Big data en sciences sociales et protection des données personnelles

Émilie Debaets
Université Toulouse 1 capitole, Institut Maurice Hauriou

Abstract : *Big data brings new possibilities for researches led in social sciences. Aggregating vast quantities of data, and in particular personal data, facilitates the emergence of new knowledges. In consequence, it raises many specific questions on how to reconcile the right to protection of personal data with a freely led scientific research. Can one be prioritized over another? If data protection laws, and especially the new European regulation known as General Data Protection Regulation (GDPR), is applicable to research led in social sciences based on Big data, those laws do not necessarily constitute an excessive constraint for researchers. The rights of data subjects (such as the right to information, the right of access, the right to rectification and the right to erasure) and the obligations and responsibilities for researchers are rendered more flexible to facilitate their researches. The difficulty mainly resides in continuing researchers' acculturation to data protection right.*

La conservation des faits, des comportements, des événements qui forment le quotidien des individus sous forme de données numériques susceptibles d'être croisées et agrégées est devenue banale. Tous les domaines d'activité sont désormais concernés par cette « digitalisation de la vie » (Rouvroy 2010 : 63 et s.). L'analyse de ces masses gigantesques de données, plus ou moins signifiantes, relevait jusque-là des domaines commerciaux. Les données étaient traitées pour établir des statistiques et/ou des informations utiles aux annonceurs publicitaires. Elles sont aujourd'hui utilisées dans des domaines non commerciaux. Les possibilités de croisement et d'agrégation favorisent en effet la production d'une nouvelle sorte de savoir exploitable en épidémiologie, en économie, en sociologie, etc. (Rouvroy 2014 : 407) Quel que soit le terme utilisé pour désigner ce phénomène, ses effets sont ambivalents [1] : s'il constitue un accès démultiplié au savoir, il présente aussi de nombreux risques pour les droits et libertés des individus.

Dans le débat public comme dans le débat académique, le terme *big data* est régulièrement mobilisé, sans pour autant être clairement défini, pour désigner ce phénomène aux potentialités et aux risques multiples. Il n'est cependant pas consacré en droit positif. Employé parfois au singulier, parfois au pluriel, ce terme

1 Sur cette ambivalence générale des nouvelles technologies, voir par exemple Burgorgue-Larsen (2009).

renvoie à des choses très différentes. Au pluriel, il est utilisé, de manière littérale, pour dénommer des données très hétérogènes : les données collectées par les plateformes[2], les données produites par les puces RFID et les capteurs[3], les données issues du mouvement d'*open data*[4], etc. Mais les données très diverses ainsi visées ne sont qu'un aspect d'une transformation plus importante : la multiplication exponentielle des données numériques et la capacité croissante à les utiliser (Ollion et Boelaert 2015). Au singulier, le terme sert à évoquer, de manière contextuelle, cette évolution de la société. C'est d'ailleurs souvent en ce sens qu'il est employé dans les rapports publics[5] et dans la doctrine juridique française[6].

Les enjeux associés au *big data*, présentés traditionnellement par la formule des *5 v* (volume, variété, vélocité, véracité et valeur)[7], posent des problèmes spécifiques selon les domaines concernés. Le domaine des sciences sociales est confronté à une transformation profonde dans la manière d'étudier et de connaître les comportements humains individuels ou collectifs. Le *big data* permet d'appréhender des sujets peu ou pas explorés et de renouveler l'analyse des sujets effectuée jusque-là au moyen d'enquêtes par questionnaires. Des études pourraient ainsi être menées sur la mobilité des populations urbaines à travers l'analyse des données produites par les cartes d'abonnement aux transports en commun, sur la conjugalité à travers l'analyse du comportement des internautes sur les sites de rencontres, sur l'origine et la diffusion de certaines découvertes scientifiques à travers l'analyse de données bibliométriques, etc.[8] Mais le *big data* suscite aussi de nombreuses critiques car il ne s'agit plus, par exemple, d'identifier des relations causales explicatives. Il conduit principalement à établir des corrélations statistiquement signifiantes entre des éléments *a priori* sans rapport[9].

Le traitement statistique de ces données massives semblerait de prime abord échapper à la protection des données personnelles telle qu'elle est prévue par les textes français et européens. Au niveau national, elle résulte de la loi relative à

---

2 Les plateformes désignent tous les sites servant de portail d'accès à des contenus fournis par des sites tiers et recouvrant notamment les moteurs de recherches, les réseaux sociaux, les sites de partage de contenus.

