Données massives, données ouvertes et protection de la confidentialité

Jean-Pierre Le Gléau

Inspecteur général honoraire de l'Insee, ancien chef du département de la coordination statistique.
Membre du comité d'éthique de l'Ined
Membre du comité d'experts de l'Institut des données de santé

jean-pierre.le-gleau@orange.fr

Résumé

L'irruption des techniques de données massives et de données ouvertes ouvre un nouveau contexte pour la protection de la confidentialité. Auparavant, les fichiers de données étaient structurés, confinés dans des espaces dédiés et isolés les uns des autres. La protection de la confidentialité est très difficile à obtenir pour les données relatives aux entreprises. Pour les personnes physiques, la vie privée est protégée par des lois françaises et par une directive européenne. Ces textes définissent ce qu'est une donnée à caractère personnel. Ils fixent les règles que doivent respecter le traitement de telles données. Notamment : la finalité déclarée, le droit à l'oubli, le droit d'opposition et de rectification. Ces exigences sont peu compatibles avec la diffusion sous forme de données ouvertes. Or, les ensembles de données massives comportant des données individuelles conduisent presque immanquablement à la constitution de données à caractère personnel, c'est-à-dire où il est possible d'identifier certains individus. Ces pourquoi ces ensembles ne peuvent être diffusés de façon ouverte, mais doivent faire l'objet d'un encadrement spécifique pour leur diffusion, en la limitant à un petit nombre de personnes. Les dispositifs juridiques et techniques existent pour respecter ces conditions.

Mots-clefs

données massives, données ouvertes, données à caractère personnel, confidentialité

Abstract

The emergence of big data and open data technologies brings a new context for privacy. Previously, the data files were structured, confined in dedicated areas and isolated from each other. The protection of confidentiality is very difficult to obtain for business data. For individuals, privacy is protected by French laws and by an European directive. They define what are personal data . They set the rules the processing of such data must comply with. Including: specified purposes, the right to be forgotten, the right of opposition and rectification. These requirements are not compatible with the release as open data. However, massive data sets with individual data almost inevitably lead to the production of personal data, that is to say where it is possible to identify individuals. That is why these sets may not be distributed openly, but should be disseminated trough a specific framework, by limiting it to a small number of people. Legal and technical means exist to meet these conditions.

Keywords

big data, open data, personal data, privacy

Les conditions de la protection de la confidentialité ont connu un important bouleversement avec l'irruption des données massives et leur mise à disposition d'un public très large sous forme de données ouvertes.

Une nouvelle donne

Jusqu'à présent, les données recueillies ou disponibles portant sur les acteurs de la société (individus ou entreprises) étaient réunies de façon **structurée**, classées à l'aide de nomenclatures, reliées de façon ordonnée les unes aux autres. Les données numériques, les données textuelles, les images, formaient des ensembles distincts, ayant chacun un mode de gestion propre de la confidentialité. La protection de cette dernière, dans ce cadre, relevait d'une analyse logique rigoureuse, relevant souvent de la mathématique traditionnelle.

Ces informations, lorsqu'elles portaient sur des unités identifiables avaient vocation à **rester confinées** dans un espace clos, de façon en général encadrée par la loi. Qu'il s'agisse de données collectées par l'administration ou par des organismes privés, elles restaient le plus souvent accessibles uniquement à celui qui les avait recueillies. Leur mise à disposition de tiers ne pouvait se faire que de façon très réglementée, comme c'est le cas par exemple pour les résultats individuels des enquêtes statistiques, qui ne sont accessibles que selon un protocole permettant le plus souvent d'identifier la personne qui accède à ces données et de tracer les requêtes qu'elle effectue sur les fichiers [1].

L'ensemble des données collectées par différents organismes, publics ou privés, se trouvaient ainsi **isolées les unes des autres**, sans moyen de les raccorder ni de les apparier de façon déterministe, pas même de façon probabiliste. Cette isolation des informations recueillies par divers acteurs était un élément fort de la protection de leur confidentialité, en interdisant l'enrichissement d'un fichier par un autre. L'appariement de fichiers correspondant à des intérêts publics différents est d'ailleurs strictement encadré par la loi.

Dans le nouveau contexte, caractérisé par une diffusion beaucoup plus large des données et leur rassemblement dans des ensembles de grande taille, la protection de la confidentialité se pose de façon nouvelle, puisque les données sont maintenant accumulées de façon non structurée, qu'elles sont accessibles à un très vaste public et qu'elles peuvent être reliées les unes aux autres.

Les données sur les entreprises

La protection de la confidentialité ne concerne pas seulement la protection de la vie privée, mais elle touche aussi d'autres aspects come par exemple le secret des entreprises, qu'il soit commercial ou industriel.

