REFERENTIEL AFCDP DES DISPOSITIFS D'ANONYMISATION

Paris, le 22 mai 2008

Dans le cadre de ses travaux sur le thème de l'anonymisation de données, le groupe *Référentiels et Labels* de l'AFCDP propose la création d'un référentiel AFCDP portant sur les outils d'anonymisation de données.

Le premier objectif visé est de permettre de distinguer ce qui peut être désigné sous le vocable « d'outil d'anonymisation ».

Ce référentiel vise également à favoriser la diffusion des technologies de protection des données à caractère personnel. Il s'applique à la sécurisation des données de tests et de maintenance, utilisées soit en interne soit dans un contexte d'externalisation (*near* et *offshore*).

Un niveau minimal d'exigence est clairement établi (« référentiel de base »), tandis que des fonctionnalités avancées font partie d'un niveau d'exigence supérieur (« référentiel étendu »). Toutes les valeurs ajoutées des produits d'anonymisation n'ont pas vocation à figurer dans le référentiel dans la mesure où elles ont vocation à faire partie du domaine de la concurrence.

Le document pourra faciliter la tâche des acheteurs (conception d'un cahier des charges, constitution d'une *short list*, etc.) et celle des vendeurs (amélioration des offres, confiance et différentiation).

L'AFCDP affecte à ce projet les caractéristiques suivantes ;

- Simplicité : ces travaux s'inscrivent dans les efforts de « pédagogie de l'anonymisation » engagés par l'AFCDP.
- Neutralité : ce référentiel est accessible aux progiciels, aux développements internes comme aux logiciels libres.
- Accessibilité : ce référentiel peut permettre de réduire le coût d'accès à un éventuel label associé, en prenant en compte dès sa conception la tâche des auditeurs.
- Ouverture : les travaux donnent lieux à publication et appel à améliorations. La démarche n'est pas propriétaire.

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

On pourra se reporter avec profit aux deux précédents livrables du groupe ;

- Glossaire « Anonymisation des Données » édité par le même groupe de travail
- Liste des questions que doit se poser le CIL pour un projet d'anonymisation de données

Les membres du groupe doivent être remerciés pour le travail accompli.

Arnaud Belleil
Cecurity.com
Administrateur AFCDP

AFCDP, Association loi 1901 créée le 10 sept. 2004 et déclarée à la Sous-préfecture de Grasse sous le n° 20040040
www.afcdp.org

Référentiels AFCDP des dispositifs d'anonymisation

Ces référentiels se focalisent sur les caractéristiques indispensables qui font d'un outil un moyen d'anonymisation.

Le référentiel de base correspond au niveau d'exigence minimal.

Un second référentiel, intitulé « référentiel étendu » comprend en sus d'autres caractéristiques simplement optionnelles.

Mais au-delà des caractéristiques de l'outil d'anonymisation, on prendra soin de ne pas méconnaitre l'importance de deux facteurs primordiaux ;

- L'expertise de l'opérateur
- La méthode utilisée pour opérer

On soulignera qu'il est très facile de produire un jeu de données insuffisamment anonymisé en utilisant un outil répondant à ce référentiel.

L'objet de ce projet n'est donc pas de proposer une grille d'évaluation permettant de décerner un quelconque prix d'excellence à un outil ou à un autre, mais bien d'identifier les fonctionnalités que doit présenter un outil d'anonymisation, outil qu'un opérateur averti saura utiliser avec discernement.

Dans l'attente de la publication d'un document focalisé sur les *best pratices* (méthode à utiliser pour anonymiser des données personnelles, pièges à éviter, traitement des difficultés principales), voici quelques éléments sur ces aspects ;

Meilleures pratiques :

Faire appel à son Correspondant Informatique & Libertés en amont du projet
Adopter une démarche « projet », en quatre étapes ;
Analyse : identification des données candidates, des objectifs et contraintes
Conception : définition des stratégies (d'extraction, de transformation, de chargement, de sécurisation, de gestion, etc.)
Mise en œuvre : construction de la plate-forme outillée et déploiement
Maîtrise-contrôle : formalisation, formation des personnels, documentation, traçabilité, exploitation sécurisée, etc.
Avoir une bonne connaissance de l'existant (schéma des données, logique des applications concernées, etc.)
Avec une idée claire des objectifs à atteindre
Mobiliser tous les acteurs concernés (sécurité, production, études – développement, systèmes, DBA, etc.)