3 Les puces RFID, contenues par exemple dans les cartes d'abonnement ou les cartes bancaires, attribuent à un objet un identifiant unique reconnaissable à distance qui permet de suivre les personnes qui les détiennent. Les capteurs de toute sorte, dans le cadre par exemple de la « voiture connectée » ou de la « maison connectée » produisent en temps réel des informations précises, récurrentes et massives sur d'innombrables pratiques et comportements.

4 Le mouvement d'*open data* est une démarche de communication des documents publics afin qu'ils soient diffusés de manière structurée selon une méthode garantissant leur libre accès et leur réutilisation par tous, sans restriction technique, juridique ou financière injustifiée. Ainsi, par exemple, depuis 2010, le Centre d'accès sécurisé distant (CASD) donne accès aux chercheurs, de façon très encadrée, à des données individuelles (INSEE et Services statistiques ministériels).

5 Voir par exemple, *Le Numérique et les droits fondamentaux* (Conseil d'État 2014).

6 Voir par exemple Cytermann (2015).

7 Conseil d'État (2014 : 48) ou Commissariat général à la stratégie et à la prospective (2013 : 2). Ces enjeux sont parfois exposés par une formule réduite à *3 v* : volume, variété, vélocité voir par exemple CNIL (2012 : 80) ou Commission de réflexion et de propositions sur les droits et libertés à l'âge du numérique (2015 : 108).

8 Pour plus d'explication sur ces exemples, voir Ollion et Boelart (2015 : § 8 et s.).

9 Sur ce renversement de la démarche classique des sciences sociales voir par exemple Cardon 2015.

l'informatique, aux fichiers et aux libertés[10], récemment modifiée d'abord à la marge par la loi pour une République numérique et plus profondément par la loi relative à la protection des données personnelles les recherches en sciences sociales[11]. Au niveau européen, elle découle, dans le cadre de l'Union européenne, des dispositions du règlement européen n° 2016/679[12] – qui s'est substitué à la directive n° 95/46[13] – et dans le cadre du Conseil de l'Europe, des dispositions de la convention n° 108 en cours de révision[14].

La problématique de la conciliation entre deux valeurs aussi importantes consacrées par la jurisprudence constitutionnelle française – la liberté de la recherche[15] et la protection des données personnelles[16] – soulève beaucoup d'interrogations. Les chercheurs en sciences sociales peuvent-ils librement accéder à ces masses gigantesques de données et à quelles conditions ? Comment résoudre l'apparente contradiction entre le besoin de données des chercheurs et la protection des individus dont le comportement est examiné ? Le second doit-il primer sur le premier pour des raisons éthiques ? Le premier peut-il primer sur le second si la recherche est considérée comme relevant de l'intérêt général ?

Si la protection des données personnelles est applicable aux recherches en sciences sociales fondées sur le *big data* (I), celle-ci ne saurait constituer une contrainte, ni même un handicap, limitant la liberté opérationnelle des chercheurs (II). La difficulté principale réside davantage la poursuite du travail d'acculturation et, désormais, de responsabilisation des chercheurs à la protection des données personnelles (III).

L'applicabilité de la protection des données personnelles aux recherches en sciences sociales fondées sur le *big data*

Les recherches en sciences sociales fondées sur le *big data* n'échappent pas à la protection des données personnelles. D'une part, les données auxquelles les

---

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

11 Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique : voir par exemple les dispositions sur l'ouverture des données publiques (art. 1 à 16), le statut des données d'intérêt général (art. 17 à 24), sur les possibilités d'appariement du NIR (art. 34) ; Loi n° 2018-493 du 20 juin 2018 relative à la protection des données personnelles : voir par exemple les dispositions permettant de déroger à l'interdiction d'utilisation des données sensibles dans le cadre de la « recherche publique » (art. 8), les dispositions simplifiant l'utilisation du NIR (art. 11) ou encore les dispositions renvoyant au pouvoir réglementaire les dérogations aux droits des personnes concernées (art. 14).

12 Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE.

13 Directive 95/46/CE du Parlement européen et du Conseil, du 24 octobre 1995, relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données.