Pour ce dernier type de données, les statisticiens ont depuis longtemps tranché : il n'est pas possible de diffuser des données à la fois individuelles et anonymes sur les entreprises. Dès que l'on connaît, pour une entreprise, son secteur d'activité, une indication de sa taille, voire une localisation, même grossière, il est pratiquement impossible de conserver son anonymat et l'entreprise en question sera le plus souvent identifiable. Mais quel serait par ailleurs l'intérêt de diffuser des informations sur une entreprise en occultant son secteur d'activité, sa taille et sa localisation géographique ? C'est pourquoi, les statisticiens ont décidé que les résultats individuels sur les entreprises ne pouvaient jamais être considérés comme anonymes et leur diffusion doit toujours passer par un avis du comité du secret statistique (celui-ci avait d'ailleurs été créé initialement dans ce but, et son premier nom était « Comite du secret statistique concernant les entreprises » [2]). On peut à la rigueur admettre que l'anonymat reste respecté, si le fichier de données individuelles ne comporte que de très petites entreprises et que la localisation et l'activité économique sont codées de façon grossière. Mais alors, on est pratiquement en présence d'un fichier relatif à des personnes. La loi réserve d'ailleurs un traitement particulier aux entreprises individuelles, les assimilant parfois à des personnes physiques.

L'encadrement légal de la protection de la vie privée

Mais, naturellement, la question la plus aiguë reste celle de la protection de la vie privée. En France, celle-ci est régie par différents textes de loi, dont deux paraissent particulièrement pertinents pour le cas des données massives. Il s'agit de la loi du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistique [3] et le la loi du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés [4].

La première fixe un cadre très strict pour la communication de renseignements individuels recueillis au moyen d'enquêtes statistiques. Le but est d'obtenir des informations sincères lors de ces collectes, en garantissant aux répondants que leurs réponses ne pourront en aucun cas leur porter tort, ni vis-à-vis de leur entourage, ni vis-à-vis de la puissance publique (fisc, police, etc.). Un cas particulier est fait pour les renseignements « ayant trait à la vie personnelle et familiale et, d'une manière générale, aux faits et comportements d'ordre privé ». Ceux-ci ne peuvent, pendant une période de soixante-quinze ans, faire l'objet d'aucune communication, sauf à des fins de statistique publique ou de recherche scientifique ou historique [4]. Et même dans ces cas, la communication est extrêmement encadrée, pour éviter tout risque de dissémination, par imprudence, laxisme ou volonté de nuire. On est évidemment très loin des données ouvertes.

La seconde loi concerne a priori toutes les données à caractère personnel. Le législateur a produit une définition précise de ce terme. Une donnée à caractère personnel est une « *information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres. Pour*

déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens en vue de permettre son identification dont dispose ou auxquels peut avoir accès le responsable du traitement ou toute autre personne ». Cette définition résulte de la transposition d'une directive européenne du 24 octobre 1995 [5]. Il est à noter que la loi française a retenu une définition plus large que celle qui figurait dans la directive européenne. En effet, cette dernière stipule dans son 26ème considérant que « *pour déterminer si une personne est identifiable, il convient de considérer l'ensemble des moyens susceptibles d'être* ***raisonnablement*** *mis en œuvre, soit par le responsable du traitement, soit par une autre personne, pour identifier ladite personne* » La suppression de l'adverbe « *raisonnablement* » dans la loi française, obtenue après un débat de moins de cinq minutes, tant en commission qu'en séance plénière de l'Assemblée nationale, élargit considérablement le champ de ce qu'il convient de comprendre derrière l'expression « données à caractère personnel » Une information qui, pour permettre d'identifier une personne, devrait mettre en jeu des moyens immenses, en allant au-delà de tout ce qui est raisonnable, sera considérée comme une donnée à caractère personnel au sens de la loi française, alors qu'elle ne l'aurait pas été en application stricte de la directive européenne. La France manifeste ainsi son originalité par rapport à la plupart de ses voisins européens, en mettant des conditions très strictes sur la protection de la vie privée (la loi Informatique et libertés a été l'une des toutes première au monde à traiter de ce sujet), alors qu'elle est beaucoup plus ouverte pour ce qui concerne l'accès aux données d'entreprises, comme l'a montré la création du Comité du secret statistique concernant les entreprises, dès 1984. La directive de 1995 devrait être prochainement (?) remplacée par un règlement, qui a déjà été adopté par le Parlement européen et qui doit encore être examiné par le Conseil. Ce règlement s'imposera, sans besoin de transposition dans le droit national, à tous les pays de l'Union européenne. Il maintient la définition des « données à caractère personnel » conforme à celle de la directive. La position française pourrait, de ce fait, évoluer, sous l'effet de la réglementation européenne.

Quoi qu'il en soit de la définition exacte des « données à caractère personnel », leur mise à disposition est régie par un certain nombre de règles qui, pour le coup, sont dans l'ensemble communes à la France et à l'Union européenne. Ces règles reposent sur les principes suivants :

- les données ont été collectées et traitées de manière loyale et licite ;
- elles sont collectées pour des finalités déterminées, explicites et légitimes et ne sont pas traitées ultérieurement de manière incompatible avec ces finalités ;
- elles sont adéquates, pertinentes et non excessives au regard des finalités pour lesquelles elles sont collectées et de leurs traitements ultérieurs ;
- elles sont exactes, complètes et, si nécessaire, mises à jour ;
- elles ne sont conservées que pendant une durée qui n'excède pas la durée nécessaire aux finalités pour lesquelles elles ont été collectées et traitées.

