TELECOM
REVUE DE L'ASSOCIATION TELECOM PARISTECH ALUMNI

TELECOM ParisTech
ALUMNI

N°169 - JUILLET 2013

BIG DATA
NOUVEAUX DÉFIS

Editorial

Fleur Pellerin, Ministre Déléguée chargée des PME,
de l'Innovation et de l'Économie numérique

Cher lecteur,

Je remercie Télécom ParisTech de me donner l'occasion, à travers cet éditorial, de partager avec vous mes convictions sur ce sujet stratégique des Big data.

Derrière le « buzz », je suis convaincue que nous avons devant nous une nouvelle révolution du numérique, avec des implications sociétales et économiques.

Nous vivons au quotidien un « déluge de données », que l'explosion des terminaux mobiles et l'émergence attendue de l'Internet des Objets vont amplifier. Mais les enjeux sont devant nous : comment créer de la valeur (une valeur d'usage, une valeur économique) à partir du traitement de ces grandes masses de données, tout en respectant des contraintes de confidentialité ?

Des standards de fait ont déjà émergé et permettent de « passer à l'échelle », je pense à Hadoop. Sur cette base, nos chercheurs, nos entrepreneurs, nos startups ont une vraie carte à jouer pour faire de la France la référence pour cette nouvelle économie.

Pourquoi suis-je si optimiste ?

Plus que dans d'autres domaines de l'économie numérique, la barrière technologique reste un élément important pour les Big data, par exemple avec la nécessité de disposer d'algorithmes performants de *crawling* ou de *text mining* (en bon français). Avec nos compétences en mathématiques, en statistiques ou en informatique, nous avons tous les atouts pour faire de cette barrière un avantage concurrentiel.

Un autre élément majeur est la capacité à disposer de données, le carburant de cette nouvelle économie. Parmi les sources de données, il y a bien sûr les données ouvertes (« open data »), largement ancrées dans nos territoires, donc à activités induites moins délocalisables. Avec le travail d'Etalab, sous la direction renouvelée d'Henri Verdier, nous avons un autre atout dans notre manche.

Enfin, les secteurs économiques et industriels « classiques » sont en train de réaliser que leurs modèles vont changer avec les Big data. C'est toute ma responsabilité que de veiller à ce qu'aucun domaine ne rate le train en marche.

La politique que je mène pour soutenir l'économie numérique est à l'écoute des acteurs que vous êtes. Du nouvel appel à projet du FSN dédié aux Big data (à partir de septembre) aux quartiers numériques que nous sommes en train de mettre en place pour dynamiser les écosystèmes numériques de croissance, en passant par ma politique en faveur de l'innovation, nous construisons les éléments dont vous avez besoin pour réaliser vos ambitions. Et les Big data sont un axe central de cette politique.

Ce numéro montre tout le dynamisme de Télécom ParisTech, plus largement de la communauté française du numérique, je pense par exemple à des entreprises comme Talend ou Exalead, à des entrepreneurs comme Tariq Krim, et à bien d'autres encore !

Bonne lecture !

BIG DATA : NOUVEAUX DÉFIS

03. **Editorial de Fleur Pellerin,**
Ministre Déléguée chargée des PME, de l'Innovation et de l'Économie numérique

SITUATION ET PERSPECTIVES

06. **Big data : une nouvelle science des données.** Eric MOULINES (86)
10. **Le Concept « Big data » - Nouveaux enjeux technologiques.** Jean-François MARCOTORCHINO
18. **Big data : un nouvel Eldorado pour les telcos ?** Élie GIRARD
22. **De Big data à Big Brother**
Entretien avec Gérard MEMMI (76), responsable du département INFRES de Télécom ParisTech
24. **Simplexité des Big data : la simplicité au service de la complexité.** François STEPHAN (91)

COLLECTER ET TRAÎTER

26. **La Data va-t-elle dévorer le monde ?** Ivan DE LASTOURS
28. **Big data et protection des données personnelles : un défi (quasi) impossible ?**
Claire LEVALLOIS-BARTH
31. **Soif de données ? Les sourceurs arrivent.** François BANCILHON
34. **Une Sécurisation des Big data est-elle envisageable ?** Tarik MOATAZ