14 Convention STE n° 108 pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel, 28 janvier 1981.

15 CC, 20 janvier 1984, déc. n° 83-165 DC.

16 CC, 22 mars 2012, déc. n° 2012-652 DC.

chercheurs souhaitent accéder peuvent souvent être qualifiées de données personnelles tant la définition de celles-ci est englobante (1.1). D'autre part, leur collecte et leur utilisation, quelles qu'elles soient, constituent des traitements de données personnelles (1.2). Dès lors, la protection des données personnelles, instituée par les textes français et européens, s'applique.

La nature des données utilisées

L'une des difficultés posées par l'exploitation des données du *big data* est de déterminer la réglementation applicable tant les données concernées sont diverses. Différents régimes juridiques sont en effet susceptibles d'être appliqués selon la nature des données utilisées.

Premièrement, les recherches en sciences sociales prenant appui sur le *big data* sont libres si les données sont anonymes, c'est-à-dire si elles ne permettent pas d'identifier ou de réidentifier une personne. Cette liberté d'utilisation des données anonymes a d'ailleurs été consacrée par la jurisprudence constitutionnelle française [17]. La complexité provient néanmoins de ce que les nouvelles possibilités de croisement et d'agréation peuvent conduire à réidentifier la personne alors même que les données n'étaient pas identifiantes à l'origine.

Deuxièmement, les recherches en sciences sociales prenant appui sur le *big data* sont encadrées si les données sont personnelles, c'est-à-dire si les données permettent d'identifier directement ou indirectement une personne.

D'une part, les chercheurs ne peuvent pas arguer que les données utilisées ne comportent pas le nom et l'adresse de la personne ou qu'ils n'ont pas l'intention de l'identifier. La définition des données personnelles retenue par le règlement européen n° 2016/679 n'innove pas réellement par rapport à celle de la loi française relative à l'informatique, aux fichiers et aux libertés. Mais, elle tente peut-être de mieux rendre compte du caractère identifiable. Le règlement précise notamment que l'identification peut être faite soit directement au moyen d'un nom ou d'un numéro d'identification, soit indirectement au moyen des données de localisation, d'un identifiant en ligne ou encore par exemple d'un ou plusieurs éléments spécifiques, propres à l'identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale [18]. Les procédés d'identification indirecte, tels que la corrélation et l'inférence [19], sont donc également concernés. Or, ces procédés sont particulièrement en jeu dans le cadre du *big data*. La suppression des données directement identifiantes n'est pas suffisante ; différents travaux ont montré que des données considérées anonymes pouvaient être réidentifiables par ces procédés d'identification indirecte comme ce fut par exemple le cas de la base de données partagée par Netflix [20].

---

17 CC, 23 juillet 1999, déc. n° 99-416 DC, § 50.

18 Art. 4 § 1 du règlement (UE) 2016/679.

19 Sur leur distinction, voir G 29, avis 05/2014, *Les Techniques d'anonymisation*, sp. p. 13. La corrélation consiste dans la capacité à relier entre elles au moins deux informations se rapportant à une même personne. L'inférence est, quant à elle, la possibilité de déduire, avec un degré de probabilité élevé, une information à partir d'un ensemble d'autres informations.

20 Exemple analysé par le G 29, avis 05/2014, *Les Techniques d'anonymisation*, sp. p. 13 et 33 à partir des travaux de Narayanan et Shmatikov (2008).

D'autre part, les chercheurs ne peuvent pas arguer non plus que les données utilisées sont codées ou pseudonymisées. Si, en France, ces données constituaient toujours des données personnelles, tel n'était pas le cas au Royaume-Uni ou en Grèce. Avec le règlement européen n° 2016/679, une telle divergence entre les États membres disparaîtra. Dans la continuité des analyses du G 29[21], le règlement a spécifiquement clarifié le statut de ces données qui ne peuvent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires conservées séparément et soumises à des mesures techniques et organisationnelles[22]. La possibilité de réidentification, porteuse de dangers pour la personne concernée, justifie le maintien du cadre protecteur ainsi institué.

Troisièmement et dernièrement, les recherches en sciences sociales prenant appui sur le *big data* sont très encadrées si les données, qui peuvent être rattachées à une personne, se rapportent à des éléments considérés comme sensibles tels l'état de santé actuel ou futur, les préférences sexuelles, les convictions religieuses, les opinions politiques, etc. Ces éléments considérés comme sensibles sont limitativement énumérés par les textes français et européens[23] et font en principe l'objet d'une interdiction d'utilisation sauf consentement de la personne concernée[24]. La complexité provient cependant là aussi de ce que les nouvelles possibilités de croisement et d'agréation peuvent conduire à révéler des éléments sensibles de la personne alors même que les données n'étaient pas sensibles à l'origine. On peut en effet penser par exemple que de telles informations pourraient être déduites par un algorithme en faisant appel à la géolocalisation et aux métadonnées d'appels (temps passé au téléphone, nombre de personnes appelées, répartition entre appels et envois de SMS, temps mis à répondre à ces derniers, etc.).