Par ailleurs, toute personne physique dispose, à l'égard des données à caractère personnel la concernant, de droits :

- d'opposition, c'est-à-dire de refuser que ces données fassent l'objet d'un traitement (notamment leur mise à disposition dans un ensemble de données ouvertes) ;
- d'accès, pour connaître la totalité des données à caractère personnel la concernant et figurant dans un ensemble de données ;
- de rectification, pour le cas où ces données ne correspondraient pas à la réalité.

Les données ouvertes sont-elles compatibles avec les données à caractère personnel ?

Évidemment, ces exigences sont peu compatibles avec la présence de telles données dans un ensemble de fichiers mis à disposition dans le cadre de données ouvertes et accessibles à un public très large. En particulier, les notions de finalité du traitement, de « droit à l'oubli », d'opposition semblent par essence contradictoires avec ce mode de diffusion.

C'est pourquoi, on doit se poser la question de la présence ou non de données à caractère personnel au sein des ensembles mis à disposition sous forme de données ouvertes. Pour que des données individuelles ne soient pas classées dans cette catégorie, un important effort doit être fait pour les anonymiser. Ce terme ne doit bien évidemment pas être pris dans son sens premier, qui est celui de « enlever le nom ». C'est bien sûr une étape nécessaire, mais elle est loin d'être suffisante. D'abord, parce qu'il existe des identifiants directs autres que le nom. Le numéro dit de « Sécurité sociale » en est un par exemple. Mais, même si l'on a ôté le nom, ainsi que tous les identifiants directs, il est

souvent possible d'identifier une personne par croisement de données, en elles-mêmes anodines, la concernant. Par exemple son âge, son lieu de résidence et le fait qu'elle ait séjourné dans un certain hôpital entre deux dates données. Chacun de ces éléments ne suffit pas pour identifier une personne, mais leur combinaison aboutit très souvent à un individu unique, que l'on peut donc identifier, et pour lequel on disposera donc de toute l'information du fichier associée à cette personne ; par exemple, la maladie pour laquelle cette personne a été hospitalisée. Ces informations, bien que dépourvues de nom et de tout identifiant direct, constituent donc des données à caractère personnel au sens de la loi de 1978 et, s'il n'a pas été nécessaire de mettre en œuvre des moyens « déraisonnables » (ce qui est le cas dans l'exemple cité), au sens de la directive ou du futur règlement européen.

La mise à disposition d'informations individuelles sous forme de données ouvertes doit donc s'accompagner de nombreuses précautions afin d'éviter toute possibilité d'identification, même indirecte. Cela entraîne nécessairement un appauvrissement drastique des bases les contenant. Dès lors que ces bases sont massives, la tâche devient pratiquement impossible, tant les possibilités de croisement entre les variables deviennent importantes et accentuent donc la probabilité de conduire à un individu unique. À partir d'un certain niveau d'accumulation des données, la tâche deviendra donc impossible.

Comment accéder aux données massives ?

Faut-il pour autant renoncer à la construction de telles bases, dont l'utilité scientifique ou économique est largement avérée ? Sans doute pas. Mais il faut dès lors prendre des mesures permettant de respecter les grands principes repris par les lois et directives citées précédemment. Par exemple, en restreignant leur accès à des personnes ou groupes de personnes qui n'utiliseront ces informations que pour des finalités compatibles avec celles qui ont été déclarées au moment de leur collecte. Et en s'assurant que les droits des personnes concernées sont respectés tout au long du traitement. Pour cela, il est nécessaire de procéder à une identification des personnes ayant accès aux données, de mettre en œuvre des moyens pour empêcher leur dissémination vers d'autres utilisateurs, de conserver une trace des traitements qu'elles auront effectués sur ces données. Des dispositifs techniques [6] et un encadrement juridique adapté existent aujourd'hui pour assurer l'ensemble de ces fonctions.

Il conviendra de les mettre en œuvre pour la mise à disposition de données massives, car celles-ci ne sont le plus souvent pas compatibles avec une diffusion sous forme de données ouvertes.

Références bibliographiques

[1] Le Guide du secret statistique :
http://www.insee.fr/fr/insee-statistique-publique/statistique-publique/guide-secret-18-10-2010.pdf

[2] Le Comité du secret statistique :
http://www.insee.fr/fr/ffc/docs_ffc/cs128e.pdf

[3] Loi sur l'obligation, la coordination et le secret en matière de statistique :
http://legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000888573&fastPos=4&fastReqId=1953093530&categorieLien=cid&oldAction=rechTexte

[4] Loi relative à l'informatique, aux fichiers et aux libertés :
http://legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000886460&fastPos=1&fastReqId=845707906&categorieLien=cid&oldAction=rechTexte

[5] Directive européenne 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
http://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:31995L0046&rid=2

[6] Le Centre d'accès sécurisé aux données
http://www.insee.fr/fr/ffc/docs_ffc/cs130e.pdf