Aspects pratiques du traitement de données personnelles

Thomas Soubiran
Ceraps (UMR 8026 CNRS - Université de Lille)

Formation données personnelles

Lyon, 26 novembre 2018

L'application de la réglementation

L'application de la réglementation est **constituée** de :

1) de principes à respecter :
   - limitation de la finalité
   - minimisation des données
   - limitation de la conservation
   - information des personnes
   - protection dès la conception
   - ...
2) de **démarches** à réaliser auprès du représentant du responsable de traitement, le délégué à la protection des données (DPD)
3) de **mesures de protection** des données à mettre en œuvre

La présentation portera spécifiquement sur les points 2) et 3) en référence au 1)

Au menu

Présentation en deux parties :

- **Les démarches IL**
  - préparation des démarches
  - réalisation des démarches

  relation avec le DPD, description du traitement, identification des parties impliquées, conventionnement,. . .

- **La protection des données**
  - protection des systèmes d'information
  - protection contre la réidentification

  cryptographie, pseudonymisation, algorithmes de (dé|ré)identification des personnes,. . .

Les démarches IL

La préparation des démarches

Enter The DPO

- avec l'entrée en application du RGPD, le **DPD** (délégué à la protection des données) succède au **CIL** (correspondant informatique et libertés)
- sa désignation est **obligatoire** notamment lorsque :

  *le traitement est effectué par **une autorité publique** ou **un organisme public** (à l'exception des juridictions agissant dans l'exercice de leur fonction juridictionnelle)* (**RGPD art. 37 § 1**)

- cette disposition concerne **aussi** les équipes de recherche des EPSCP, EPST,...
- en conséquence,
  - les traitements de DCP qui y sont réalisés doivent faire l'objet d'un **enregistrement** dans le registre du DPD

    enquêtes, inscriptions aux événements organisés, fonctionnement des équipes, ...

  - et ce **avant** le début de chaque traitement

Note : les modalités des enregistrements varient en fonction du type de traitement, certains pouvant être enregistrés une fois pour toute (cf. normes simplifiées)

Le DPD dans vos projets de recherche

Le DPD doit donc être associé **systématiquement** à **tous** vos traitement de DCP

- et cela, dès **la conception du projet**

Car, **au-delà** de l'inscription au registre

- l'application de la réglementation peut en effet **impacter** tous les aspects de vos investiguations :
  - **ce que vous pouvez collecter**
  - mais aussi **la façon** dont vous pouvez le collecter et le traiter
  - donc, plus vous tardez, plus les choses risquent de se compliquer
- de plus,
  - la réglementation est avant tout constituée **de principes généraux**

    dont les implications pratiques ne se donnent souvent pas de façon évidente
  - le RGPD **renforce** considérablement les obligations du RdT
  - en marquant le passage à un régime dit **de responsabilisation**
  - qui se traduit par **l'inversion** de la charge de la preuve

    désormais, c'est au RdT de prouver qu'il est en conformité avec la réglementation

Le DPD dans vos projets de recherche

la première action consiste donc
À PRENDRE CONTACT AVEC SON DPD

- et cela, dès **la conception du projet** (bis)
- après l'avoir **désigné officiellement** si ce cela n'a pas déjà été fait

la désignation du DPD doit être enregistrée auprès de la CNIL pour être valide

L'enquête en questions

Pour entreprendre des démarches IL, il faut être en mesure de répondre précisément aux **questions suivantes** :

- **qui :** quel(s) est|sont le(s) **responsable(s) de traitement** (RdT), les **destinataires de données**
- **quoi : quels** renseignements seront collectés et **auprès de qui**
- **pourquoi :** quelles sont les **finalités** (*modus essendi*)
- **quand, où, comment :** quelles sont **modalités** de collectes (*modus oprandi*)
- **pendant combien de temps :** limitation de **la durée de conservation** des données

Avec une **question subsidiaire :** quels sont les **effets** que le traitement de DCP peut avoir sur les personnes concernées

L'enquête en questions

Autrement dit, il faut être **au clair** sur :

- **la finalité :** la problématique précise, la population enquêtée
- **les moyens de la collecte :** entretiens, questionnaires, aspirations de données,. . .

  et fournir tous les éléments correspondants : grille d'entretien, questionnaires,. . . et pouvoir justifier de leur proportionnalité et de leur pertinence

- ainsi que les éventuels **transferts** et **croisement** de données
- mais aussi avoir **identifié** :
  - le(s) responsable(s) de traitement

    notamment pour déterminer le DPD compétent

  - les destinataires de données
  - les partenaires
  - sous-traitants
- et réaliser **une étude d'impact**

  publication, rediffusion de base de données,. . .