La nature des opérations effectuées

Les chercheurs peuvent rencontrer des difficultés techniques et juridiques à collecter certaines données du *big data* notamment parce que celles-ci ne sont pas rendues publiques ou parce que les sites Internet sur lesquels elles sont rendues publiques bloquent leur extraction afin de les revendre aux chercheurs au même titre qu'aux annonceurs publicitaires. Quelles que soient les modalités par lesquelles les chercheurs récupèrent les données – qu'elles leur aient été cédées directement par les sites Internet ou indirectement par le biais d'entreprises d'études du web social[25], qu'elles aient été mises à leur disposition, qu'elles aient été extraites par le biais de techniques de *crawling* ou de *scrapping*, etc. –, toutes ces opérations de collecte et leur exploitation ultérieure s'analysent comme des traitements de données personnelles.

---

21 G 29, avis 4/2007, *Le Concept de données à caractère personnel*, sp. p. 19 à 21 ; avis 05/2014, *Les Techniques d'anonymisation*, sp. p. 11.

22 Art. 4 § 5 du règlement (UE) 2016/679.

23 Art. 9 du règlement 2016/679 (UE) et art. 8 de la loi n° 78-17.

24 Art. 9 du règlement (UE) 2016/679 ; art. 8 de la loi n° 78-17.

25 Par exemple, la société Linkfluence, basée à Paris, en France ou la société Spinn3r, basée à San Francisco, aux États-Unis, qui analysent des centaines de millions de sources et captent des centaines de millions de publications chaque jour.

Les traitements de données personnelles sont définis de manière très large par les textes français et européens. Il s'agit selon le règlement européen n° 2016/679 de toute opération ou tout ensemble d'opérations effectuées ou non à l'aide de procédés automatisés et appliqués à des données ou des ensembles de données personnelles, telles que la collecte, l'enregistrement, l'organisation, la structuration, la conservation, l'adaptation ou la modification, l'extraction, la consultation, l'utilisation, la communication par transmission, la diffusion ou toute autre forme de mise à disposition, le rapprochement ou l'interconnexion, la limitation, l'effacement ou la destruction. Cette définition se retrouve aussi dans la loi française [26].

Les traitements de données personnelles sont également appréhendés de manière très large par la jurisprudence. La Cour de justice de l'Union européenne a par exemple considéré que l'activité des moteurs de recherches, qui repose sur la technique de *crawling*, constitue un traitement de données personnelles [27]. Son raisonnement pourrait ainsi être transposé aux recherches en sciences sociales prenant appui sur le *big data* : tout comme l'exploitant d'un moteur de recherche, le chercheur « collecte » les données personnelles qu'il « extrait », « enregistre » et « organise ». Il les « conserve » et les « exploite ». Le cas échéant, il « communique » ses résultats et les « met à disposition » du public. Selon la Cour, ces opérations qui étaient visées de manière explicite et inconditionnelle par la directive européenne n° 95/46, et désormais par le règlement européen n° 2016/679, constituent des « traitements » de données personnelles. Peu importe que ces opérations visent indistinctement des données personnelles et d'autres types d'informations [28]. Peu importe également que les données personnelles aient déjà fait l'objet d'une publication sur Internet et n'aient pas été modifiées [29]. Dans un cas comme dans l'autre, cela ne remet pas en cause, selon la Cour, la qualification de traitements de données personnelles.

Dès lors, les recherches fondées sur le *big data* constituent des traitements de données personnelles protégés par les textes français et européens.

L'adaptabilité de la protection des données personnelles aux recherches en sciences sociales fondées sur le *big data*

La protection des données personnelles est souvent perçue par les chercheurs comme constituant un frein à leur travail. Or, les textes français et européens prévoient un certain nombre de dérogations [30] pour les recherches scientifiques. Ces dérogations concernent aussi bien l'interdiction de réutilisation des données collectées à

26 Art. 4 § 2 du règlement (UE) 2016/679 ; art. 2 de la loi n° 78-17.

27 CJUE, 13 mai 2014, *Google Spain SL et Google Inc. c. Agencia Española de Protección de Datos (AEPD) et Mario Costeja González*, aff. C-131/12.

