AFCDP

Association Française des Correspondants à la Protection des Données a Caractère Personnel

Anonymisation : en route vers le label ?

OSSIR - Journée Sécurité des Systèmes d'Information

Paris – 23 mai 2008

Arnaud Belleil (abl@cecurity.com)

et Bruno Rasle (bruno_rasle@halte-au-spam.com)

Sommaire

- **La place de l'anonymat dans l'univers de la confiance**
- **Retours d'expérience sur les projets d'anonymisation**
- **Du lexique de l'anonymisation au Référentiel**
- **Robustesse et des-anonymisation**
- **Vers une labellisation des technologies d'anonymisation ?**

Sommaire

- **La place de l'anonymat dans l'univers de la confiance**
- **Retours d'expérience sur les projets d'anonymisation**
- **Du lexique de l'anonymisation au Référentiel**
- **Robustesse et des-anonymisation**
- **Vers une labellisation des technologies d'anonymisation ?**

Quelques usages et mésusages de l'anonymat*

- Lettre de dénonciation anonyme
- Secret bancaire
- Secret du vote
- Alcoolique anonyme
- Revendication d'anonymat pour les 500 signatures de l'élection présidentielles
- Accouchement sous X
- Anonymisation des requêtes sur les moteurs de recherche
- Argent liquide
- Télécarte, téléphone mobile sans abonnement
- Anonymat des gagnants du Loto
- Anonymat des contributeurs de Wikipedia
- Anonymat des sources
- Protections des témoins
- Centre d'appels info Sida ; numéro d'appel maltraitance
- L'identité à la Légion Étrangère
- Anonymat du don d'organes
- Anonymat des agents en fonction (GIGN, RAID)
- Anonymat des copies d'examens
- Anonymat de l'offre dans les marchés publics
- Anonymat des enquêtes, questionnaires, sondages
- Bons au porteur
- CV anonyme
- Liberté d'aller et venir

(*) Sur la base du travail réalisés par des élèves de l'EHESP (École des Hautes Études en Santé Public) de Rennes lors d'une formation dispensée le 6 mars 2008 par A. Belleil et Y Le Hegarat.

L’anonymat une valeur ambivalente et méconnue

- Une mauvaise réputation
  - Les lettres anonymes de dénonciation calomnieuse des corbeaux (exemple récent : polémique Note2b)
  - Une aide aux comportements déviants voire délictueux ; obstacle pour l’action de la police
- Des pratiques généralisées, de nouvelles pratiques
  - Secret du vote, anonymat des copies d’examens
  - CV anonyme
- Une confusion fréquente entre anonymat, pseudonymat et données indirectement nominatives

La position de la CNIL

- « *La Cnil (…) soutient les initiatives qui peuvent concourir à protéger, dans le respect évidemment de l'ordre public, un certain anonymat sur Internet* »

Rapport d'activité 1997

La position de la CNIL

- Une action constante en faveur de l'anonymat
  - Études épidémiologiques
  - Paiement électronique sur Internet
  - Anonymisation des décisions de justice publiées sur Internet
  - Mesure de la diversité
  - Anonymisation des données du recensement
  - Archivage électronique
  - Navigo anonyme
  - Etc.
- Une exception
  - Réticence par rapport à l'anonymat de l'appelant dans les dispositifs d'alerte professionnelles (*Whistleblowing)*

Les techniques d'anonymisation ne sont pas bonnes ou mauvaises : elles correspondant à des usages différents

Sommaire

- **La place de l'anonymat dans l'univers de la confiance**
- **Retours d'expérience sur les projets d'anonymisation**
- **Du lexique de l'anonymisation au Référentiel**
- **Robustesse et des-anonymisation**
- **Vers une labellisation des technologies d'anonymisation ?**

Best Practices

- Adopter une démarche « projet applicatif ».
  - Analyse : identification des données candidates
  - Conception : définition des stratégies
  - Mise en œuvre : construction d'une plate-forme outillée et déploiement
  - Maîtrise-contrôle : formalisation, formation, documentation, audit, traçabilité, etc.
- Chaque fois que c'est possible
  - Éviter la collecte de données identifiantes
  - Supprimer ces données dès qu'elles ne sont plus indispensables
  - Ne les garder que pour le délai déclaré
- Au moindre doute, contacter la CNIL

