GLOSSAIRE « ANONYMISATION DE DONNEES »

Dans le cadre de ses travaux sur le thème de l'anonymisation de données, le groupe *Référentiels et Labels* de l'AFCDP a jugé utile de produire un glossaire des termes rencontrés dans la rare littérature dédiée à ce sujet.

Ce glossaire n'a pas l'ambition de se poser en référence académique. Son unique objectif est de faciliter, par le Correspondant Informatique & Libertés ou toute personne concernée par la protection des données à caractère personnel, la compréhension des divers documents, livres blancs, brochures ou articles faisant référence aux techniques d'anonymisation.

L'ambition des auteurs était de parvenir à un document accessible aux non informaticiens sans que ces derniers puissent considérer le texte comme simpliste ou réducteur. L'accueil qui sera accordé à ce document nous dira si cet objectif a été atteint.

Les éléments du glossaire comprennent maints néologismes et anglicismes, dont de nombreux termes empruntés à d'autres spécialisations du monde informatique (exemple : *data masquerading*, utilisé à l'origine dans le domaine du réseau), voire à d'autres métiers (exemple : *floutage*, tiré du jargon des graphistes).

À chaque fois que c'est possible, la ou les sources principales ont été indiquées.

Ce document est perfectible et a naturellement vocation à évoluer. Le groupe *Référentiels et Labels* invite les membres de l'AFCDP à faire part de toute remarque et proposition visant à l'améliorer.

Les membres du groupe doivent être remercié pour le travail accompli :
- Eric Barbry, Alain Bensoussan Avocats
- Yann le Hegarat, +SELF+
- Bernard Lombardo, Ugap
- Alain Rouffiat, Princeton Softech
- Michel Simion, Compuware

Une mention toute particulière sera décernée à Gilles Trouessin, Oppida, qui a éclairé le groupe sur les travaux antérieurs réalisés par le groupe sécurité/santé de l'AFNOR et à Bruno Rasle, Cortina, qui a assuré l'essentiel du travail de rédaction de ce Glossaire.

Arnaud Belleil
Cecurity.com
Administrateur AFCDP

Abduction.- Voir inférence.

Adduction.- Voir inférence.

Aging.- Voir vieillissement.

Agrégation.- Action qui consiste à regrouper des informations relatives à des personnes ou à des ménages dans des groupes d'une taille suffisante pour qu'il ne soit plus possible de retrouver des informations relatives à une seul personne ou à un seul ménage. Cette approche est notamment mise en œuvre pour la mise à disposition des données de recensement de la population avec une agrégation qui s'effectue au niveau de l'îlot.

Anonyme.- Du grec anônumos, ανωνυμία, sans nom, dont le nom est inconnu. Dans le domaine informatique, se dit d'une façon générale d'une donnée qui ne peut être rattachée à une identité, à un individu spécifique.

Anonymat.- L'anonymat est la qualité de ce qui est sans nom, l'état d'une personne dont on ignore l'identité. En informatique, c'est l'impossibilité de déterminer le véritable nom de l'utilisateur d'un outil informatique (révélation de l'identité) ou d'un individu auquel sont rattachées des données à caractère personnel. L'anonymat peut être définie comme l'impossibilité (pour un tiers) de déterminer le véritable nom de cet utilisateur.

Anonymisation.- *Syn.* Data Masking, Data Cloaking, Data Masquerading. Processus par lequel des données sont rendues anonymes, processus à l'issue duquel elles ne peuvent plus être affectées ou rattachées à une personne en particulier, à un individu.

Dans une acception assez stricte, l'anonymat consiste en la possibilité de suivre une personne unique dans la durée (caractéristiques, comportements, …) sans avoir la moindre possibilité de connaître sa véritable identité.

Dans une acception plus large – et il est même possible de contester qu'il s'agisse d'une véritable anonymisation – l'anonymisation peut résulter de la suppression des données à caractère personnel. Dans ce cas de figure, la notion de suivi dans la durée n'est plus possible.