28 *Idem*, § 28.

29 *Idem.*, § 29-30 ; voir aussi CJCE, G. Ch., 16 décembre 2008, *Tietosuojavaltuutettu c. Satakunnan Markkinapörssi Oy et Satamedia Oy*, aff. C-73/07, § 48 et 49.

30 Prévues de manière globale par les art. 36 de la loi n° 78-17 et 89 du règlement européen (UE) 2016/679.

d'autres fins [31] que la limitation de la durée de conservation des données, l'obligation d'information des personnes concernées [32], les droits d'accès [33], de rectification [34], et d'opposition des personnes concernées… Elles s'appliquent dès lors que ces droits risqueraient de rendre impossible ou d'entraver sérieusement la réalisation des finalités spécifiques et où de telles dérogations sont nécessaires pour atteindre ces finalités. En dépit du sentiment d'entrave des chercheurs, bon nombre de règles sont donc susceptibles de dérogations lorsque les données personnelles sont traitées à des fins de recherches scientifiques, entendues largement comme toute production de connaissances nouvelles [35]. Dès lors, les recherches en sciences sociales fondées sur le *big data* peuvent bénéficier de cette adaptabilité de la protection des données personnelles qui se manifeste tout particulièrement au niveau de l'information des personnes concernées (2.1) et de l'exercice de leurs droits (2.2).

L'information des personnes concernées

L'information des personnes concernées sur la manière dont leurs données sont collectées, utilisées, consultées, traitées constitue un préalable indispensable à la maîtrise effective des usages qui en sont faits [36]. Ce n'est en effet qu'à travers une information simple, accessible, exhaustive sur les traitements de données personnelles que les personnes concernées peuvent être mises à même d'exercer les droits qui leur sont reconnus. Cette prise de conscience de l'importance de la transparence sur les traitements de données personnelles a conduit à un renforcement progressif du devoir d'information auquel les recherches en sciences sociales fondées sur le *big data* sont également soumises.

Le devoir d'information ne se limite pas aux cas où les données personnelles sont directement collectées auprès de la personne concernée. Ainsi que le prévoyaient déjà la directive européenne n° 95/46 et la loi française, celui-ci s'applique aussi, sous des modalités différentes, aux cas où les données personnelles sont collectées auprès d'un tiers [37]. Ce devoir d'information lorsque les données personnelles ne sont pas collectées auprès de la personne concernée a été renforcé par le règlement européen n° 2016/679 [38]. Les éléments d'information à fournir sont désormais plus nombreux. La personne concernée doit par exemple être informée de la période de conservation des données (à tout le moins, des éléments permettant de la déterminer), des intérêts légitimes du responsable du traitement (lorsque la licéité de celui-ci est basée sur un équilibre des intérêts), de l'ensemble des droits reconnus

31 Art. 6 de la loi n° 78-17.

32 Art. 32, III, de la loi n° 78-17.

33 Art. 39, II, de la loi n° 78-17.

34 Art. 40, II, de la loi n° 78-17.

35 Peu importe que les recherches menées soient fondamentales ou appliquées. Peu importe également le domaine dans lequel les recherches sont menées. Peu importe enfin la structure publique ou privée au sein de laquelle les recherches sont effectuées. Voir cons. 159 du règlement européen (UE) 2016/679.

36 Voir la consécration de cet objectif au sein de l'art. 1er al. 2 de la loi n° 78-17 (tel que modifié par la loi pour une République numérique).

37 Art. 11 de la directive 95/46/CE ; art. 32, III de la loi n° 78-17.

38 Art. 14 du règlement européen (UE) 2016/679.

à la personne (dont le droit à la portabilité des données ou le droit au retrait du consentement), du droit d'introduire une réclamation auprès d'une autorité de contrôle, de l'existence d'une décision automatisée comprenant un profilage (avec une explication significative de la logique sous-jacente et les conséquences pour les personnes). Le règlement précise également que le responsable doit fournir ces informations à la personne concernée soit dans un délai raisonnable qui n'excède pas un mois après la collecte, soit, s'il est envisagé de communiquer les informations à un autre destinataire ou de les utiliser pour communiquer avec la personne concernée, au plus tard lorsque les informations sont communiquées pour la première fois.