USAGES ET APPLICATIONS

38. **Big data et marketing : une application au marketing politique**
Amirhossein MALEKZADEH (97)
42. **Champs pétroliers digitaux : « Big Production Data »**
Antoine TRIHOREAU et Aymeric PRÉVERAL-ETCHEVERRY
44. **Big data au service de la sécurité du transport aérien : l'analyse des données de vol**
Pierre JOUNIAUX
48. **La pub en ligne boostée par le Big data.** Vincent LEPAGE (2006) et Nicolas GRISLAIN
50. **Du Big data à l'Open Data, le nouvel enjeu du partage d'information par la géographie**
Jean-Thomas ROUZIN
52. **Hub de données pour services intelligents.** David THOUMAS

RETOMBÉES ÉCONOMIQUES ET ACADÉMIQUES

54. **Télécom ParisTech lance la première formation en France dédiée au « Big data »**
Stéphan CLÉMENÇON
58. **Enjeux économiques du Big data.** Patrick WAELBROECK

Notre réseau

63. **Présentation du Livre de Jacques FLEURET (67)**
62. **Retour sur les remises de diplômes Mastères spécialisés, Docteurs et Ingénieurs**
68. **Retour sur La 15ème cérémonie du Prix des Technologies Numériques**
77. **Les Rencontres Télécom ParisTech alumni d'Orange**
80. **La Fondation Télécom**
82. **Les actualités de Télécom ParisTech**

LE PROCHAIN NUMÉRO SERA CONSACRÉ AUX TRANSPORTS INTELLIGENTS.

TÉLÉCOM n°169 - Juillet 2013
est édité par l'Association Télécom ParisTech alumni.
Dépôt légal à parution.

Directeur de la publication : Dominique Jean (73)
Secrétaire de rédaction : Amélie Pageard
Numéro préparé par : Gérard Cambillau (73) et Louis-Aimé de Fouquières (82)
Comité de rédaction :
Marylin Arndt (81), Céline Beillouin (2011), Skander Benattia (2009), Christine Chardon (95), Michel Cochet (73), Jean-Pierre Dardayrol (77), Stéphane Debarbieux (84), Hélène Delahousse (93), David Fayon (93), Ayoub Figuigui (2011), Grégoire Galievsky (2000), Philippe Hilsz (80), Michel Mairal (97), Mejdi Mrad (2011), Pascal Verveur (88)

Direction artistique et réalisation : Valérie Mounier
Dessinateur : Gédéon
Banque d'image : Thinkstock
Les illustrations des articles sont fournies par les auteurs, sous leur responsabilité concernant les droits de reproduction. Les idées exprimées dans cette revue engagent la seule responsabilité de leurs auteurs. Reproduction autorisée avec mention d'origine après accord de la publication.

Rédaction & Abonnements :
46 rue Barrault
75634 Paris Cedex 13
Tél. 01 45 81 74 77
Courriel : revue@telecom-paristech.org
Site : www.telecom-paristech.org

Régie publicitaire : FFE
15 rue des Sablons - 75116 Paris
Directeur de la publicité :
Patrick Sarfati - Tél. 01 53 36 20 40
Chef de publicité : David Sellam
Tél. 01 48 05 26 65 - Email : david.sellam@ffe.fr
Assistante de fabrication :
Aurélie Vuillemin - Tél. 01 53 36 20 35

Imprimé en france

CCP n° 29854 N Paris
Abonnements annuels 2011 : 53 € TTC
Prix au numéro : 21 € TTC
ISSN 0040-2478

Big data et protection des données personnelles : un défi (quasi) impossible ?