Appauvrissement.- *Syn*. floutage. Action qui consiste à rendre les données moins précises, en effectuant des suppressions sélectives. Par exemple ; un libraire conserve le détail des titres des livres commandés par ses clients durant une certaine période, et supprime par la suite cette précision en ne conservant que la catégorie des ouvrages commandés (histoire, biographie, etc.). Par exemple ; la suppression de l'adresse physique dans un annuaire téléphonique, ou le fait de ne laisser que l'initiale du prénom de l'abonné.

Chiffrement.- En cryptographie, le chiffrement est le procédé qui s'appuie sur un algorithme de chiffrement et une clé, grâce auquel on peut rendre la compréhension d'un document impossible à toute personne qui n'a pas la clé de (dé)chiffrement. Le chiffrement peut être utilisé lors d'un processus d'anonymisation réversible (voir ce terme), mais il ne conserve pas, en général, le format des données. On préférera le terme chiffrement au terme « cryptage »,

qui serait un anglicisme tiré de l'anglais encryption même si on peut le trouver dans de nombreux usuels.

Collision.- Par analogie avec son sens premier (un choc entre deux objets), on appelle collision le fait que deux individus donnent, après processus d'anonymisation par hachage, le même résultat, ce qui ne doit pas se produire. Une bonne technique d'anonymisation doit être résistante aux collisions, c'est-à-dire que deux messages distincts doivent avoir très peu de chances de produire le même résultat, la même signature.

Concaténation.- Remplacement par une valeur issue de la combinaison de champs figurant dans la source. C'est l'une des techniques d'anonymisation utilisée pour conserver le format et la validité au sein d'un jeu de données utilisé pour tester une application informatique.

Data Cloaking.- *Syn.* anonymisation (voir ce terme).

Data Masquerading.- Voir anonymisation. Dans la littérature anglo-saxonne, ce terme recouvre soit le processus d'anonymisation, soit, dans le domaine des réseaux informatiques, la translation d'adresse (NAT, pour network address translation) où, à une adresse IP, on en fait correspondre une autre.

Data Masking.- *Syn.* anonymisation (voir ce terme).

Déduction.- *Syn.* inférence (voir ce terme).

Désidentification.- *Syn.* Anonymisation (voir ce terme).

Données indirectement nominatives.- Les données indirectement nominatives sont celles qui permettent d'identifier une personne bien qu'elles ne soient pas accompagnées d'une identité : toute forme de numéro ou d'immatriculation, (téléphone, voiture, adresse IP, n° de sécurité sociale, numéro fiscal…). Ces données sont indirectement nominatives car il faut pouvoir rapprocher l'information d'une table de conversion afin de faire le lien entre un n° et une personne. Cette notion fait référence à l'article 4 de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, qui s'applique aux « *informations qui permettent, sous quelque forme que ce soit, directement ou non, l'identification des personnes physiques auxquelles elles s'appliquent, que le traitement soit effectué par une personne physique ou par une personne morale* ».

Des objets sont souvent considérés comme des données indirectement nominatives permettant d'identifier le propriétaire et/ou l'utilisateur de l'objet : plaque d'immatriculation des véhicules, numéro d'un téléphone, adresse IP d'un ordinateur, numéro de référence figurant dans une puce RFID insérée dans un vêtement, etc.

Effacement.- *Syn.* Suppression (voir ce terme).

Encodage.- *Syn.* encodage, transcodage. En sémantique, un encodage est un procédé de transformation d'un langage formel en un autre langage formel. On préférera utiliser le terme codage qui est plus correct, le terme « encodage » s'étant répandu dans le milieu informatique sous l'influence de l'anglais *encoding*. (Source Wikipedia)

Floutage.- Terme issu du jargon des graphistes : fait de rendre floue la totalité ou une partie d'une image. En anonymisation de données, soit *Syn.* d'appauvrissement, soit réellement appliqué aux images, pour rendre impossible l'identification de personnes figurant dans un film ou sur une photo.

Génération de données.- Technique d'anonymisation qui remplace la valeur par des données fictives, aléatoires ou extraites d'une table externe (librairie) – par exemple des libraires de prénoms, noms, type de voie, adresses email, noms de ville, n° de Sécurité Sociale, n° de cartes bleues, etc.