Ce devoir d'information fait cependant l'objet d'exceptions, exceptions élargies par le règlement européen. L'une d'entre elles vise spécifiquement les traitements à des fins de recherche scientifique. L'information ne doit pas être fournie si elle s'avère impossible ou nécessite des efforts disproportionnés. Elle ne doit pas être fournie également si elle est susceptible de rendre impossible ou de compromettre gravement la réalisation des objectifs du traitement. Or, eu égard au « nombre de personnes concernées », les recherches en sciences sociales prenant appui sur le *big data* sont directement visées[39]. En pareils cas, il appartient alors au responsable du traitement de prendre « des mesures appropriées pour protéger les droits et libertés ainsi que les intérêts légitimes de la personne concernée, y compris en rendant les informations publiquement disponibles ». Le devoir d'information semble ainsi pouvoir être restreint à une information générale et collective sur l'objet et le protocole de l'étude envisagée ainsi que sur les modalités d'exercice des droits. Cette information pourrait alors emprunter différents canaux : campagne d'affichage dans les établissements fréquentés par les personnes concernées (établissements scolaires, entreprises, maisons départementales des personnes handicapées, etc.) et sur leur site Internet et Intranet, articles dans la presse générale et/ou spécialisée, etc. En toute hypothèse, la détermination de ces mesures appropriées ne peut être faite qu'au cas par cas.

L'exercice des droits des personnes concernées

Les droits reconnus aux personnes concernées tendent à rétablir un équilibre entre celles-ci et les responsables de traitement. Le droit d'accès, le droit de rectification, le droit à l'effacement, le droit à la limitation du traitement ou encore par exemple le droit d'opposition contribuent à assurer aux personnes concernées une maîtrise sur les usages qui sont faits de leurs données personnelles. Ces droits s'imposent aux recherches en sciences sociales prenant appui sur le *big data*.

Des dérogations aux droits reconnus aux personnes concernées étaient déjà prévues par la directive n° 95/46 et par la loi française pour les traitements de données personnelles à des fins de recherche scientifique. Le règlement européen n° 2016/679 clarifie en partie le régime spécifique aux recherches scientifiques en regroupant, dans son article 89, la plupart de ces dérogations[40]. Il ressort de cet

39 C'est d'ailleurs l'un des critères qui devrait être pris en considération dans cette appréciation : v. cons. 62 du règlement européen (UE) 2016/679.

40 Certaines dérogations ne sont pas mentionnées dans cet article et demeurent, malgré la tentative de clarification, énoncées dans les articles consacrés aux droits en question. Voir par exemple l'article 17 sur le droit à l'effacement et à l'oubli.

article et de l'article 17 que les États peuvent apporter des dérogations au droit d'accès, au droit de rectification, au droit à l'effacement, au droit à la limitation du traitement et au droit d'opposition lorsque les données personnelles sont traitées à des fins de recherche scientifique[41]. Mais ces dérogations sont alors soumises à une double condition. D'une part, l'application de ces droits doit rendre impossible ou compromettre sérieusement l'accomplissement des finalités poursuivies. D'autre part, les dérogations doivent être nécessaires à l'accomplissement de ces finalités.

L'appréciation de cette double condition peut s'avérer difficile pour les recherches en sciences sociales prenant appui sur le *big data*. Le droit à l'effacement, qui participe à ce que l'on qualifie communément de « droit à l'oubli », l'illustre bien. Ce droit, nouvellement consacré par le règlement européen, a été contesté dans son principe même en ce qu'il constituerait un frein à la recherche. Alors que l'effacement était conditionné jusque-là à l'inexactitude ou l'équivocité des données personnelles[42], la personne concernée peut désormais demander l'effacement de ses données lorsqu'elle a retiré son consentement ou lorsqu'elle exerce son droit d'opposition[43]. C'est pourquoi le règlement européen a expressément prévu la possibilité d'y déroger pour les traitements de données personnelles à des fins de recherche scientifique. Néanmoins, cela suppose que le non-effacement compromette sérieusement l'accomplissement des finalités et qu'il soit nécessaire à l'accomplissement de ces finalités. Or, eu égard au nombre de personnes concernées par les recherches en sciences sociales prenant appui sur le *big data*, cette dérogation semble directement compromise.