Quelques points de vigilance :

Prise en compte des relations (au sein des lignes et des tables, entre les tables) au moment de l'extraction des données sources (dépendances fonctionnelles)

AFCDP, Association loi 1901 créée le 10 sept. 2004 et déclarée à la Sous-préfecture de Grasse sous le n° 20040040
www.afcdp.org

Traitement des homonymies[1]
Réflexion quant à l'ordre des traitements d'anonymisation
Présence de méta-données
Présence de données signifiantes (ex. NIR[2] et données avec *checksum*[3])
Présence de texte libre (ex. zone commentaires ou bloc note)
Isolated Case Phenomena[4]
Cas spécifiques (ex. un employé ayant quitté l'entreprise, puis embauché à nouveau)

Connaissances indispensable à l'opérateur :

Objectifs poursuivis
Contraintes (ex. propriétés des champs ; type, format, contrainte d'intégrité, etc.)
Schéma de données (bases de données sources et cibles)
Logiques des applications impliquées
Avantages et inconvénient de chaque technique d'anonymisation

Intrinsèquement, aucune technique d'anonymisation n'est « bonne » ou « mauvaise ». Elles présentent toutes des avantages et des inconvénients. L'opérateur doit donc les utiliser avec discernement, en tenant compte des objectifs poursuivis et des contraintes existantes.

Quelques exemples :

- La technique du mélange (« shuffle »), qui semble très frustre, présente de grande qualité lorsqu'il est exigé le maintien des distributions dans le jeu de données.
- La technique de la variance permet de banaliser des salaires ou des dates de naissance, tout en conservant une vraisemblance (cette technique peut également permettre de conserver les distributions).
- La technique du hash, qui présente un haut niveau de sécurité, ne permet pas de conserver les caractéristiques de la donnée source (format, type) et perd la lisibilité. Par contre, elle donne un résultat unique.
- L'utilisation non maîtrisée de la technique du vieillissement donne quelquefois des résultats étonnants : un PDG âgé de 4 ans ou un collaborateur décédé avant d'avoir rejoint l'entreprise.

[1] Après anonymisation, un père et un fils devront-ils toujours avoir le même nom ? Deux homonymes devront-ils le rester après anonymisation ?
[2] Le NIR – plus connu sous l'appellation de « N° de Sécurité Sociale » est signifiant dans la mesure où il permet de prendre connaissance du sexe, du mois, de l'année et de la commune de naissance de son propriétaire. A l'inverse, le N° Siren ne porte aucune signification en lui-même sur l'entreprise identifiée.
[3] Technique qui permet de vérifier que la donnée n'a pas été altérée, par exemple durant un envoi. Le *checksum* est calculé à partir de la donnée. Le destinataire la recalcule pour valider l'intégrité de l'information. Le procédé d'anonymisation doit respecter cette contrainte, sous peine de voir l'application refuser la donnée, pour défaut d'intégrité.
[4] Ex. PDG « révélé » par le salaire le plus élevé

AFCDP, Association loi 1901 créée le 10 sept. 2004 et déclarée à la Sous-préfecture de Grasse sous le n° 20040040
www.afcdp.org

Référentiels AFCDP des dispositifs d'anonymisation

Référentiel « de base »

1 - Un dispositif d'anonymisation doit permettre l'extraction de données à caractère personnel (données sources) en conservant les relations et dépendances fonctionnelles (au sein d'une ligne, au sein d'une même table, entre tables).

2 - Un dispositif d'anonymisation doit proposer un large choix[5] de méthodes d'anonymisation parmi les méthodes suivantes :

- Appauvrissement[6]
- Masquage
- Suppression
- Chiffrement
- Vieillissement ou décalage
- Génération de données ou remplacement par des données fictives
- Mélange de données
- Calcul d'empreinte (hash)
- Variance[7]
- Concaténation[8]
- Obfuscation[9]

3 - Un dispositif d'anonymisation doit permettre de choisir – pour chaque type de données – la technique qui semble appropriée, en fonction des objectifs poursuivis et contraintes constatées.

4 - Un dispositif d'anonymisation doit permettre une utilisation souple de ces méthodes d'anonymisation.

Exemple : « Pour ce type spécifique de données, appliquer la méthode d'anonymisation a tant que [condition A remplie], appliquer la méthode b si [condition B remplie] et ne rien faire quand la condition C est observée ».

Autres exemples :

- entre deux types de données : masquer l'information A s'il s'agit d'une femme.
- pour une même donnée : appliquer une variance de 10% sur le champ Salaire pour toute valeur > valeur de référence.

[5] Une réflexion est en cours pour préciser cette notion. Dans le contexte d'une labellisation, une formulation plus précise permettrait de réduire la marge d'appréciation de l'auditeur.
[6] Perte de sens (ex. Au lieu d'indiquer la date de naissance précise, on indique une fourchette)
[7] La donnée (type numérique) subit une variation au sein d'une fourchette prédéfinie (ex. Faire varier le salaire de – 20 à +15%)
[8] La donnée cible est obtenue à partir de « l'association » de plusieurs données source (ex. Faire la moyenne des dix salaires précédents et des dix salaires suivants)
[9] 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.

