien scolaire, à des fins d'identification des enseignants « interdits », auxquels la société ne souhaite plus attribuer de cours, est constitutif d'un détournement de finalité par rapport à la finalité déclarée qui est d'accomplir les formalités auprès des organismes de sécurité sociale imposées par la loi. (Délibération n° 2010-115 du 22 avril 2010) A ce titre a été prononcé un avertissement à l'encontre d'une société qui avait des pratiques similaires. (Délibération n° 2010-113 du 22 avril 2010).

92 L'argumentaire était que la finalité déclarée de ce fichier est la liquidation des prestations dues aux bénéficiaires du régime obligatoire de sécurité sociale agricole. L'envoi de messages publicitaires et plus généralement l'envoi d'informations, sans rapport avec l'objet des prestations sociales agricoles, sont des finalités étrangères aux missions confiées aux caisses.

93 Selon l'article 6.2 de la loi, les données « *sont collectées pour des finalités déterminées, explicites et légitimes et ne sont pas traitées ultérieurement de manière incompatible avec ces finalités. Toutefois, un traitement ultérieur de données à des fins statistiques ou à des fins de recherche scientifique ou historique est considéré comme compatible avec les finalités initiales de la collecte des données, s'il est réalisé dans le respect des principes et des procédures prévus au présent chapitre, au chapitre IV et à la section 1 du chapitre V ainsi qu'aux chapitres IX et X et s'il n'est pas utilisé pour prendre des décisions à l'égard des personnes concernées* ».

94 On peut citer pour exemple la décision IMS Health du Conseil d'Etat, du 26 mai 2014. Le Conseil d'Etat admet la compatibilité des finalités de statistiques et de recherche scientifique en vue de la réalisation d'études relatives à la consommation de médicaments avec les finalités initiales du traitement de données issues des feuilles de soins électroniques anonymisées. Ce traitement ne constitue pas un détournement de finalité. Il ressort par ailleurs de la décision du Conseil d'Etat relative au fichier « ELOI », rendue le 30 décembre 2009, qu'un traitement automatisé de données à caractère personnel relatives aux étrangers faisant l'objet d'une mesure d'éloignement peut permettre l'établissement de statistiques à propos des mesures d'éloignement et de leur taux d'exécution, à condition qu'il ne permette pas de prendre de décisions à l'égard des personnes concernées.

personnes concernées (sauf consentement spécifique des personnes pour la mise en œuvre de ce traitement)[95].

S'agissant des traitements visés par cette présomption de compatibilité, s'il peut sembler plus aisé de déterminer les traitements réalisés à des fins statistiques ou à des fins de recherche historique, il est moins évident de déterminer quels traitements peuvent entrer dans ce régime particulier car ils sont mis en œuvre à des **fins de recherche scientifique**. L'analyse des délibérations adoptées jusqu'à présent par la CNIL révèle une conception plutôt stricte de la notion de recherche scientifique[96], mais cette analyse est uniquement fondée sur les délibérations concernant des traitements soumis à autorisation préalable dans le domaine de la recherche scientifique. Cela ne signifie pas, pour autant, que la CNIL ne pourrait pas avoir une conception plus large des traitements visés par cette présomption de compatibilité, au-delà par exemple des recherches menées dans le domaine médical ou mises en œuvre par des organismes publics de recherche.

D'autres autorités retiennent en effet une **définition** assez large de la notion de recherche scientifique. Par exemple, l'autorité belge de protection des données considère qu'une « *recherche scientifique vise à établir des permanences, des lois de comportements ou des schémas de causalité qui transcendent tous les individus qu'ils concernent* »[97]. Par ailleurs, dans son rapport *« big data and data protection* », l'Information Commissioner's Office (ICO), indique que lorsque des données personnelles collectées pour une finalité initiale sont ensuite réutilisées pour de la recherche, celle-ci ne constitue pas une finalité incompatible[98]. Selon l'ICO, bien que le terme de « recherche » ne soit pas défini par le Data Protection Act (DPA) il n'inclut pas seulement les recherches scientifiques ou historiques mais également les recherches à des fins commerciales, comme une étude de marché.

Sur les critères dégagés par le G29 et le futur règlement européen pour effectuer un test de compatibilité pour la réutilisation des données pour d'autres finalités

Les **travaux du G29** permettent également d'apporter un éclairage sur la question des finalités compatibles. Depuis 2013, afin de déterminer si des finalités sont compatibles, celui-ci a proposé **quatre critères** qui ont vocation à s'appliquer à tout traitement ultérieur de données. Dans une appréciation au cas par cas, il s'agit de prendre en compte :

- la relation entre les finalités,
- le contexte et les attentes des personnes concernées,
- la nature des données personnelles et l'impact sur les personnes concernées d'un traitement ultérieur de leurs données,
- les garanties adoptées[99].

Le projet de règlement européen suit également cette logique en reprenant ces critères à son article 6.3 a) et dans son considérant 40.

[95] Le G29, la Commission pour la protection de la vie privée en Belgique et le projet de règlement européen à l'article 83 prévoient également pour les traitements ultérieurs à des fins statistiques, de recherche historique ou scientifique des mesures de protection additionnelles : anonymisation, pseudonymisation ou séparation fonctionnelle. Le G29 définit la séparation fonctionnelle comme la mise en place de mesures techniques et organisationnelles afin de s'assurer que les données ne puissent être utilisées pour prendre des décisions ou d'autres actions dans le respect des individus, ce qui est déjà prévu par l'article 6.2° de la loi du 6 janvier 1978.

[96] Voir le document de synthèse « La CNIL et la régulation du Big data » (pages 34 à 39) préparé par Margaux Deuchler et disponible sous U dans le dossier thématique sur le Big data.

[97] Voir le site internet de la Commission de protection de la vie privée http://www.privacycommission.be/fr/lexique/fins-historiques-statistiques-ou-scientifiques.

[98] Big data and data protection, ICO https://ico.org.uk/media/for-organisations/documents/1541/big-data-and-data-protection.pdf

[99] Opinion 03/2013 WP 29 on purpose limitation.