Qu'il s'agisse de l'information des personnes concernées ou de l'exercice de leurs droits, le règlement met en œuvre un principe de proportionnalité. C'est donc au terme d'une analyse concrète pour chacune des recherches en sciences sociales fondées sur le *big data* que des dérogations pourront être envisagées.

La responsabilité des chercheurs en sciences sociales travaillant sur le *big data* en matière de protection des données personnelles

Les recherches en sciences sociales fondées sur le *big data* doivent dorénavant systématiquement intégrer dans leurs protocoles la question de la protection des données personnelles et développer une culture de conformité au règlement européen n° 2016/679 (3.1). Ces recherches mériteraient donc, peut-être, d'être sécurisées à travers le développement de comités éthiques tels que ceux spécifiquement prévus en matière de recherches médicales (3.2).

---

41 Les droits auxquels des dérogations peuvent être apportées varient selon la finalité poursuivie (des fins de recherches scientifiques ou des fins archivistiques). L'article 89 prévoit aussi des dérogations aux obligations de notification (art. 19) et au droit à la portabilité des données (art. 20) en cas de traitements à des fins archivistiques.

42 Art. 40 de la loi n° 78-17 ; art. 12 de la directive 95/46/CE.

43 Art. 17 du règlement (UE) 2016/679.

La mise en œuvre de la notion d'*accountability*

Le règlement européen n° 2016/679 procède à un changement de paradigme. Au contrôle *a priori* fondé sur des formalités préalables est substitué un contrôle *a posteriori* fondé sur la notion d'*accountability*. En vertu de ce principe, tous les responsables de traitements, y compris les chercheurs, doivent mettre en œuvre des mesures techniques et organisationnelles appropriées en vue d'effectuer ces traitements dans le respect du règlement et être à même de le démontrer.

L'article 89 du règlement relatif aux traitements de données personnelles à des fins de recherche scientifique traduit cette notion d'*accountability*. Il soumet ces traitements à la prise de mesures de sauvegarde des droits et libertés des personnes en termes notamment de minimisation des données et de pseudonymisation. La minimisation des données[44] implique que seules les données nécessaires à la poursuite des finalités puissent faire l'objet d'un traitement. Or, les recherches fondées sur le *big data* pourraient apparaître incompatibles avec ce principe car « toute donnée triviale, y compris ce qui passerait, dans le contexte de traitements statistiques plus classiques, pour du "bruit", peut concourir à la production de profil » (Rouvroy 2014 : 412). La pseudonymisation implique, quant à elle, que l'identité de la personne concernée soit dissimulée afin d'atténuer le risque de mise en corrélation d'un ensemble de données[45]. L'article 89 précise d'ailleurs que, si la poursuite des finalités le permet, le responsable doit privilégier les traitements qui ne permettent pas ou plus l'identification des personnes concernées. Or, là aussi, les recherches fondées sur le *big data* pourraient apparaître incompatibles avec ce principe. Les chercheurs sont ainsi contraints de réfléchir à leur méthodologie pour s'assurer qu'ils ne collectent pas des données non nécessaires et que celles-ci ont été, si possible, pseudonymisées.

D'autres articles du règlement, qui traduisent aussi cette notion d'*accounatbility*, sont également applicables aux traitements de données personnelles à des fins de recherches scientifiques. C'est par exemple le cas de l'article 24 qui impose de prendre en compte les droits et les intérêts des personnes concernées dès la conception du traitement et dans les paramétrages par défaut. Or, ces techniques de protection de la vie privée qui se traduisent par des outils de chiffrement ou de calcul sécurisé sont spécialement requises dans le cadre du développement du *big data*. C'est aussi par exemple le cas de l'article 35 qui impose de réaliser une analyse d'impact pour les traitements susceptibles d'exposer les personnes à un risque élevé au regard de leurs droits et libertés. Le responsable doit alors évaluer, en particulier, l'origine, la nature, la portée, le contexte, la particularité et la gravité de ce risque. Or, cette analyse d'impact est spécialement requise pour les traitements qui servent à traiter un volume considérable de données personnelles et qui peuvent affecter un nombre important de personnes[46].

---

44 Art. 5 c du règlement (UE) 2016/679.

45 Art. 4 § 5 du règlement (UE) 2016/679.

46 Cons. 91 du règlement (UE) 2016-679 et l'un des critères développés par le G 29 pour déterminer si le traitement présente un risqué élevé. Voir G 29, Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, 4 avril 2017, sp. p. 8 et 9.