5 - Un dispositif d'anonymisation doit permettre de créer son propre corpus de données banalisées, de l'importer et de l'utiliser pour remplacer certaines données (méthode concernée : remplacement par des données fictives).

Remarque : Cette facilité est indispensable pour pouvoir répondre à toute contrainte spécifique, voir nationale. Certains outils mettent à disposition de tels corpus (liste de prénoms féminins italiens, liste de faux numéro de Cartes bancaires utilisées aux Etats-Unis, liste de villes européennes, faux numéro de téléphones français, etc.)

6 - Un dispositif d'anonymisation doit permettre – sur une même donnée – l'association de plusieurs techniques.

Exemple : Le remplacement couplé à du masquage, utilisé notamment pour les numéros de téléphone (02.54.98.56.23 donne 01.31.06.XX.XX).

7 - Un dispositif d'anonymisation doit permettre à son opérateur de maîtriser le séquencement (l'ordre) dans lequel les opérations d'anonymisation sont effectuées (pour se plier aux relations entre les tables de la base source)[10].

8 - Un dispositif d'anonymisation doit permettre le chargement de données après banalisation (données cibles) en conservant les relations (au sein d'une ligne, au sein d'une même table, entre tables).

Remarque : Ce référentiel s'applique à la sécurisation des données de tests et de maintenance, utilisées soit en interne soit dans un contexte d'externalisation. Les données banalisées obtenues doivent impérativement être manipulables par les applications. Dans ce contexte, obtenir des données banalisées, mais totalement inutilisables, ne répond pas aux objectifs poursuivis.

9 - Un dispositif d'anonymisation doit proposer à son opérateur un choix de techniques d'anonymisation et une souplesse de configuration qui lui permettent de concevoir une stratégie de transformation qui offre une résistance raisonnable[11] aux tentatives de réversibilité (obtenir les données sources à partir des données banalisées).

Remarque : Comme nous l'avons indiqué en préambule, la robustesse résulte d'un ensemble de critères très divers : le choix des données à banaliser, le choix de la technique utilisée pour anonymiser chaque type de données, la taille du corpus, les éventuelles indications qui accompagnent le fichier, etc. Elle est donc fortement aux *best practices* et à l'expérience de l'opérateur. Toutefois un outil proposant un large choix de techniques d'anonymisation et offrant toute souplesse dans leur mise en œuvre concoure à la conception d'un jeu de test banalisé et robuste.

[10] Dans certains cas, si on anonymise une donnée A, il devient impossible de transformer la donnée B qui était « liée » à A (infraction aux contraintes d'intégrité).

[11] A notre connaissance, il n'existe pas à ce jour de méthode permettant de s'assurer de la robustesse d'un jeu de test banalisé, *a fortiori* il n'existe pas de méthode permettant du *benchmarking* (« Cette stratégie d'anonymisation A donne un résultat plus robuste que cette stratégie B, sur un même jeu de données source et en respectant les mêmes contraintes d'exploitation »).

10 - Un dispositif d'anonymisation doit permettre l'enregistrement des stratégies d'anonymisation mise en œuvre par l'association des différentes méthodes afin de pouvoir le rejouer ultérieurement.

11 - Un dispositif d'anonymisation doit disposer d'une documentation rédigée en français

Remarque : Cette documentation doit permettre notamment de maîtriser les manipulations mises en jeu lors de l'anonymisation des données. Il n'est pas pour autant demandé de traduire ou d'adapter la totalité de la documentation technique.

Référentiel « étendu »

Le référentiel étendu comprend l'intégralité des points composant le référentiel de base, auxquels s'ajoutent les exigences suivantes ;

12 - Un dispositif d'anonymisation doit permettre de suivre le dossier d'une même personne, non identifiable, dans la durée.

Remarque : Cette facilité de « suivi », qui dépend beaucoup plus de l'organisation globale mise en œuvre que de l'outil utilisé, est indispensable pour tout projet d'anonymisation mené dans le cadre de suivi statistiques (par exemple sur la population) ou de santé. La technique du hash permet – entre autres – de répondre à cet objectif.

13 - Le dispositif d'anonymisation doit proposer un choix de techniques qui permette à son opérateur en charge de la conception de la stratégie de banalisation des données à caractère personnel d'obtenir un résultat

- **Exempt de collisions (données source différentes donnant le même résultat)**
- **Où les cas d'homonymie sont maîtrisés**

Remarque : La taille du corpus doit être considéré – une technique sera efficace pour un corpus de taille X, mais on commencera à observer des collisions à partir d'une taille Y (par exemple à l'échelle de la population d'un pays).