Dans l'avis du G29, le raisonnement suivi pour chaque critère est le suivant. Pour le premier critère, il s'agit d'analyser la **relation entre les finalités** pour lesquelles les données ont été collectées et les finalités pour lesquelles elles sont traitées ultérieurement. L'analyse peut couvrir des situations dans lesquelles soit le traitement ultérieur était déjà **plus ou moins implicite** par rapport aux finalités initiales, soit il s'inscrit dans la **suite logique** de la collecte des données et le but de cette collecte, soit le traitement ultérieur n'a qu'un **lien partiel voire inexistant** avec les finalités initiales. Dans tous les cas, plus la distance entre la finalité de la collecte des données et celle de leur traitement ultérieur est grande, plus il sera difficile d'apprécier leur compatibilité (même s'il reste possible de conclure à la compatibilité, par exemple lorsque les données sont anonymisées, exemple 15 de l'avis du G29).

Le second critère est le **contexte** dans lequel les données ont été collectées et les **attentes** raisonnables des personnes concernées par rapport à un autre usage de ces données. Il s'agit d'étudier la **relation et la répartition des pouvoirs** entre le gestionnaire de données et la personne concernée (en particulier, lorsque la communication des données était imposée par la loi ou prévue dans un contrat), la **manière dont les données ont été communiquées** (librement ou non) et **ce qui pourrait être habituellement et généralement attendu** dans ce contexte et compte tenu de la relation donnée. Plus la différence est importante entre la finalité initiale d'un traitement et les opérations ultérieures, plus le contrôle de compatibilité devra impliquer un examen approfondi et détaillé du contexte.

Le G29 cite de nombreux **exemples** dans son avis pour illustrer comment réaliser ce test de compatibilité. Il mentionne ainsi le recours par une société à des algorithmes à l'insu des personnes pour identifier, à partir de leur historique d'achats, les clientes enceintes, le stade de leur grossesse, et leur proposer des offres ciblées. Aucune information n'est fournie aux clientes lors de la souscription de la carte de fidélité, si ce n'est que les données de la carte de fidélité seront utilisées à des fins marketing, pour fournir notamment des offres spéciales et des coupons de réduction. La réutilisation des données pour cette finalité est jugée incompatible.

Le troisième critère est relatif à la **nature des données** personnelles et **l'impact** sur les personnes concernées d'un traitement ultérieur de leurs données. Le fait que les données concernées soient des **données sensibles** apporte un indice sur les exigences à appliquer quant à l'évaluation des compatibilités. Dans cette évaluation, il est nécessaire de prendre en compte **toutes les conséquences du traitement des données, aussi bien positives que négatives** : décisions ou actions potentielles d'un tiers, exclusion ou discrimination d'individus, impact émotionnel etc. et aussi la **manière dont les données sont traitées ultérieurement** : par un gestionnaire extérieur, accessibilité à un large nombre de personnes, grande quantité de données traitées et combinées avec d'autres, etc. De manière générale, plus l'impact sur les personnes concernées d'un traitement ultérieur de leurs données est négatif ou incertain, plus il sera difficile de considérer qu'il s'agit d'un usage compatible.

Le quatrième critère concerne les **garanties** adoptées par le gestionnaire de données pour assurer un traitement juste et pour prévenir tout impact excessif sur les personnes concernées. Cela correspond à l'examen des **mesures techniques et organisationnelles** permettant une **séparation fonctionnelle**[100] afin de s'assurer que les données **ne puissent être utilisées pour prendre des décisions** ou d'autres actions à l'encontre des individus (anonymisation, pseudonymisation et agrégation de données) mais aussi des **mesures supplémentaires** au profit des personnes concernées (transparence, possibilité de proposer

[100] Selon le G29, la séparation fonctionnelle implique que les données utilisées à des fins de recherche, à des fins statistiques et de détection de tendances ou de corrélations ne doivent pas être utilisées pour la prise de décision ou de mesures à l'égard des personnes dont les données sont collectées.

ou d'objecter sur un contenu spécifique). **S'agissant des garanties appropriées, le futur règlement européen cite le chiffrement et la pseudonymisation**.

L'avis du G29 contient également des développements spécifiques sur le *« big data »* et propose des garanties différentes pour les deux types de traitement ultérieur compatible (qui correspondent aux deux cas identifiés au premier point de cette partie) : premièrement, quand ce traitement ultérieur est fait pour **détecter des tendances ou des corrélations** et deuxièmement ; lorsqu'il est fait **pour en apprendre davantage** sur les individus et **pour prendre des décisions** les concernant.

Dans le premier cas, le G29 recommande une **séparation fonctionnelle stricte** entre les deux analyses. Les gestionnaires de données doivent pour cela **garantir la confidentialité et la sécurité des données**. Dans le second, « *un **consentement** libre, spécifique, éclairé, non-ambigu et exprès sera presque toujours requis, sinon un traitement ultérieur ne peut être considéré compatible* ». Afin de donner un consentement éclairé et d'assurer la transparence des traitements effectués, les personnes concernées doivent disposer de moyens de contrôle effectifs[101].

Cette analyse nous invite donc à nouveau à établir un statut différencié et des conditions de mise en œuvre distinctes pour les traitements « *big data* » en fonction des deux objectifs principaux identifiés (cf. tableau récapitulatif en annexe 2). Dans tous les cas, le fait qu'il soit possible de réutiliser les données pour des finalités compatibles ne signifie pas que cette réutilisation ne doit pas être faite dans le respect des principes de protection des données.

Dans son avis, le G29 indique qu'en **cas d'incompatibilité**, un retour vers les personnes concernées et une nouvelle base légale, notamment le consentement, peuvent permettre dans certains cas de compenser l'incompatibilité. Toutefois, en principe, le G29 considère qu'un traitement incompatible ne saurait être légitimé en s'appuyant sur un nouveau fondement légal, qui viserait à « couvrir » l'incompatibilité (l'exigence de compatibilité posée par l'article 6 1. (b) de la directive et celle d'une base juridique appropriée posée par l'article 7 sont cumulatives). Sur ce point, **le futur règlement européen adopte une approche plus ouverte en restreignant le recours au test de compatibilité pour les traitements n'étant pas fondé sur le consentement des intéressés ou l'existence d'une législation nationale ou européenne** (article 6.3 (a) et considérant 40). Votre rapporteur propose d'adopter d'ores et déjà cette approche.