Si la notion d'*accountability* ne s'oppose pas aux recherches en sciences sociales prenant appui sur le *big data*, elle fait néanmoins peser sur les chercheurs et les équipes de recherche une responsabilité nouvelle.

Pour un développement des comités d'éthique pour les recherches en sciences sociales ?

Les pouvoirs publics n'ont pas mené de réflexion sur la généralisation des comités d'éthique pour la recherche scientifique à l'instar de ce qui existe déjà pour la recherche en matière médicale. Ne faudrait-il pas instaurer dans le domaine des sciences sociales des comités d'éthiques agissant aux côtés de la CNIL pour conseiller les chercheurs ? Ces comités d'éthiques permettraient-ils de consolider juridiquement les recherches menées et de les légitimer ?

Si les données personnelles de santé, en tant que données sensibles, ne peuvent en principe pas être traitées, le législateur français a spécifiquement autorisé leur utilisation à des fins de recherche médicale. Ce traitement repose sur trois principes essentiels[47]. D'une part, le traitement de ces données personnelles est possible dès lors que la personne dûment informée ne s'y est pas opposée. D'autre part, le traitement n'est possible qu'après autorisation de la CNIL qui prend sa décision après une évaluation par des comités spécifiques (le comité de protection des personnes ou le comité d'expertise pour les recherches, les études et les évaluations dans le domaine de la santé). Enfin, la transmission des données personnelles n'est possible que si elles ne permettent pas l'identification de la personne ou s'il existe des garanties assurant la confidentialité de ces données.

Seule la recherche médicale fait donc l'objet d'une évaluation par des comités d'éthiques. Ceux-ci ne devraient-ils pas être généralisés à l'ensemble de la recherche scientifique dès lors qu'elle porte sur la personne humaine comme c'est le cas de la sociologie ou de la psychologie par exemple ? Cependant, les chercheurs semblent majoritairement hostiles à l'adaptation de tels principes aux sciences sociales en ce qu'ils limiteraient directement leur liberté de recherche, voire même en ce qu'ils permettraient une forme de censure au nom de l'éthique[48]. Le modèle des « Institutional Review Boards », des comités d'éthique mis en place dans les universités américaines, modèle désormais exporté en Grande-Bretagne, en Australie ou en Afrique du Sud, est en effet souvent perçu comme l'« importation d'une analogie ruineuse avec l'univers médical » (Laurens et Neyrat 2010 : 9 et s., sp. p. 30) conduisant à « la montée inéluctable et paralysante des régulations bureaucratiques » (Fassin 2008 : 124 et s., sp. 127).

Ces réticences ne doivent pas pour autant exclure toute réflexion sur la pertinence d'une régulation éthique en complément de la régulation juridique et susceptible d'être prise en compte par la CNIL dans son appréciation[49]. Les recherches fondées sur

---

47 Chap. IX de la loi n° 78-17.

48 Sur ce constat s'agissant de l'exemple américain, voir Bonnet et Robert (2009 : 87 et s., sp. 101).

49 Sur la prise en compte au titre des garanties de la création d'un comité éthique, composé de représentants d'associations, de statisticiens et de personnes qualifiées, voir par exemple la délibération n° 2017-126 du 20 avril 2017 autorisant le centre universitaire de recherche sur l'action publique et le politique, épistémologie et sciences sociales (CURAPP-ESS) à mettre en œuvre un

les données issues du *big data* présentent, quel que soit le domaine des recherches, des risques similaires auxquels peuvent s'ajouter des risques spécifiques selon les données utilisées. Dès lors, de tels comités, dont les modalités seraient à discuter au vu des expériences étrangères (comités centralisés ou décentralisés, critères de l'évaluation, évaluation *a priori* ou *a posteriori*) permettraient notamment d'établir une déontologie commune sur les conditions dans lesquelles les données issues du *big data* peuvent être utilisées à des fins de recherche scientifique. Ces comités d'éthiques auraient ainsi pour mission d'associer les participants à la recherche et les chercheurs, d'assurer la qualité des travaux par l'analyse des protocoles et des méthodologies de recherche, de protéger les participants, de veiller au respect de règles de gouvernance concernant l'information des personnes et les politiques de publication des résultats dans les revues scientifiques.

traitement de données à caractère personnel ayant pour finalité la réalisation d'une étude sur la citoyenneté politique des personnes handicapées.