Le responsable de traitement

- la règle déterminant le RdT pour les UMR est le produit d'**une (longue) négociation** entre le CNIL et les instances représentant l'ESR
- elle **n'apparaît pas** dans la LIL ou le RGPD
- dans les situations qui n'ont pas fait l'objet de décision, la détermination du RdT n'est **pas toujours aisée**

> *multiplicité des tutelles, caractère informel de nombreux dispositifs et activités réalisées dan dans le cadre de l'ESR*

Conventionnement

- si **plusieurs responsables de traitement** déterminent conjointement les finalités et les moyens du traitement (p. ex. dans le cas d'un projet de recherche associant plusieurs entités) :
- ils sont les responsables **conjoints** du traitement (**RGPD art. 26 § 1**)
- les responsables conjoints du traitement définissent de manière transparente **leurs obligations respectives** (*ibid*)
- par une **convention de recherche**

Note : la convention constitue aussi **une protection** en cas de dissensions au sein du projet

Conventionnement

- le conventionnement peut aussi se révéler nécessaire dans la relation **au terrain d'enquête** :
  - il peut arriver que des fichiers soient **transmis** par des institutions mais aussi que des fichiers leurs soient transmis (éventuellement en retour)
  - là encore, ces transferts doivent faire l'objet d'une convention et, le cas échéant d'une information des personnes concernées par les DCP transmises
- mais aussi dans les situations de **sous-traitance**

  Exemple : la transcription d'entretiens
  nécessite l'ajout d'une annexe au contrat de travail stipulant les obligations du prestataire quant aux traitement de DCP

Aparté : Propriété des données

- le RdT est parfois vu comme **le « propriétaire des données »**
- or, le cadre applicable aux DCP ne dit **rien** sur la propriété des données
- et les deux notions sont **distinctes**
- même si elles peuvent **coïncider** en pratique

  Note : les deux notions coïncident rarement dans l'ESR qui est avant tout composée d'agents publiques

- toutefois, la convention peut constituer, le cas échéant, l'occasion de régler les questions de propriété

Destinataires de données

- **destinataires de données :** « toute personne habilitée **à recevoir communication** de ces données autres que la personne concernée, le responsable du traitement, le sous-traitant et les personnes qui, en raison de leurs fonctions, sont chargées de traiter les données » (LIL art. 3)
- liste des personnes qui auront **accès aux données** collectées

  Exemple : les membres d'un projet de recherche

- cette liste doit figurer dans **l'information** faite aux personnes sur les traitement de DCP les concernant

La réalisation des démarches

The Art of The Deal

- la réalisation des démarches IL nécessite de de saisir **les implications pratiques**
- **des notions et principes** constitutifs de la réglementation sur le traitement
- pour prendre **les mesures adéquates** pour assurer la protection des données
- mais aussi afin de pouvoir **argumenter**
- notamment pour **justifier** :
  - la finalité du traitement
  - ainsi que sa proportionnalité et sa pertinence

Le traitement

- sans que les termes soient pour autant traités de façon identique, la distinction **« quali »-« quanti »** n'est pas aussi structurante (et clivante)

  la question est d'abord de savoir quelles informations vont être collectées

- le traitement est **un tout** :
  - pas de distinction entre **collecte**, **stockage**, **analyse** ou encore **publication** : toutes ces opérations font parties du traitement
  - le fait que les données à caractère personnel collectées ne soient **pas utilisées** du tout ou seulement dans une phase du traitement comme l'analyse ne change rien (puisque le stockage est un traitement)
  - pas plus que le **nombre** de personnes identifiables