Transports parisiens

- Passage d'une carte sans contact à l'entrée d'un moyen de transport :
  - Enregistre le numéro de carte (donnée à caractère personnel)
  - Enregistre le lieu, la date et l'heure de passage.
- Finalités : régler les transports, affecter des moyens, établir des statistiques
- N° de Pass → borne → transporteur → régulateur (avec le lieu et l'heure)
- Le numéro de carte est « haché » chez le transporteur et chez le régulateur (double anonymisation)
- Partage des responsabilités

Assurance*

- Une application, 60 tables DB2, 15 fichiers VSAM
- 200 données anonymisées
- Projet de 60 jours (25 d'analyse, 10 de conception et 25 de déploiement),
- Cette assurance a décidé de mettre en œuvre un environnement dédié à l'anonymisation (outil de pilotage, serveurs dédiés, etc.).

* Source : Compuware

Ministère*

- Montée de version d'une application
- Recette de cette migration confiée à une SSI
- Nécessité de disposer d'un jeu de test banalisé, mais permettant une recette effective de la nouvelle version
- Base de données Oracle
- > 50 millions d'enregistrements
- Anonymisation menée à bien en moins d'une semaine

* Source : Cortina

Les principaux écueils

- Facilité apparente
- Tout baser sur la technique (« panacée »)
- Manque de maîtrise :
  - Exemple 1 : Les personnes qui ont conçu et qui maîtrisaient la logique de certaines applications sont parties à la retraite – sans avoir documenté leurs développements… certaines données sont « codées » et le personnel en place ne maîtrise pas totalement la signification de chaque champ
  - Exemple 2 : Organisme ne maîtrisant la structure de la base de données
- Prévoir la conduite du changement

Conditions du succès

- Préparation et Gestion de projet
- Bonne connaissance de l'existant
- Imaginer le processus futur (avoir une idée claire de l'objectif)
- Mobilisation : Tous les acteurs sont concernés (sécurité, production, études-développement, systèmes, DBA, etc.)
- Avoir un « Champion », un pilote de projet
- Bien valider la recette : une erreur provient-telle de l'application ou bien du processus d'anonymisation ?

L'apport du CIL*

- L'anonymisation permet de « sortir » du périmètre (et donc d'éviter une demande d'autorisation, avec les risques de refus associés et les délais d'attente)
- L'impliquer en amont, en appui du chef de projet
- Interface CNIL

cf. document « *Les questions que doit se poser un Correspondant Informatique & Libertés dans le cadre d'un projet d'anonymisation* » www.afcdp.org)


AFCDP

LISTE DES QUESTIONS QUE DOIT SE POSER LE CIL
POUR UN PROJET D'ANONYMISATION DE DONNEES

Dans le cadre de ses travaux sur le thème de l'anonymisation de données, le groupe *Aplitweiki et Lobeli* de l'AFCDP suggère une liste de points que le Correspondant Informatique & Libertés doit aborder et valider dans le cadre d'un projet d'anonymisation de données.

Le Correspondant Informatique & Libertés peut apporter son aide aussi bien en amont du projet (avis sur le périmètre et sur le processus envisagé, participation au choix de l'outil, recommandations quant à sa mise en œuvre, sensibilisation des personnels, etc.) qu'en aval (contrôles réguliers, vérification du respect des procédures, propositions d'améliorations, etc.). Naturellement, cette démarche s'effectue en pleine coopération avec les fonctions concernées par le projet ; responsable des développements, RSSI, Risk manager, responsable informatique, etc.

Il concourre ainsi à la protection des données personnelles et à la réduction effective du risque qui pèsent sur l'entité et le responsable du traitement.

Les membres du groupe doivent être remerciés pour le travail accompli :
- Eric Barbry, Alain Bensoussan Avocats
- Yves le Hégarat, «SELP»
- Bernard Lombarde, Ugap
- Bruno Rasle, Cortina
- Alain Rouffiat, Princeton Softech
- Michel Simion, Compuware
- Gilles Trouessin, Oppida

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

On pourra se reporter avec profit au Glossaire « Anonymisation des Données » édité par le même groupe de travail

Arnaud Belleil
Cecurity.com
Administrateur AFCDP

23 mai 2007 - Page 1 sur 3

* Correspondant Informatique & Libertés

Sommaire

- **La place de l'anonymat dans l'univers de la confiance**
- **Retours d'expérience sur les projets d'anonymisation**
- **Du lexique de l'anonymisation au Référentiel**
- **Robustesse et des-anonymisation**
- **Vers une labellisation des technologies d'anonymisation ?**