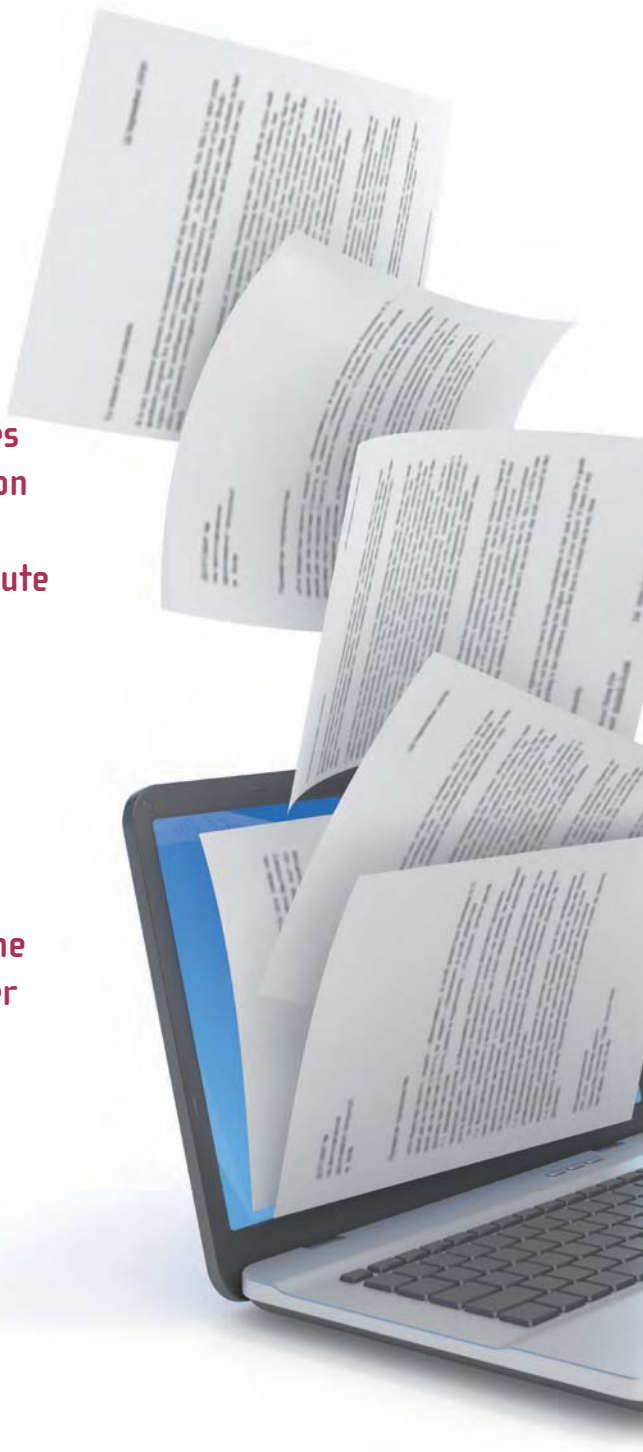
par Claire LEVALLOIS-BARTH

Les avancées technologiques récentes ont étendu le champ des informations disponibles produites par l'individu lui-même ou par les machines.

A cet égard, le phénomène « Big data » se caractérise par le volume des informations traitées et leur variété. Le traitement d'énormes volumes de données structurées *via* le datamining par les assureurs et les sociétés de télécommunications n'est pas récent. La nouveauté réside dans l'hétérogénéité des sources et des formats de données (formes structurées ou non structurées telles que les flux d'images de caméras de vidéosurveillance). Au volume et à la variété s'ajoute également la vitesse d'analyse de la donnée et la compréhension des relations entre les données. Ainsi, les entreprises, les administrations et les individus disposent à bas coût d'informations combinées, partagées et dupliquées.

Cette ère des données omniprésentes et de leurs usages en cours de définition interroge notre système de protection des droits fondamentaux, en particulier le droit à la protection des données personnelles. En France, l'usage de ce type de données est règlementé par la loi Informatique et Libertés[1] qui, dans sa version modifiée, transpose la directive européenne Protection des données[2].

Comment cette législation s'applique-t-elle ? Est-elle adaptée aux enjeux posés par le big data ?



Qu'est-ce qu'une donnée « personnelle » ?

Selon l'article 2 de la loi Informatique et Libertés, la donnée personnelle concerne « toute 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 », étant précisé que « *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 large couvre à la fois des **informations directement nominatives** (nom et prénom d'une personne, son adresse postale et de courrier électronique de type toto.dupond@abc.fr) et des **informations indirectement nominatives** : le numéro de téléphone d'un individu, sa géolocalisation, le segment comportemental qui lui est rattaché, etc. A cet égard, le Groupe de travail de l'article 29, qui regroupe l'ensemble des autorités de protection des données personnelles de l'Union européenne dont, pour la France, la Commission Nationale de l'Informatique et des Libertés (CNIL), considère que l'adresse IP est une donnée personnelle, sauf exception[3].

Ainsi, dès qu'une information est « relative » de près ou de loin à un individu, elle constitue une donnée personnelle. Une piste de réflexion porte alors sur son anonymisation.

Une donnée personnelle est-elle anonymisable ?

L'anonymisation suppose de détruire le lien entre l'information et l'identité d'une personne à travers diverses méthodes (anonymisation, pseudonymisation, chiffrement irréversible, etc.).