Hachage.- *Syn*. Hash. Technique d'anonymisation qui remplace la valeur par une empreinte (utilisée dans les expérimentations d'anonymisation du Dossier Médical Personnalisé). Une fonction de hash (anglicisme) ou fonction de hachage est une fonction irréversible qui associe à un grand ensemble de données un ensemble beaucoup plus petit (de l'ordre de quelques centaines de bits) qui est caractéristique de l'ensemble de départ. Le résultat de cette fonction est appelé un hash ou une empreinte. Le Hachage sera considéré comme d'autant plus efficace que le taux de collision sera faible.

Homonymie.- Caractère de personnes portant le même nom et pouvant être confondues Ce cas doit être correctement traité par le processus d'anonymisation. Le fait de posséder un prénom et un nom courant assure une forme d'anonymat par rapport aux recherches effectuées avec un moteur de recherche sur les contenus du Web.

Identification par croisement, par recoupement.- Voir inférence.

Induction.- Voir inférence.

Inférence.- Dans sa définition classique, opération mentale qui consiste à tirer une conclusion d'une série de propositions reconnues pour vraies. Ces conclusions sont tirées à partir de règles de base. C'est pourquoi l'inférence est souvent *Syn.* de déduction. Appliqué au domaine de l'anonymisation de données, ce terme correspond à la découverte de données identifiantes par la mise en correspondance de plusieurs données légitimement accessibles (anonymisées).
Les spécialistes distinguent trois types d'inférences : la déduction, l'induction, l'abduction et adductive (ou probabiliste).

- Inférence déductive - la déduction est un raisonnement logique (par exemple, *si* un certain patient fait un test de dépistage puis dans les quelques jours qui suivent, fait un test de dosage, *alors* le résultat du dépistage était positif)
- Inférence inductive – l'induction tire des conclusions générales sur base de cas particuliers et n'a de validité qu'en termes probabilistes (par exemple, un tel patient est très probablement atteint de telle pathologie compte tenu du fait qu'il lui est prescrit tels médicaments comme il est d'usage pour cette pathologie).
- Inférence abductive - L'abduction est un procédé consistant à introduire une règle à titre d'hypothèse afin de considérer ce résultat comme un cas particulier tombant sous cette règle (par exemple ; « *et s'il avait une maladie grave, cela expliquerait pourquoi il s'absente si souvent pour se rendre à l'hôpital Paul Brousse de Villejuif* … ». C'est aussi une forme de raisonnement intuitif qui consiste à supprimer les solutions improbables.

- Inférence adductive ou probabiliste - L'adduction consiste à estimer la vraisemblance d'une information sensible en utilisant les informations accessibles (par exemple, « *puisque P est traité à l'hôpital H, et puisque H est spécialisé dans les maladies M1 et M2, et puisque à son âge, la probabilité d'avoir M1 est très faible (10%), alors on peut déduire qu'à 90%, P est atteint de M2* »).

(Sources : Wikipedia et Anas Abou El Kalam, Yves Deswarte, Gilles Trouessin, and Emmanuel Cordonnier *in* « *Une démarche méthodologique pour l'anonymisation de données personnelles sensibles* », 2004)

Inversibilité.- Propriété d'un processus, d'une technique d'anonymisation qui permet de « remonter », depuis les données anonymisées, jusqu'aux données nominatives originelles en appliquant une procédure exceptionnelle et formalisée, sous surveillance d'une instance légitime (par exemple médecin-conseil, médecin inspecteur, dans le cas de données de santé) garante du respect de la vie privée des individus concernés. Il s'agit donc d'une pseudonymisation (voir ce terme).

Irréversibilité.- Propriété d'un processus, d'une technique d'anonymisation qui rend impossible toute « remontée » aux données nominatives originelles à partir des données anonymisées. C'est le cas réel de l'anonymisation ; une fois remplacés par des identifiants anonymes, les identifiants nominatifs originels ne sont plus recouvrables ; cependant, avec les techniques d'attaques par inférence, les identifiants anonymes, s'ils sont trop universellement utilisés, risquent de permettre la découverte d'identités mal cachées; pour ce type d'anonymisation, la technique communément utilisée est une fonction de hachage ;

Masquage.- Dissimulation d'une partie des champs des données identifiantes (par exemple par des X ou par des zéro). Technique utilisée depuis plusieurs années pour sécuriser les numéros de téléphone composés (sur les factures détaillées) ou les numéro de cartes bleues.