Synthèse/proposition :

- Le principe est qu'il est possible de réutiliser des données personnelles pour d'autres finalités à la condition que ces finalités soient compatibles. Il est donc possible de favoriser les possibilités d'exploitation et de réutilisation des données dans un contexte « *big data* », dans le respect de la législation.

- Sur le principe de la compatibilité des finalités, la doctrine de la CNIL contient des enseignements intéressants en matière de « *big data* ». Il est possible de réfléchir à partir de la notion de famille de finalités et d'usages compatibles. Par ailleurs, au vu des décisions relatives à des détournements de finalité, il existe une présomption d'incompatibilité lorsque les données sont réutilisées pour mener des actions pouvant porter préjudice aux personnes, ou avoir des incidences négatives à leur encontre, ou pour effectuer un ciblage particulier (prospection politique ou commerciale).

[101] Le G29 recommande notamment que les personnes aient accès à leur « profil » ainsi qu'à la logique de l'algorithme qui a généré ces profils. De plus, la source de données qui a mené à la création de ce profil doit aussi être divulguée. Compte tenu du risque d'inférences inexactes en particulier, il est crucial que les personnes concernées puissent corriger et actualiser leur profil si elles le souhaitent.

- Il y a en revanche une présomption de compatibilité pour le traitement ultérieur des données à des fins statistiques ou à des fins de recherche scientifique ou historique. Cette possibilité est particulièrement intéressante en matière de « *big data* » puisque nombre des usages ne visent pas les personnes en tant que telles mais l'exploitation statistique des données les concernant. Des garanties de protection de la vie privée doivent toutefois être mises en place pour accompagner cette ouverture à une réutilisation statistique des données, visant notamment à garantir que ces traitements ne sont pas utilisés pour prendre des décisions à l'égard des personnes concernées (sauf consentement spécifique des personnes pour la mise en œuvre de ce traitement).

- Quatre critères principaux permettent d'effectuer un test de compatibilité pour le traitement ultérieur des données pour d'autres finalités :
1) la relation entre les finalités,
2) le contexte et les attentes des personnes concernées,
3) la nature des données personnelles et l'impact sur les personnes concernées d'un traitement ultérieur de leurs données,
4) les garanties adoptées (séparation fonctionnelle, anonymisation, pseudonymisation, transparence et droits des personnes, etc.).

- Affranchissement de la notion de compatibilité et du test susmentionné quand le traitement ultérieur des données est réalisé avec le consentement de la personne ou en application d'une législation nationale ou européenne (dispositions du règlement européen).

C. Identification des conditions pour un traitement loyal et licite des données

Comme cela a été précédemment mentionné, votre rapporteur considère que le développement du « *big data* » ne saurait se faire en méconnaissant les principes fondamentaux de la protection des données, et notamment les obligations en termes d'information préalable, qui conditionnent l'effectivité des droits garantis aux personnes. En effet, **favoriser les possibilités d'exploitation et de réutilisation des données dans un contexte « *big data* » ne doit pas se faire au détriment des droits des personnes**, et une approche ouverte de ces possibilités doit être contrebalancée par une transparence réelle à l'égard des personnes concernées. Cette approche est conforme au **droit à l'autodétermination informationnelle**, dont l'introduction à la loi du 6 janvier 1978 modifiée est proposée par le projet de loi pour une République numérique (article 26). Cette notion consacre le droit à la libre disposition de ses données, c'est-à-dire le droit de l'individu de maîtriser les usages qui sont faits de ses données à caractère personnel. L'exercice de ce droit, ainsi que des droits « informatique et libertés » traditionnels, nécessite que les traitements « *big data* » soient mis en œuvre de manière transparente pour les personnes.

Cette transparence passe d'abord par l'information des personnes concernées et la possibilité d'exercer leurs droits « informatique et libertés ».

A cet égard, votre rapporteur souhaite préciser en premier lieu que, **lorsque l'application « classique » de la loi est possible[102], celle-ci doit être respectée** et cet impératif doit être rappelé aux responsables de traitements concernés.

[102] Il convient de rappeler que certains traitements (régaliens notamment) sont soumis à des règles particulières : articles 32.V, 32.VI, paragraphe 3 de l'article 38, articles 41 et 42.

Dans de nombreux cas, les modalités d'information et d'exercice des droits des personnes ne posent pas de difficulté particulière lorsque les données sont directement collectées par le responsable de traitement auprès de la personne concernée pour la mise en œuvre de son projet « *big data* ». Qu'il s'agisse d'utilisation statistique d'informations relatives à un client ou de la mise en œuvre d'un traitement tel que « *Pay as you drive* », par exemple, le responsable de traitement est en mesure d'informer la personne conformément à l'article 32 de la loi[103] - et le cas échéant de recueillir son consentement pour la mise en œuvre du traitement - au moment de la collecte des données, comme le prévoit la loi. Votre rapporteur souhaite également souligner qu'il est possible de collecter et de traiter des données pour différentes finalités (exemple précité des finalités visées par la norme simplifiée n° 48 relative à la gestion des clients et des prospects) et d'en informer les personnes concernées au moment de cette collecte. Dans le considérant 25 du règlement européen, traitant du consentement, il est d'ailleurs précisé que, lorsque le traitement a plusieurs finalités, un consentement devrait être donné pour l'ensemble des finalités du traitement.

Lorsque les données sont réutilisées par le responsable de traitement pour d'autres finalités que celles pour lesquelles il les avait initialement collectées, votre rapporteur considère que le traitement doit par principe être mis en œuvre dans le respect des droits des personnes, qui doivent être informées et mises en mesure d'exercer leurs droits, notamment d'opposition. Aucun obstacle n'empêche le responsable de traitement de contacter les intéressés, avec lesquels il est ou a été précédemment en relation pour le traitement initial. Le futur règlement européen prévoit d'ailleurs explicitement ce cas de figure[104]. Selon la finalité du traitement ultérieur, le responsable de traitement devra également revenir vers la personne concernée pour recueillir son consentement à la mise en œuvre des traitements (cf. développements précédents). En cas de collecte dite indirecte des données personnelles, par exemple auprès de partenaires commerciaux, la personne doit avoir été initialement informée, au moment de la collecte des données, des destinataires ou catégories de destinataires. Selon l'article 32.III de la loi du 6 janvier 1978 modifiée « *Lorsque les données à caractère personnel n'ont pas été recueillies directement auprès de la personne concernée, le responsable du traitement ou son représentant doit fournir à cette dernière les informations énumérées au I dès l'enregistrement des données ou, si une communication des données à des tiers est envisagée, au plus tard lors de la première communication des données* ». Le futur règlement européen contient également des dispositions similaires sur ce point[105].