En pratique, la « dé-identification » est difficile à obtenir car il ne s'agit pas de considérer uniquement une information isolée mais de prendre en compte les croissements possibles entre les informations. **Or, le big data, ainsi que l'open data, accroissent considérablement les possibilités de recoupement et donc d'identification d'une personne.** Un cas célèbre est celui du site web collaboratif d'évaluation et de recommandation de films de la société américaine Netflix. Netflix a publié, dans le cadre d'un concours visant à améliorer son algorithme de recommandation, cent millions de données d'évaluation anonymes. Des chercheurs ont recoupé ces données avec d'autres notations de films non anonymes : la connaissance de deux notes leur a permis d'identifier 68 % des utilisateurs. Face au risque de condamnation pour non respect de la vie privée de ces clients, Netflix a mis fin au concours.

Est-ce à dire que l'anonymat devient une impossibilité algorithmique puisque la quantité de données augmente le repérage des personnes ? Une conclusion possible serait d'estimer que toutes les données devraient être considérées comme des données personnelles et donc relevant du champ d'application de la loi Informatique et Libertés.

Cette conclusion nous parait contre-productive. D'une part, elle inciterait les organismes à écarter l'anonymisation ce qui augmenterait les risques d'atteinte aux données personnelles, à la vie privée des individus ou à leur liberté d'expression. Or la dé-identification est devenue une composante-clé de nombreux business models, en particulier dans le domaine de la santé (on pense ici aux essais cliniques) ou de la publicité comportementale en ligne. D'autre part, une information dé-identifiée, bien que comportant un risque de ré-identification, portera toujours moins atteinte aux droits fondamentaux de la personne.

Une solution consiste à trier quasiment en tant réel les informations qui seront enregistrées pour être analysées. A cet égard, dans un avis rendu sur des panneaux publicitaires munis de caméras et de dispositifs d'analyse du comportement des passants, la CNIL considère que « même si ces données sont anonymisées à très bref délai et si seules des données statistiques sont conservées à l'issue du traitement, il n'en demeure pas moins que celui-ci est réalisé à partir d'informations permettant d'identifier des personnes ». Dès lors, la loi Informatique et Libertés s'applique alors même que les images ne sont pas enregistrées[4].

Quelles données personnelles collecter ?

Selon l'article 6 de la loi Informatique et Libertés, les données personnelles doivent être collectées et traitées pour des finalités (c'est-à-dire des usages) « déterminées, explicites et légitimes » ; elles ne doivent pas être traitées ultérieurement de manière incompatible avec ces finalités. **Ce principe de finalité est un préalable au principe de qualité des données.** D'une part, seules les données nécessaires et pertinentes pour atteindre les finalités doivent être collectées. D'autre part, **la durée de conservation des données ne doit pas excéder la durée nécessaire aux finalités** pour lesquelles elles sont collectées. Passé ce délai, les données doivent être détruites. Apparaît ainsi le droit à l'oubli.

Mais que signifient ces règles dans un monde de données massives ? Dans ce monde où les utilisations possibles des informations peuvent difficilement être

[1] *Loi n°78-17 du 6.01.1978 relative à l'informatique, aux fichiers et aux libertés, JORF 7.01.1978, p. 227 (Loi dite Informatique et Libertés).*
[2] *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, JOCE n° L 281 du 23/11/1995, p. 31.*
[3] *Groupe Article 29, Avis 2/2007 sur le concept de données à caractère personnel, WP 136 adopté le 20 juin 2007.*
[4] *CNIL, Dispositifs d'analyse du comportement des consommateurs : souriez, vous êtes comptés ! 19 avril 2010 http://www.cnil.fr/linstitution/actualite/article/article/dispositifs-danalyse-du-comportement-des-consommateurs-souriez-vous-etes-comptes-2/*

anticipées et où le principe d'une collecte minimaliste des données et de leur suppression est en contradiction avec la définition même du big data.

Cette difficulté pose la question du contrôle, par la personne elle-même, de ses propres données et de l'image informationnelle qui en résulte.

Quel contrôle pour la personne concernée ?

Cette question du contrôle suppose que la personne soit informée de l'existence même de la création des données. Reconnu par l'article 32 de la loi Informatique et Libertés, **ce droit d'information porte aussi bien sur la collecte des données que leur utilisation**[5]. Il est primordial car il conditionne l'exercice du droit d'accès aux données personnelles et du droit d'opposition. Soulignons que le droit d'information est allégé lorsque les données collectées sont très vite anonymisées ou lorsque les données ne sont pas recueillies directement auprès de la personne. Il est même exclu « lorsque l'information de la personne se révèle impossible ou très difficile ».