- **Exemple :** l'analyse de questionnaires
  - le fait que l'analyse de données d'enquêtes par questionnaires soit le plus souvent anonyme **ne change rien**
  - et cela même si les données à caractère personnel ne sont utilisées que pour la collecte et ne sont **jamais croisées** avec les réponses

  cf. la présentation ***Protection des données à caractère personnel et qualité des enquêtes statistiques*** à la journée CJADCP pour une proposition de **« méthodologie de référence »** dans ce cas précis (Soubiran, 2017)

Principes de proportionnalité et pertinence

- avoir de (bonnes) raisons (clairement définies) de collecter des données **ne suffit pas**
- ***The Name of the Game :*** vous faire collecter **le moins d'informations possible** (minimisation des données)
- en pratique, un des aspects **les plus délicat** de l'application de la réglementation en sciences sociales :
  - la finalité n'est pas toujours facile à établir précisément **au préalable** et donc ce qui est strictement nécessaire à la finalité
  - dépasse l'aspect **procédural**
  - peut toucher **au contenu** des recherches elle-mêmes
  - particulièrement lors de la collecte de **données sensibles**

La minimisation des données

Exemple : la limitation **du croisement des données**

- ne se limite pas aux croisement de source (p. ex. des bases des données)
- et peut conduire à **un cloisonnement thématique**
- **cas pratique (tiré d'un cas concret) :** enquêtes par questionnaire sur **les déplacements**
  - l'application stricte du principe de minimisation impliquerait de ne collecter des renseignements **exclusivement sur les déplacements** (fréquence, modes de transports,. . .)
  - et exclurait donc la collecte d'autres informations comme, p. ex., la composition du ménage
  - néanmoins, on peut ici arguer que, p. ex., **les caractéristiques du ménage** (sa composition, ses revenus,. . .) ont un effet sur les déplacements **pour établir la proportionnalité et la pertinence** de la collecte d'information sur le ménage et les individus qui le compose relativement à la finalité
- **autre cas :** les indicateurs

  Exemple : propriété du logement, équipements du ménage (réfrigérateur, bibliothèque),. . .

La minimisation des données

Cas pratique (plus délicat) : la religion

- là aussi, l'application stricte du principe de minimisation impliquerait que l'on ne puisse poser **des questions relatives aux pratiques religieuses** des individus que dans le cadre **d'enquêtes sur les pratiques religieuses**
- or, d'un point de vue sociologique, la religion apparaît comme un **fait social total** et touche donc à **de nombreux autres domaines** comme la fécondité, l'éducation, les consommations, la participation politique et associative...
- ainsi, l'étude de la religion implique souvent de s'intéresser à **d'autres pratiques** et, réciproquement, l'études de certaines pratiques nécessite parfois l'intégration de **la dimension religieuse**

La minimisation des données

Problèmes :

- tout ce qui a trait à la religion est considéré comme une **donnée sensible**
- encore mieux (ou pire) : la réalisation de la finalité nécessite de croiser pratiques religieuses et pratiques politiques (**autres données sensibles**)

Toutefois,

- dans ce cas particulier, on ne peut que se féliciter de ce que **G. Michelat et M. Simon** aient réalisé leurs enquêtes AVANT le vote de la LIL et permettent d'étayer la proportionnalité et la pertinence de la collecte et du traitement de données liant pratiques politiques et religieuses
- préparez-vous néanmoins à devoir batailler...

Fazit

La finalité des traitements (et surtout leur indétermination) peut **parfois** causer des difficultés dans les démarches relatives aux données à caractère personnel :

- il ne s'agit cependant pas du point le plus problématique
- sous condition que vos interlocuteurs aient une **familiarité suffisante** avec les enquêtes en sciences sociales

Mais, en règle générale,

la proportionnalité et la pertinence de la collecte constituent un des principaux points d'achoppement dans l'application de la réglementation relative aux DCP en sciences sociales

et ce, particulièrement lorsque la finalité implique la collecte et, *a fortiori*, le croisement **de données sensibles**

Note : il est important de souligner que ce n'est pas toujours le cas et que la proportionnalité et la pertinence des traitements peuvent être établis dans de très nombreuses situations

Et après ?