Les techniques

- **Suppression ou masquage**
  - Total : 01.22.00.67.99 → XX.XX.XX.XX.XX
  - Partiel (floutage) 01.22.00.67.99 → 01.22.00.XX.XX*
- Par **ajout** d'éléments (créer des personnes inexistantes, par exemple)
- **Remplacement**
  - Substitution par des données fictives
    - préservant l'unicité Gilles → René, Pierre → Georges, Gilles → René
    - sans préservation de l'unicité Gilles → René, Pierre → Georges, Gilles → Paul
  - Translation (fixe – ex. Vieillissement - ou variable)
  - Chiffrement

* Peut s'ajouter à la substitution 01.22.00.67.99 → 01.43.26.XX.XX

– **Remplacement** (suite)
  - « Hachage »
  - Génération de données fictives
  - Remplacement par des données aléatoires
  - Concaténation : remplacement par une valeur issue de la combinaison de plusieurs champs figurant dans la source
  - Mélange : les données sont « brassées » sans être modifiées

Choisir la technique adéquate

- Unicité
- Lisibilité
  - Applicative
  - Humaine
- Réversibilité*
- Répétitivité

* Donc pseudonymat plutôt qu'anonymat en cas de réversibilité

Les cas difficiles

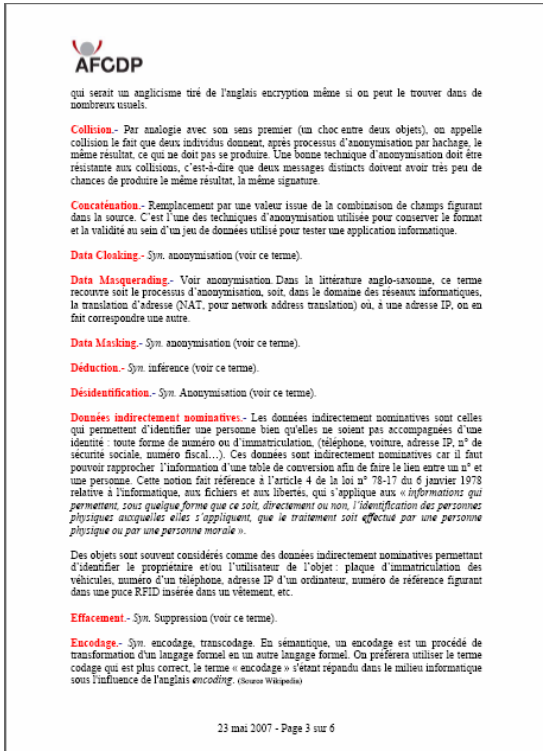
- Les données signifiantes (ex. NIR, data avec checksum)
- Les zones de texte libre (ex. Commentaires)
- « *Isolated Case Phenomena* » : PDG identifié par le salaire le plus élevé ?
- Séquencement des opérations d'anonymisation
- Cas très spécifiques : ex. Un employé a quitté l'entreprise, puis est à nouveau embauché

Outils de génération de jeux banalisés

- Développements spécifiques (notamment dans le secteur santé)
- Outils libres (ex : anonymisation des CV)
- Outils/offres commerciales
  - Compuware
  - Cortina
  - IBM (Preston Softech)

Glossaire

- Lexique en ligne www.afcdp.org
- Quelques exemples :
  - Pseudonymat
  - Shuffle
  - Inférence
  - Appauvrissement
  - Obfuscation
- Défini « perfectible »



AFCDP

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 préfèrera 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)

23 mai 2007 - Page 3 sur 6

Sommaire

- **La place de l'anonymat dans l'univers de la confiance**
- **Retours d'expérience sur les projets d'anonymisation**
- **Du lexique de l'anonymisation au Référentiel**
- **Robustesse et des-anonymisation**
- **Vers une labellisation des technologies d'anonymisation ?**

- La *robustesse* d'un système d'anonymisation est constituée de l'ensemble des caractéristiques à satisfaire pour éviter la levée de l'anonymat de façon non-autorisée.
- La CNIL n'impose pas de méthode, **seul le résultat compte**
- Quels sont les critères qui jouent sur la robustesse ?
  - La taille du corpus ? (Dix mille entrées ou deux ?)
  - Informations sur la nature du jeu (« *Il s'agit d'un fichier paye récent de la ste X* »)
  - L'outil utilisé ?
  - Les techniques retenues par l'administrateur de l'outil ?
  - La démarche globale ?
- Quelle robustesse doit-on avoir et vis-à-vis de quoi ?