Votre rapporteur considère, qu'**au cas par cas, il pourrait être décidé d'utiliser les dispositions de l'article 32.III de la loi** du 6 janvier 1978 modifiée prévoyant que l'obligation d'information des personnes ne s'applique pas lorsque la « *personne concernée est déjà informée ou quand son information se révèle impossible ou exige des efforts*

---

[103] « *La personne auprès de laquelle sont recueillies des données à caractère personnel la concernant est informée, sauf si elle l'a été au préalable, par le responsable de traitement ou son représentant : 1° De l'identité du responsable de traitement et, le cas échéant, de son représentant ; 2° De la finalité poursuivie par le traitement auquel les données sont destinées ; 3° Du caractère obligatoire ou facultatif des réponses ; 4° Des conséquences éventuelles, à son égard, d'un défaut de réponse ; 5° Des destinataires ou catégories de destinataires des données ; 6° Des droits qu'elle tient des dispositions de la section 2 du présent chapitre (droits des personnes à l'égard des traitements de données) ; 7° Le cas échéant, des transferts de données à caractère personnel envisagés à destination d'un Etat non membre de la Communauté européenne* ».

[104] Article 14.1b *"Where the controller intends to further process the data for a purpose other than the one for which the data were collected the controller shall provide the data subject prior to that further processing with information on that other purpose and with any relevant further information as referred to in paragraph 1a".*
Article 14.5 *"Paragraphs 1, 1a and 1b shall not apply where and insofar as the data subject already has the information".*

[105] L'article 14a du futur règlement européen définit de manière similaire les informations qui doivent être fournies par le responsable de traitement lorsque les données n'ont pas été recueillies auprès des personnes concernées, en envisageant dans cet article également un alinéa spécifique pour le traitement ultérieur des données pour d'autres finalités (alinéa 3).

disproportionnés par rapport à l'intérêt de la démarche »[106]. Cette possibilité pourrait être intéressante lorsque les données utilisées ont vocation à être anonymisées à bref délai[107] (les personnes n'ont donc plus de droit à exercer une fois l'anonymisation réalisée) pour les traitements ayant vocation à détecter des tendances ou des corrélations entre les données, à établir des statistiques (cas n°1 du facteur de différenciation objectif du traitement). Le futur règlement européen va d'ailleurs en ce sens[108]. L'article 32.IV[109] prévoit en outre une information allégée en cas d'anonymisation des données.

Au-delà de ce rappel des règles générales, et de leur application possible dans le cadre de nombreux traitements « *big data* », votre rapporteur souhaite souligner le **cas particulier de la collecte et l'exploitation des données personnelles publiquement accessibles sur Internet**.

La position de la CNIL est que les données personnelles rendues publiques ne changent pas de nature et méritent protection. Aussi, la CNIL a considéré que la collecte de données personnelles dans l'espace public de l'Internet est déloyale lorsqu'elle est réalisée à l'insu des personnes ou lorsque l'information n'est pas satisfaisante. La position de la CNIL a été confirmée par la jurisprudence notamment en matière de lutte contre le spam[110] ou lors de la confirmation de l'avertissement Pages Jaunes.

Dans un contexte de banalisation de la ré-exploitation massive des données mises en ligne sur internet, et face aux difficultés de régulation de ces pratiques, votre rapporteur observe que le maintien d'une position stricte et classique en matière de collecte déloyale dans un contexte « *big data* » (et notamment imposer une obligation d'information directe), pour l'ensemble des traitements quelle que soit leur finalité, produit des conséquences importantes :
- imposer une obligation d'information directe et préalable des personnes concernées reviendrait, dans de nombreux cas, à rendre la mise en œuvre du traitement impossible : en effet, les responsables de traitements concernés ne sont pas en relation directe avec les

[106] La Commission a déjà autorisé le recours à cet article, voir par exemple la délibération n°2013-031 du 24 janvier 2013 autorisant l'INAVEM à mettre en œuvre un traitement automatisé de données à caractère personnel ayant pour finalité le suivi des activités des associations d'aide aux victimes adhérentes et l'établissement de statistiques ou le courrier du 29 octobre 2008 adressé à la société Orange pour son traitement « info trafic ». Il convient aussi de préciser que pour les traitements spécifiquement mentionnés, l'article 32.III prévoit également que « *lorsque les données à caractère personnel ont été initialement recueillies pour un autre objet, les dispositions de l'alinéa précédent ne s'appliquent pas aux traitements nécessaires à la conservation de ces données à des fins historiques, statistiques ou scientifiques, dans les conditions prévues au livre II du code du patrimoine ou à la réutilisation de ces données à des fins statistiques dans les conditions de l'article 7 bis de la loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques ».*

[107] Cette dispense ne couvrirait que les traitements big data dynamiques (pour lesquels les données sont corrélées en quasi temps-réel). Les traitements conservant les données sur un temps plus long en seraient exclus.

[108] L'article 14a.4 prévoit que "*Paragraphs 1 to 3a shall not apply where and insofar as :*
a) the data subject already has the information; or
b) the provision of such information proves impossible or would involve a disproportionate effort; in particular for processing for archiving purposes in the public interest, or scientific and historical research purposes subject to the conditions and safeguards referred to in Article 83(1) or in so far as the right referred to in paragraph 1 is likely to render impossible or seriously impair the achievements of the objectives of the archiving purposes or the statistical purposes; in such cases the controller shall take appropriate measure to protect the data subjects' rights and freedoms and legitimate interests, including making the information publicly available; or
c) obtaining or disclosure is expressly laid down by Union or Member State law to which the controller is subject, which provides appropriate measures to protect the data subject's legitimate interests; or
d) where the data must remain confidential subject to an obligation of professional secrecy regulated by Union or Member State law, including a statutory obligation of secrecy"

[109] « *Si les données à caractère personnel recueillies sont appelées à faire l'objet à bref délai d'un procédé d'anonymisation préalablement reconnu conforme aux dispositions de la présente loi par la Commission nationale de l'informatique et des libertés, les informations délivrées par le responsable du traitement à la personne concernée peuvent se limiter à celles mentionnées au 1° et au 2° du I* ».