- par défaut, les DCP ne peuvent être conservées que **le temps nécessaire** à la réalisation de la finalité

  dans un état permettant l'identification des personnes

- les données doivent ensuite être ensuite **détruites** ou **anonymisées**
- la conservation est toutefois possible dans le cadre d'**archives publiques**
- pour la réutilisation, c'est plus compliqué...

La protection des données

La protection des données à caractère personnel

La protection des données à caractère personnel peut être déclinée selon **deux aspects** (liés) :

- **la protection des systèmes d'information**

  protection physique ou logicielle contre les accès non autorisés aux données

- **la protection contre la réidentification des personnes**

  concerne les données elles-mêmes

- lors **des différentes étapes** du traitement

  collecte, conservation, analyse ou (re)diffusion

La protection des données à caractère personnel

- importance de **la sécurisation** des données collectées, particulièrement lors de la collecte **de données sensibles**
- exemples de mesures prescrites par le RGPD :
  - **minimisation**, **anonymisation**
  - **la pseudonymisation** et **le chiffrement** des données à caractère personnel (RGPD art. 32 § 1 (a))
- ainsi que :
  - des moyens permettant de garantir **la confidentialité**, l'intégrité, la disponibilité et la résilience constantes des systèmes et des services de traitement (RGPD art. 32 § 1 (b))
  - une procédure visant à tester, **à analyser et à évaluer** régulièrement **l'efficacité** des mesures techniques et organisationnelles pour assurer la sécurité du traitement (RGPD art. 32 § 1 (d))
  - **notification**, dans les 72h, des incidents de sécurité (« violation de données à caractère personnel ») à l'autorité de contrôle ainsi qu'aux personnes concernées (RGPD art. 33 et art. 34)
- **rappel :** la protection des données est **la responsabilité** du RdT

La sécurisation des données

La sécurisation des données

Sujet très vaste, les mesures à prendre dépendent du type de données , de leur mode de collecte, du contexte de leur utilisation, des risques,. . .

- *a minima*, recourir au **chiffrement** systématique des ressources
- chiffrement **des périphériques** de stockage (chiffrement par blocs) :
  - partitions, `DD` externe, clefs `USB`,. . .
  - soit en utilisant des logiciels proposés par les systèmes d'exploitation : `dm-crypt` sous Linux, `Bitlocker` sous Windows ou `FileVault` sous Mac OS X
  - soit en utilisant des logiciels portables comme `VeraCrypt` (*fork* de `TrueCrypt`)
- chiffrement des **transferts** de données (chiffrement asymétrique) : `GnuGPG`

Note : la meilleure sécurité est évidemment de ne disposer d'aucune données à caractère personnel ou de s'en débarrasser (moins de données à caractère personnel, moins de contraintes)

La cryptographie asymétrique

- distinction entre chiffrement **symétrique** et **asymétrique** :
  - **symétrique :** une seule clef k sert à chiffrer et déchiffrer le message m

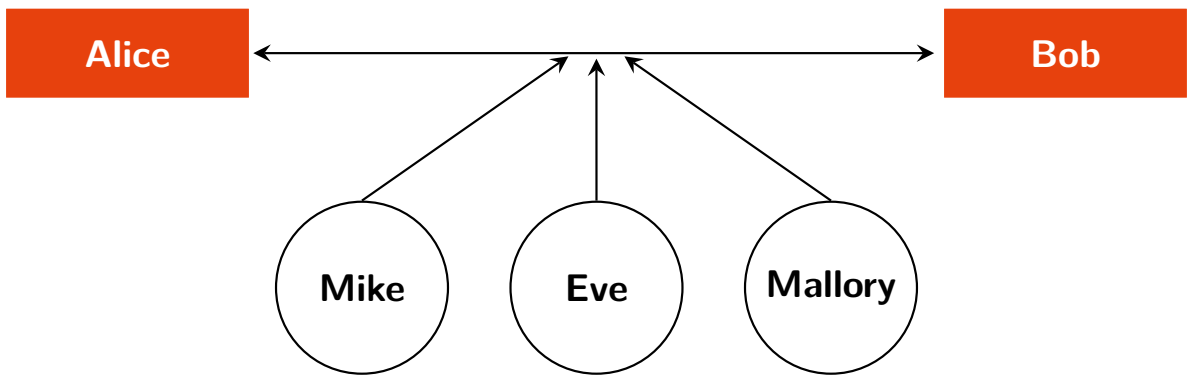
    Exemple : le chiffre de César
    décalage du numéro d'ordre des lettres de l'alphabet, k correspondant au décalage
  - **asymétrique :** on génère deux clefs
    - une clef publique k_{pub} qui sert **à chiffrer** m
    - une clef privée k_{priv} qui sert **à déchiffrer** m
    - k_{pub} peut être diffusée sans restriction alors que k_{priv} doit restée cachée
- Exemple d'alogrithmes de chiffrement symétrique :