Mais c'est sans nul doute le **consentement** qui offre – ou est supposer offrir – à la personne un véritable pouvoir[6]. Le consentement désigne « toute manifestation de volonté libre, spécifique et informée »[7]. Cela signifie que la permission doit être fournie dans un contexte spécifique et déterminé ce qui ne correspond pas aux usages actuels, notamment selon la CNIL aux pratiques de Google[8]. On constate également que le consentement peut être biaisé. Des chercheurs ont ainsi démontré qu'en créant un simple sentiment de contrôle, la personne est encouragée à donner son autorisation, indépendamment de la question de savoir si elle a effectivement acquis un contrôle réel sur la donnée personnelle[9].

Le principe de consentement est-il approprié ? Nous pensons que oui, mais en déterminer les contours s'avère souvent difficile. Sans doute convient-il d'élargir quelque peu la liste des exceptions dressées par la loi Informatique et Libertés en tenant compte du fait que des traitements big data peuvent poursuivre un objectif « d'intérêt général » et apporter un avantage important à l'ensemble de la société, par exemple en prédisant des épidémies ou des accidents de voitures. En exploitant des données GPS, des chercheurs ont montré qu'ils pouvaient prédire la position d'une personne à 80 semaines, avec une précision de plus de 80 %. Mais il convient sur ce point d'être vigilant car les prédictions peuvent également permettre de déterminer et d'analyser la personnalité d'un individu, en particulier son comportement, sa situation économique. Ce faisant elles laissent entrevoir des possibilités importantes de discrimination.

La réforme à venir

A cet égard, le profilage constitue sans nul doute l'un des principaux challenges pour la protection des libertés fondamentales car les big data fonctionnent à une échelle qui dépasse la compréhension humaine et ne permettent pas d'expliquer la base de la prédiction. Comment alors éviter qu'elles ne deviennent des boites noires en dehors de toute traçabilité et de toute responsabilité ?

C'est à cette question et bien d'autres que tente de répondre la révision de la direction Protection des données[10]. L'objectif est de relever les défis posés par le développement des nouvelles technologies tout en réalisant une application effective des règles que nous venons brièvement de présenter.

La Chaire de recherche de l'Institut Mines-Télécom « Valeurs et politiques des informations personnelles »

Créée en partenariat avec le Groupe Imprimerie Nationale, BNP Paribas et Dassault Systèmes, la chaire se propose d'aider les entreprises, les citoyens et les pouvoirs publics dans leurs réflexions sur la collecte, l'utilisation et le partage des informations personnelles, à savoir les informations concernant les individus (leurs vies privées, leurs activités professionnelles, leurs identités numériques, leurs contributions sur les réseaux sociaux, etc.) incluant celles collectées par les objets communicants qui les entourent (smartphones, compteurs intelligents, etc.).

La chaire regroupe une équipe pluridisciplinaire de chercheurs travaillant à la fois sur les aspects juridique de régulation et de conformité, technique de sécurité des systèmes et des données, économique de partage des informations personnelles, et philosophique de responsabilisation et d'anticipation des conséquences sociétales.

Cinq axes de recherche ont été déterminés : identités numériques, gestion des informations personnelles, contributions et traces, informations personnelles dans l'internet des objets et politiques des informations personnelles.

Pour plus d'informations : **www.informations-personnelles.org**

L'AUTEUR

Claire LEVALLOIS-BARTH est docteur en droit, et coordinatrice de la Chaire de recherche de l'Institut Mines-Télécom Valeurs et politiques des informations personnelles.

5 *Art. 32-I de la loi Informatique et Libertés selon laquelle responsable de traitement doit notamment informer la personne concernée de son identité, de la finalité poursuivie par le traitement, des destinataires des données et du caractère obligatoire ou facultatif des réponses. L'information doit également porter sur les transferts de données envisagés à destination d'un État non membre de la l'Union européenne.*

6 *Art. 7 de la loi Informatique et Libertés.*

7 *Art. 2- h de la directive Protection des données.*

8 *CNIL, 16 oct. 2012, Règles de confidentialité de Google : une information incomplète et une combinaison de données incontrôlée*

9 *L. Brandimarte, A. Acquisti & G. Loewenstein, Misplaced Confidences: Privacy and the Control Paradox, sept. 2010.*

10 *Voir Commission européenne, Preparing Data Protection Reform, http://ec.europa.eu/justice/data-protection/review/actions/index_en.htm.*