[110] Arrêt Cass. crim. 14 mars 2006 : *« est déloyal le fait de recueillir, à leur insu, des adresses électroniques personnelles de personnes physiques dans l'espace public d'Internet, ce procédé faisant obstacle à leur droit d'opposition ».*

personnes, et ne disposent pas toujours du moyen de les informer individuellement de la réutilisation de leurs données ;
- lorsque des coordonnées pourraient être utilisées pour contacter la personne, imposer une obligation d'information directe, qui pourrait par exemple être faite par courriel, pour des traitements mis en œuvre à des fins de recherche ou d'établissements de statistiques pourrait être vécu comme plus intrusif par les personnes concernées que l'utilisation même de leurs données pour ces finalités. Avec la multiplication éventuelle de ces traitements, cette démarche pourrait même être perçue comme du spam ;
- *in fine*, cela pourrait contribuer à la généralisation de l'utilisation de ces données à l'insu des personnes ou à une information dégradée fournie dans les conditions générales d'utilisation des sites internet et des réseaux sociaux, rarement lues par les utilisateurs, information qui pourraient permettre à ces acteurs de soutenir que les personnes ont d'ores et déjà été informées de la collecte au sens de l'article 32-III de la loi du 6 janvier 1978 modifiée.

Il ne semble pour autant pas opportun de généraliser le recours à la dérogation à l'obligation d'information des personnes (passage de l'article 32.III précité), l'importance de maintenir l'individu au cœur du processus, pour qu'il conserve la maîtrise de l'usage de ses données, ayant déjà été soulignée, et les enjeux des traitements « *big data* » ayant pour objet de cibler les individus et de prendre des décisions à leur égard ayant déjà été évoqués.

Votre rapporteur encourage cependant la **recherche de solutions innovantes et de nouvelles modalités d'application des principes fondamentaux** pour les traitements « *big data* » visant à détecter des tendances et des corrélations et, le cas échéant, pour certaines réutilisations des données lorsque les finalités ont été jugées compatibles. Les modalités pourraient donc à nouveau être différenciées selon les traitements.

Il propose ainsi la piste de réflexion suivante.

Le consentement de la personne à une réutilisation de ses données pour certaines finalités pourrait être donné au moment de la mise en ligne des informations la concernant, par exemple lors de la création d'un compte sur un réseau social. Lors de la première étape, au moment de la diffusion des informations, la personne pourrait donner son consentement à la réutilisation de ses données par type de finalités/famille de finalités (à des fins statistiques, à des fins de recherche, ou encore, pour un réseau social professionnel, à des fins de recrutement) et être informée (voire consentir) des destinataires ou catégories de destinataires (instituts de recherche, sociétés commerciales, etc.). Un signe distinctif pourrait permettre d'identifier sur la page sur laquelle les informations sont diffusées le consentement donné par l'individu pour une réutilisation de ses données (dans la même logique que l'ajout d'un pictogramme dans les annuaires permet d'identifier les personnes ayant choisi de s'inscrire sur la liste « anti-prospection »). La personne pourrait à tout moment revenir sur ses choix.

Dans un second temps, une plate-forme accessible sur le site source permettrait aux personnes de savoir quels organismes ont utilisé leurs données, pour quelle finalité/famille de finalités (pour laquelle elle avait donc préalablement donné son consentement) et d'exercer leurs droits auprès de cet organisme. Il serait également possible de prévoir un système de notification par le biais du site, sur la plate-forme, au moment de la réutilisation des données par le responsable de traitement. Une telle solution constituerait un compromis acceptable dès lors qu'elle permettrait de rejeter dans la clandestinité l'aspiration sans information et consentement des données publiées.

La mise en œuvre de ces pistes de solution dépend en grande partie de l'adhésion des acteurs à leur pertinence. Il est possible qu'elles ne correspondent pas à leurs attentes ou à leurs besoins, ou que les acteurs aient des propositions qui pourraient être plus adéquates.

Aussi, votre rapporteur considère qu'il serait intéressant d'inviter les acteurs à réfléchir à cette question, et à mettre leur potentiel en matière d'innovation au service de la recherche de solutions et de nouveaux outils.

Il souligne également que la CNIL dispose de pouvoirs et de possibilités pour faire des propositions en ce sens (article 11 de la loi), et que cette mission de la CNIL est consacrée par le projet de loi pour une République numérique. Son article 29 vise à élargir les missions de la CNIL pour qu'elle joue un rôle plus en amont, en soutenant le développement de technologies respectueuses de la vie privée (« *Privacy by design* ») et en accompagnant davantage les responsables de traitements. Ce projet de loi consacre également la mission éthique de la CNIL, l'invitant à conduire une réflexion sur les problèmes éthiques et les questions de société soulevées par l'évolution des technologies.

Cet objectif de transparence à l'égard des personnes concernées passe ensuite par des moyens d'action et de contrôle renforcés.

Pour les traitements ayant un impact sur les personnes concernées, la deuxième proposition faite par votre rapporteur est de renforcer les moyens de contrôle des intéressés.

Tout d'abord, pour ces traitements, le **consentement doit rester une base légale centrale**. Mais ce consentement doit être libre, spécifique et informé. Et, pour contrebalancer les risques présentés par certains traitements « *big data* » en termes de discrimination ou de fausse inférence, l'information des personnes et la **transparence** quant à la manière dont le traitement est mis en œuvre est primordiale. Cela devrait comprendre une information et une transparence **sur les données effectivement traitées et sur la logique qui sous-tend le traitement**.

Cela permettrait aux individus, dans une certaine mesure, de vérifier et de contrôler si les conclusions qui sont tirées les concernant sont exactes et justes, et d'en contester la logique. Les personnes concernées pourraient ainsi mieux comprendre, et éventuellement rectifier, les décisions prises à leur encontre. Cela permet également un exercice effectif des droits, notamment de rectification et de suppression lorsque les données utilisées sont inexactes, incomplètes, équivoques ou périmées (article 40 de la loi).