  `AES (Rijndael)`, `Blowfish`, `Twofish`,...
- Exemple d'alogrithmes de chiffrement asymétrique :

  `RSA`, `Diffie-Hellman`,...

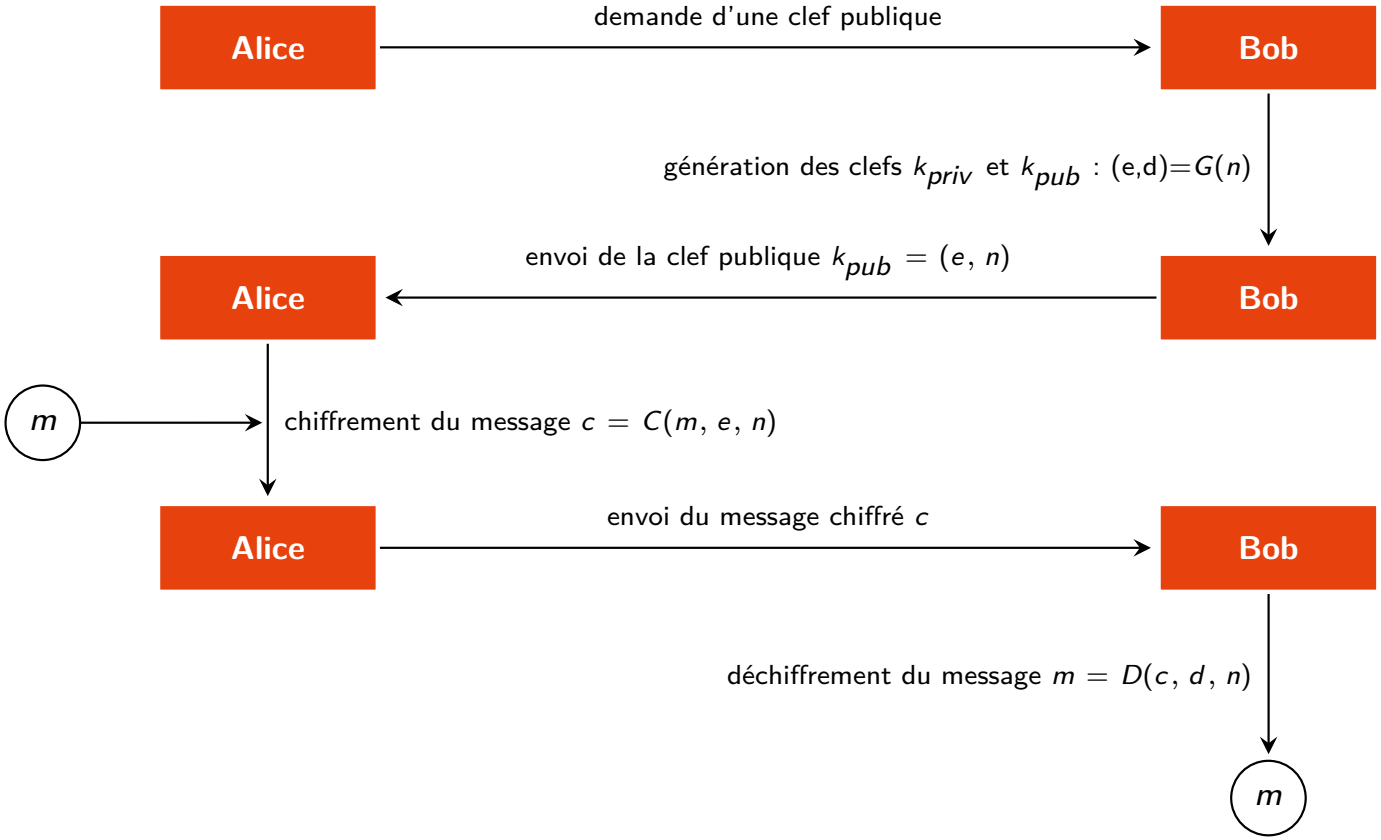
Alice, Bob, Eve et les autres

- A(lice) et B(ob) sont deux personnages fictifs souvent utilisés en cryptographie
- Alice et Bob veulent communiquer sans que Eve (*the eavesdropper*), Mike (*the microphone*), Mallory (*malicious*),... puissent connaître le contenu de leurs échanges



Chiffrement asymétrique : RSA

Alice veut écrire à Bob :



Note : pour que Bob puisse écrire à Alice, il faut réaliser l'opération inverse

Cryptographie sommaire

- la cryptographie repose sur des fonctions **inversibles**

$$c = C(m, k)$$
$$m = D(c, k) = D(C(m, k), k) = C^{-1}(c, k)$$

- exemple de fonction inversible, **l'élévation à la puissance** :

$$y = x^2 \quad x = \sqrt{y}$$

- la racine carrée peut aussi s'écrire

$$x = y^{\frac{1}{2}}$$

- en effet,

$$(x^2)^{\frac{1}{2}} = x^{\frac{1}{2}2} = x^1 = x$$

Application

- `RSA` : chiffrement

$$c \equiv m^e \, mod \; n$$

- `RSA` : déchiffrement

$$m \equiv c^d \, mod \; n \equiv (m^e)^d \, mod \; n \equiv m^{ed} \, mod \; n$$

- Notes :
  - les opération sont réalisées modulo n avec n premier
  - la cryptographie repose sur l'arithmétique modulaire, -ie : l'arithmétique des horloges

    161 $mn = 2h$ 41 $mn = 41 + 2 * 60$
  - l'arithmétique modulaire permet de définir un type de nombres particulier (les corps finis)
  - dont les propriétés sont utiles à la cryptographie