« dés-anonymisation »

- Anonyme + Données personnelles
  - Le fichier X est anonymisé, le fichier Y ne l'est pas
  - Le croisement du fichier X avec un fichier Y lève l'anonymat des données de X
- Anonyme + Anonyme
  - Les données A et B sont (semblent ?) anonymes
  - La combinaison des données A et B devient indirectement nominative

Rappel : L'interconnexion de fichier est un traitement, aux yeux de la CNIL

Anonyme + Anonyme = Identification*

Il y a eu 762 407 naissances en France en 1990
Il y a eu 45 sorties de l'hôpital le 12/12/2006

Une seule personne, née en 1990, est sortie de l'hôpital le 12/12/2006
(Source : présentation CNIL)

87 % de la population des Etats-Unis (soit 216 millions sur 248) peuvent être identifiés simplement à partir du code ZIP, combiné au sexe et date de naissance.
(Source : Bruce Schneier, "Why 'Anonymous' Data Sometimes Isn't", 13 décembre 2007)

Des ratés…

- Un moteur de recherche américain propose des données de recherche anonymes à des chercheurs pour diverses analyses
- Le moteur de recherche enregistre l'adresse IP et les mots clés recherchés pour chaque requête
- Solution adoptée : L'adresse IP est hachée
- Problème imprévu :
  - Les mots clés portent souvent sur des proches, des centres d'intérêt et toute sorte d'informations, qui corrélées … deviennent personnelles
  - Des journalistes découvrent l'identité de plusieurs personnes et publient des extraits de requêtes d'utilisateurs (ex. « *how to kill your wife*», « *pictures of dead people* », …).