S'agissant de l'objectif de transparence et de compréhension des algorithmes, **l'article 39 de la loi offre déjà des possibilités**, l'exercice du droit d'accès permettant à la personne notamment d'obtenir :
- des informations relatives aux finalités du traitement et aux catégories de données à caractère personnel traitées ;
- la communication, sous une forme accessible, des données personnelles qui la concernent ainsi que toute information disponible quant à l'origine de celles-ci ;
- les informations permettant de connaître et de contester la logique qui sous-tend le traitement automatisé en cas de décision prise sur le fondement de celui-ci et produisant des effets juridiques à l'égard de l'intéressé.

Ces principes sont repris à l'article 15 du futur règlement européen[111]. Le futur droit à la portabilité des données pourra également concourir à cet objectif du renforcement des droits et moyens de contrôle des individus.

Cependant, actuellement, ces informations sont communiquées lorsque la personne en fait la demande. **Un renforcement de la transparence envers les personnes et des moyens**

[111] L'article 15.1 (h) prévoit que la personne peut obtenir du responsable de traitement des informations concernant « *the existence of automated decision making including profiling referred to in Article 20(1) and (3) and at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject* ».

de contrôle dont elles disposent pourrait militer en faveur d'une communication spontanée et non conditionnée à la demande formulée par les personnes concernées. Le futur règlement européen va d'ailleurs dans ce sens en prévoyant une telle information[112]. Votre rapporteur propose ainsi que la CNIL pousse dès à présent une mise en conformité des responsables de traitements avec les dispositions qui seront prochainement en vigueur. Les dispositions du projet de loi pour une République numérique traitant de la transparence des algorithmes des administrations militent également en ce sens (articles 2 et 4).

Pour les acteurs de l'Internet ayant des **sites internet** sur lesquels les personnes disposent d'un **compte** auquel elles peuvent se connecter, **l'information sur leur « profil », les données traitées et inférées et la logique de l'algorithme pourrait être accessible dans cet espace**. Les personnes pourraient par ce biais corriger et actualiser aisément leur profil et les données les concernant. Pour les administrations, l'article 4 du projet de loi pour une République numérique prévoit que « *Les administrations mentionnées au premier alinéa de l'article L. 300-2, à l'exception des personnes morales dont le nombre d'agents ou de salariés est inférieur à un seuil qui ne peut être supérieur à cinquante agents ou salariés, fixé par décret, rendent publics en ligne, dans un standard ouvert et aisément réutilisable, les règles définissant les principaux traitements algorithmiques utilisés dans l'accomplissement de leurs missions lorsqu'ils fondent des décisions individuelles* ».

En outre, pour donner aux individus des garanties appropriées concernant les algorithmes prédictifs utilisés pour prendre des décisions à leur encontre, l'étude annuelle 2014 du Conseil d'Etat formule des préconisations et propositions, auxquelles votre rapporteur souscrit totalement :

- **assurer l'effectivité de l'intervention humaine dans la prise de décision**. Pour assurer l'effectivité de l'interdiction de fonder une décision sur la seule mise en œuvre d'un traitement automatisé (article 10 de la loi), l'intervention humaine dans la décision doit être réelle et pas seulement formelle[113]. Si dans 99 % des cas, la mesure prise est celle proposée par le système automatique, on peut suspecter que le système présenté comme une « aide à la décision » est en réalité un système de décision.
A cet égard, le Conseil d'Etat propose d'indiquer dans un instrument de droit souple les critères d'appréciation du caractère effectif de l'intervention humaine, via une recommandation de la CNIL ou un avis du G29. Les services de la CNIL pourraient s'atteler à de tels travaux.

---

112 L'article 14.1a prévoit que " *In addition to the information referred to in paragraph 1, the controller shall, at the time when personal data are obtained, provide the data subject with the following further information necessary to ensure fair and transparent processing (…) (h) the existence of automated decision making including profiling referred to in Article 20(1) and (3) and at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject".*

113 Le futur règlement européen contient également ce principe, tout en prévoyant des exceptions. Article 20 « *1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.*
1a. Paragraph 1 shall not apply if the decision :

a) is necessary for entering into, or performance of, a contract between the data subject and a data controller; or

b) is authorized by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
c) is based on the data subject's explicit consent
1b In cases referred to in paragraph 1a (a) and (c) the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision
3.Decisions referred to in paragraph 1a shall not be based on special categories of personal data referred to in Article 9(1), unless points (a) or (g) of Article 9(2) apply and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place »
Définition "profiling means any form of automated processing of personal data consisting of using those data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements". Voir également le considérant 58.

- **veiller à la non-discrimination et développer le contrôle des résultats produits par les algorithmes**[114]. Selon le Conseil d'Etat, dans le cadre de l'article 44 de la loi, et dans le respect du secret industriel, il serait possible de développer le contrôle des algorithmes par l'observation de leurs résultats, notamment pour détecter les discriminations illicites, en renforçant à cette fin les moyens humains dont dispose la CNIL. Les discriminations pourraient aussi être révélées par des opérations de *testing*.

- **mettre en place des garanties de procédure et de transparence**. Cette préconisation rejoint celle formulée ci-dessus par votre rapporteur. Il conviendrait d'imposer aux auteurs de décisions s'appuyant sur la mise en œuvre d'algorithmes une obligation de transparence sur les données personnelles utilisées par l'algorithme et le raisonnement général suivi par celui-ci. Comme le prévoit l'article 10 de la loi et le futur règlement européen, la personne faisant l'objet de la décision doit avoir la possibilité de faire valoir ses observations

Les actions que la CNIL pourrait engager dans ce domaine s'inscrivent pleinement dans sa mission de réflexion sur les problèmes éthiques et de soutien au développement de technologies respectueuses de la vie privée, mission consacrée dans le projet de loi pour une République numérique (article 29).

Les solutions techniques à explorer pour un traitement respectueux de la vie privée.

Plusieurs initiatives tentent de concilier « *big data* » et « *privacy* ». Si certaines d'entre elles (comme le CASD) sont opérationnelles, ces architectures techniques relèvent encore largement de la recherche.