Mélange.- *Syn.* Shuffle. Technique d'anonymisation où les données sont « brassées » sans être modifiées. Elle ne doit être utilisée que dans des corpus d'une taille suffisante pour éviter de faciliter les identifications par déduction.

Obfuscation.- *Syn.* obscurcissement, assombrissement. À l'origine, technique utilisée par certains développeurs de programmes informatiques pour rendre la compréhension de leur code très difficile pour un humain, tout en le maintenant parfaitement compilable par un ordinateur. La finalité de cette démarche est de protéger les investissements de développement et de rendre plus difficile la rétro-ingénierie. Dans le domaine de l'anonymisation de données, cette technique consiste à ajouter aux entrées réelles des données totalement fictives, afin de « noyer » les données pertinentes au milieu d'une masse d'informations insignifiante. (Source Wikipedia). Pour utiliser une formule imagée, l'obfuscation consiste à dissimuler une aiguille en l'ensevelissant sous une botte de foin.

Pseudonymat.- Situation d'anonymat qui peut être levée dans certaines conditions très strictes et formalisées par des tiers spécifiques (par exemple par les autorités judiciaires). Le pseudonymat ajoute à l'anonymat le fait que l'utilisateur peut être tenu responsable de ses actes ; par exemple, en cas de litige ou d'enquête (lutte contre le blanchiment d'argent sale, par exemple), la propriété requise est la pseudonymat (et non l'anonymat) car certaines informations personnelles doivent pouvoir être fournies aux autorités judiciaires. L'anonymat

peut être levé, a posteriori, et nécessite donc l'utilisation de techniques inversibles (voir ce terme).

Pseudonyme.- Du grec *pseudès*, faux, et *onoma*, nom. Nom d'emprunt pour dissimuler son identité. Moyen de ne pas être identifiable de tous, mais éventuellement de certains. On parle également de pseudonymes lorsqu'il y a un usage de plusieurs identités – de plusieurs noms d'emprunt - par une même personne.

Pseudonymatisation.- Processus par lequel des données perdent leur caractère nominatif sans être pour autant être anonymes. Elles ne pourront être affectées ou rattachées à une personne en particulier, à un individu que dans des conditions bien particulières.

Randomization.- (anglicisme). Technique qui voit les données identifiantes remplacées par des données aléatoires. Introduction du hasard dans le processus concerné.

Réversibilité.- Propriété d'un processus, d'une technique d'anonymisation autorisant de « remonter », depuis les données anonymisées, jusqu'aux données nominatives originelles. Le chiffrement en est un exemple, dans sa phase de déchiffrement (recouvrement des données d'origine grâce à la clé).

Réversion.- Processus inverse de la fonction d'anonymisation, qui permet, à partir des données anonymes, de retrouver les données identifiantes.

Robustesse.- La robustesse d'un système d'anonymisation est constituée de l'ensemble des caractéristiques à satisfaire vis-à-vis d'attaques ayant pour but de lever l'anonymat de façon non-autorisée. Il peut s'agir d'une robustesse à la réversion concernant la possibilité d'inverser la fonction d'anonymisation, mais il peut aussi s'agir d'une robustesse à l'inférence qui consiste à déterminer des informations nominatives à partir d'éléments d'informations purement anonymes.

Shuffle.- Voir [*Syn.*] mélange.

Suppression.- Destruction définitive et totale des données identifiantes.

Transcodage.- *Syn.* encodage (voir ce terme).

Translation.- Remplacement des valeurs sensibles par des données compréhensibles à partir d'une table de translation.

Variance.- Modification de données chiffrées dans une plage de valeurs pré-définies. Par exemple, pour le poids d'un patient, appliquer une transformation entre – et + 10%.

Vieillissement.- *Syn.* Aging. Remplacement cohérent des dates sensibles tout en maintenant le format initial