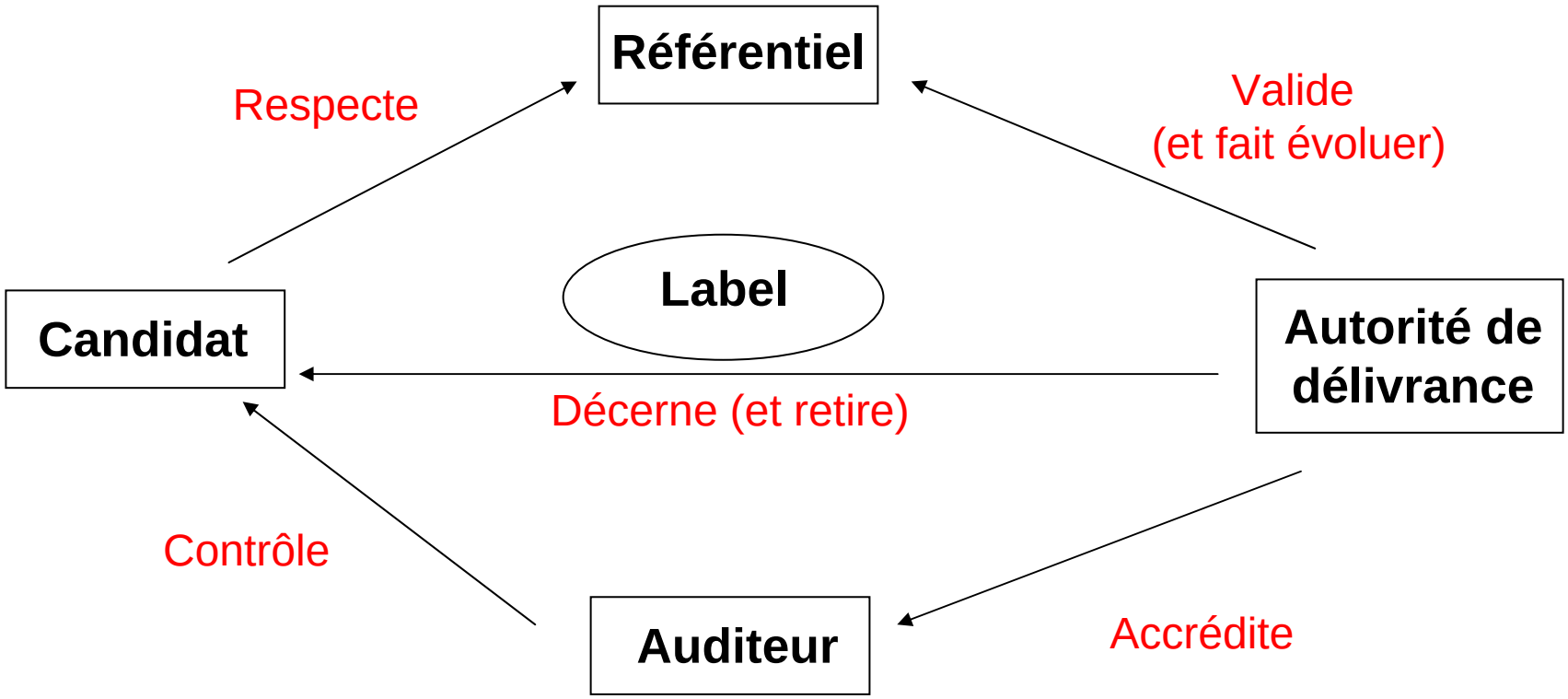
Sommaire

- **La place de l'anonymat dans l'univers de la confiance**
- **Retours d'expérience sur les projets d'anonymisation**
- **Du lexique de l'anonymisation au Référentiel**
- **Robustesse et des-anonymisation**
- **Vers une labellisation des technologies d'anonymisation ?**

Un label pourquoi faire ?

- Permettre de distinguer ce qui est de l'anonymisation et ce qui n'en est pas
  - D'où le travail préalable sur le glossaire
- Définir un niveau minimal d'exigence
  - Les niveaux élevés font partie des valeurs ajoutées, du domaine de la concurrence
- Veiller à l'homogénéité du niveau de performance & de sécurité
  - Prendre en compte tous les maillons de la chaîne
- Faciliter la tâche des acheteurs
  - *Sourcing*, constitution d'une *short list*
- Faciliter la tâche des vendeurs
  - Avant le label : amélioration de leur offre
  - Après le label : confiance et différenciation

Référentiel, label, auditeurs



Les objectifs du référentiel AFCDP

- Favoriser la diffusion des technologies de protection des données à caractère personnel
  - Domaine d'application : la sécurisation des données de tests et de maintenance par l'anonymisation dans un contexte d'externalisation croissante, voire d'*offshoring*
- Simplicité
  - Pédagogie de l'anonymisation
- Neutralité
  - Accessibles aux progiciels, services ou développements internes
- Accessibilité
  - Réduire le coût d'accès à un éventuel label associé en facilitant par construction la tâche des auditeurs
- Ouverture
  - Publication et appel à amélioration, démarche non propriétaire

La sécurisation des données de test

- Les données réelles sont utilisés pour les développements (64% des cas), les tests (44%) et la maintenance (34%)
- Parmi les entreprises qui utilisent des données réelles, 66% le font avec les données relatives aux clients
- Seules 5% des entreprises utilisent des dispositifs d'anonymisation

Source : « Insécurité liée aux données de test : un danger latent », Étude Ponemon Institute / Compuware – 31 janvier 2008

Le contenu du référentiel AFCDP

- Les domaines couverts
  - Robustesse
  - Traitement des collisions et homonymies
  - Variété des méthodes d'anonymisation
  - Association des méthodes d'anonymisation
  - Enregistrement des méthodes d'anonymisation
  - Librairie de données fictives
  - Documentation
- Première diffusion publique à l'occasion de la journée OSSIR 2008

Vers un label CNIL ?

- La labellisation des produits par la CNIL, innovation (encore méconnue) de la loi d'août 2004

Article 11

3° A la demande d'organisations professionnelles ou d'institutions regroupant principalement des responsables de traitements :

a) Elle [la CNIL] donne un avis sur la conformité aux dispositions de la présente loi des projets de règles professionnelles et des produits et procédures tendant à la protection des personnes à l'égard du traitement de données à caractère personnel, ou à l'anonymisation de ces données, qui lui sont soumis ;

c) Elle délivre un label à des produits ou à des procédures tendant à la protection des personnes à l'égard du traitement des données à caractère personnel, après qu'elles les a reconnus conformes aux dispositions de la présente loi ;

- Le modèle *Privacy Seal* de l'autorité de protection des données du Schwelsig Holstein
- Dans l'attente du décret d'application ….

Vers un label CNIL ?

- La CNIL est l'autorité de délivrance la plus légitime pour la labellisation des produits d'anonymisation
- Le référentiel de l'AFCDP et ses prochaines versions sont à la disposition de la CNIL qui peut s'en inspirer librement
- Dans l'attente du décret d'application …. Il semble déjà possible d'entamer les échanges avec la CNIL sur un avis de conformité