En France, depuis 2010, le CASD (Centre d'Accès Sécurisé Distant) offre un accès sécurisé aux chercheurs pour l'utilisation de données de la statistique française. Celui-ci rend possible des traitements statistiques sur des données relatives à des personnes individuelles ou des ménages tout en s'assurant de l'impossibilité de les détourner à d'autres fins, accidentellement ou intentionnellement. A la différence des premiers centres dans lesquels les chercheurs devaient se rendre physiquement, le CASD offre un accès distant. Le principe consiste à mettre à disposition de l'utilisateur un terminal appelé « SD Box » qui lui permet de se connecter et de travailler sur le serveur centralisé contenant les données et installé au Genes (Groupe des écoles nationales d'économie et de statistique). Il peut ainsi voir les données sur lesquelles il travaille, effectuer des traitements statistiques à l'aide de logiciels installés sur la SD Box, produire des tableaux et graphiques, *etc.* sans pour autant pouvoir enregistrer ou imprimer les données, celles-ci ne quittant jamais le serveur centralisé[115].

[114] Les auteurs de l'ouvrage « Big data : la révolution des données est en marche » proposent la création d'une profession d' « algorithmiste » composée d'expert en science informatique et en statistique, soumis à une déontologie et à des contrôles analogues à ceux des professions comme les médecins ou les commissaires au compte, et qui procéderaient à des contrôles internes aux entreprises ou externes, sous la responsabilité des gouvernements, pour vérifier la validité des algorithmes.

[115] Techniquement, la communication entre la SD Box et le serveur du Genes est assurée par une liaison sécurisée et chiffrée. De plus, l'utilisation de la SD Box nécessite l'attribution d'une carte individuelle de connexion. Celle-ci contient un gabarit des empreintes digitales de l'utilisateur nécessaire pour authentifier son accès au CASD. Les terminaux SD Box et autorisations de connexion sont délivrés au chercheur après avis du comité du secret statistique qui étudie sa demande et accomplissement des formalités requises auprès de la CNIL. De plus, dans certains cas, le chercheur peut souhaiter effectuer des appariements de données, par exemple pour enrichir les informations recueillies lors d'une enquête par des données administratives ou issues d'une autre enquête. Dans ce cas, si le chercheur ne peut effectuer ce croisement de données lui-même – car il nécessiterait l'accès à des informations directement identifiantes (comme le NIR par exemple) – celui-ci peut-être réalisé par l'Insee hors du dispositif CASD (et avec accord de la CNIL). Le fichier résultant, préalablement expurgé des informations identifiantes, est ensuite mis à disposition du chercheur dans le CASD.

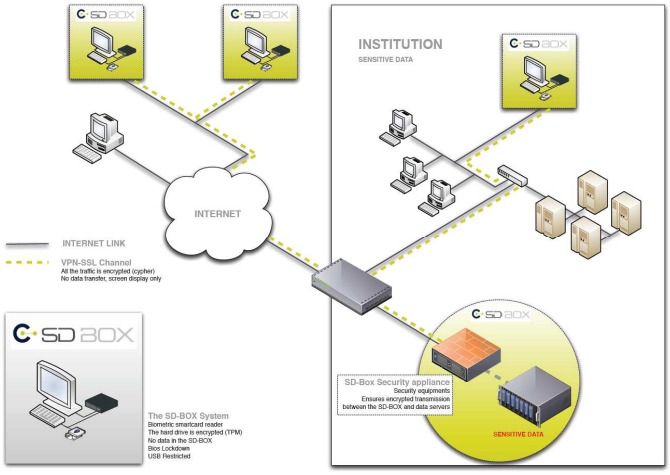

Figure 2. Architecture technique du CASD.

Actuellement, de nombreuses recherches sont menées sur le sujet du chiffrement homomorphe (*homomorphic encryption*)[116]. En cryptographie, un système est dit homomorphe lorsqu'il possède des caractéristiques algébriques lui permettant la commutation d'opérateurs mathématiques. Une telle propriété permet en particulier de confier des calculs à un agent externe, sans que les données ni les résultats ne soient accessibles à cet agent ; c'est-à-dire que des opérations peuvent être réalisées sur des données préalablement chiffrées, le résultat de ce calcul étant le résultat chiffré des opérations sur les données en clair (voir Figure 1).

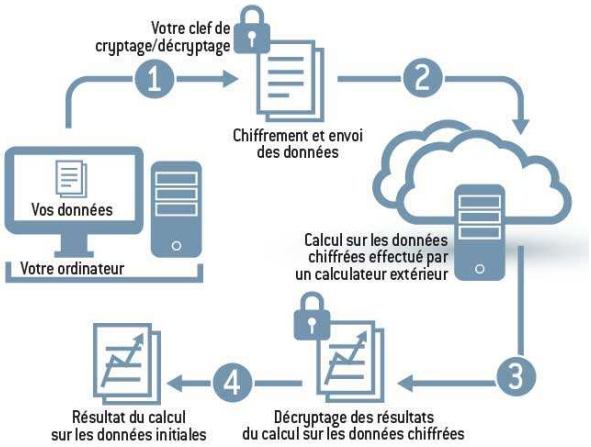

Figure 1. Schéma de calcul reposant sur le principe du chiffrement homomorphe (pour la Science n°456, octobre 2015).

En utilisant le principe du chiffrement homomorphe, une équipe du Massachusetts Institute of Technology (MIT) a développé un cadre technique permettant aux personnes de mettre à disposition leurs données personnelles tout en garantissant leur confidentialité ainsi qu'une utilisation raisonnée par les responsables de traitements. Cette solution, appelée *Enigma*[117], repose sur la technologie de la *blockchain* qui sous-tend le *bitcoin* (monnaie virtuelle mise en circulation en 2009). A l'instar de *bitcoin*, *Enigma* propose de stocker les informations de

[116] https://www.cryptoexperts.com/research/projects/heat/
https://www.cryptoexperts.com/research/projects/cryptocomp/

[117] Cf. http://enigma.media.mit.edu/

façon décentralisée et chiffrée (ici les données personnelles) et de les partager avec des nombreux tiers de manière sécurisée. Ce nouveau type d'architecture permet donc de remettre en cause le modèle actuel d'internet qui s'appuie sur des intermédiaires de confiance centralisant l'intégralité des données et contrôlant leur utilisation[118].

Si le dispositif Enigma est effectivement mis en œuvre[119], les données stockées pourront être analysées par des applications et des logiciels extérieurs, tout en maintenant ces informations sous le contrôle de leur propriétaire. Des retombées de grand intérêt pourraient être observées dans le domaine de la *data science* et du *machine learning*, permettant aux entreprises d'exécuter des calculs sur des données chiffrées et obtenir des résultats utiles sans pour autant pouvoir accéder aux données d'un utilisateur spécifique.

Un Appel à projets « Protection des données personnelles » a été lancé cet automne dans le cadre des Investissements d'avenir. Les *« Architectures innovantes de protection et de gestion des données personnelles favorisant la maîtrise par les individus »* figuraient parmi les trois axes prioritaires de cet appel à projets.

Synthèse/proposition :

- favoriser les possibilités d'exploitation et de réutilisation des données dans un contexte « *big data* » ne doit pas se faire au détriment des droits des personnes, et une approche ouverte de ces possibilités doit être contrebalancée par une transparence réelle à l'égard des personnes concernées. Lorsque l'application « classique » de la loi est possible, celle-ci doit être respectée et cet impératif doit être rappelé aux responsables de traitements concernés.

- Au cas par cas, il pourrait être décidé d'utiliser les dispositions de l'article 32.III de la loi du 6 janvier 1978 modifiée (par exemple en cas d'anonymisation à bref délai des données pour des traitements ayant pour objet de détecter des tendances ou des corrélations).

- Pour le cas particulier de la collecte et l'exploitation des données personnelles publiquement accessibles sur Internet, votre rapporteur encourage la recherche de solutions innovantes et de nouvelles modalités d'application des principes fondamentaux pour les traitements « *big data* » visant à détecter des tendances et des corrélations et, le cas échéant, pour certaines réutilisations des données lorsque les finalités ont été jugées compatibles.
Une piste de réflexion serait que le consentement de la personne à une réutilisation de ses données pour certaines finalités soit donné au moment de la mise en ligne des informations (consentement par type de finalités et avec information, voire consentement, par catégorie de destinataires). Un signe distinctif pourrait permettre d'identifier les choix faits par la personne (révocables à tout moment) et une plate-forme accessible sur le site source permettrait aux personnes de savoir quels organismes ont utilisé leurs données, pour quelle finalité, et d'exercer leurs droits auprès de cet organisme.

[118] En s'appuyant sur le chiffrement homomorphe, *Enigma* crypte les données en les divisant en éléments et en distribuant de manière aléatoire les morceaux indéchiffrables de ces éléments à des centaines d'ordinateurs du réseau *Enigma* appelés « nœuds ». Chaque nœud effectue des calculs sur son morceau discret de l'information avant que l'utilisateur recombine les résultats pour obtenir une réponse en clair. Comme pour la cryptomonnaie *bitcoin*, le registre de la *blockchain* (*ledger*), partagé par les ordinateurs membres du réseau, contrôle l'identité des utilisateurs d'*Enigma* (via un code, car ils sont anonymes) et leur donne accès ou non à tout ou partie des données. Il enregistre l'ensemble des opérations réalisées sur *Enigma* : enregistrement de nouvelles informations, consultation, opérations réalisées sur ces données, *etc.* Cependant, un obstacle considérable pour *Enigma*, est qu'il nécessite que des centaines, voire des milliers d'utilisateurs adoptent le système et exécutent son code avant de pouvoir commencer à garantir une complète sécurité.

[119] Au 05/02/2016, le lancement de la solution *Enigma* n'a pas encore été effectué.

- Pour les traitements ayant un impact sur les personnes concernées, la proposition de votre rapporteur est de renforcer la transparence et les moyens de contrôle des intéressés (obligation de transparence sur les données personnelles utilisées par l'algorithme et le raisonnement général suivi par celui-ci ; assurer l'effectivité de l'intervention humaine dans la prise de décision ; contrôle des algorithmes).

- Les solutions techniques qui se développent pour permettre un traitement des données respectueux de la vie privée doivent être exploitées et encouragées.

IV. CONCLUSION

Si la Commission partage les pistes de réflexion et les analyses développées par votre rapporteur dans la présente communication, il serait possible de les valider comme approche de référence pour les traitements « *big data* », approche qui sera éprouvée, et affinée, dans le cadre de l'instruction des projets « *big data* » à venir.

L'appropriation de cette approche de référence par les différents services de la CNIL permettrait également d'ajouter une dimension plus sectorielle à ces analyses, et de les préciser, certains secteurs d'activités étant soumis à des règles spécifiques (notamment la santé et le régalien).

Dans un second temps, il serait possible d'élaborer des **guides pratiques ou des référentiels sectoriels** pour la mise en œuvre de traitement « *big data* », à l'attention des responsables de traitements. Cette logique serait dans le sillage du développement des packs de conformité.

Ces guides ou référentiels sectoriels pourraient contenir les questions essentielles à se poser lors de la mise en œuvre de ces traitements en fonction de la nature du traitement (cf. typologie et tableau récapitulatif), permettant d'identifier les possibilités et les points de vigilance pour les différents cas de figure, les garanties à apporter, le régime juridique applicable, etc. Ils pourraient présenter des illustrations concrètes (à l'image des exemples développés dans le tableau, en détaillant la démarche). Ils pourraient aussi rappeler les traitements qui sont actuellement soumis à demande d'autorisation, et ceux qui seront demain subordonnés à la réalisation d'une étude d'impact et, le cas échéant, à la consultation préalable de l'autorité de protection des données. A cet égard, il serait possible d'envisager la création d'un PIA « *big data* » pour des secteurs d'activité précis. Ils pourraient également présenter les solutions offertes par l'anonymisation des données et rappeler, dans le contexte du « *big data* », nos recommandations sur le *cloud computing*.

Au préalable, il pourrait être opportun de s'approprier cette grille d'analyse, ces propositions et ces pistes de réflexion afin de valider et d'enrichir leur portée.

En effet, les positions dégagées dans le présent rapport ont certes été élaborées à partir d'exemples concrets de traitements « *big data* » mais elles reposent également en grande partie sur des raisonnements *in abstracto*, issus des dispositions des textes applicables, de rapports et d'analyses théoriques, de positions précédentes dégagées dans des contextes différents, etc. Il semble nécessaire que les pistes de réflexion et propositions soient éprouvées par la pratique et nourries de l'expérience des professionnels en la matière.

LISTE DES ANNEXES