    Note : le chiffre de César est un exemple d'application de l'arithmétique modulaire en cryptographie

Pour aller plus loin...

- **thèorème (Euler) :** $m^{\phi(n)} \equiv 1 \ mod \ n$

$$m^{k\phi(n)+1} \equiv m^{k\phi(n)} m \equiv 1^k m \equiv m \ mod \ n$$

où :

- m et n sont coprimes
- $\phi(n)$ est la fonction de totient d'Euler (qui compte le nombre d'entiers premiers inférieurs à n)
- pour `RSA`, $n = pq$ et $\phi(n) = (p-1)(q-1)$

  avec p et q deux nombres premiers générés aléatoirement

- e est choisi de façon à ce que $1 < e < \phi(n)$
- d est choisi de façon à ce que (inverse modulaire) :

$$d = e^{-1} \ mod \ \phi(n)$$
$$ed = 1 \ mod \ \phi(n)$$
$$ed = 1 + k\phi(n)$$

La sécurité RSA

- n est le produit de **deux nombres premiers** p et q générés aléatoirement
- la sécurité RSA repose sur le fait que, pour déterminer d, il faut **disposer** de p et de q
- cette opération est possible mais **quasi-irréalisable** dans les faits car elle nécessite de factoriser n (pour n suffisamment grand)

  avec un module de 1024 bits, le temps nécessaire à l'opération a été estimé à 2 000 ans en mobilisant plusieurs centaines de machines

- plus généralement, la cryptographie repose sur des problèmes mathématiques dont la solution est **difficile à obtenir**

Notes :

- d'autres attaques contre RSA sont possibles, p. ex. en réalisant une cryptanalyse acoustique
- un entier peut être facilement factorisé avec un ordinateur quantique

Cryptographie hybride

- `RSA` permet la diffusion des clefs de chiffrement sans restrictions
- toutefois, il n'est pas adapté pour le chiffrement **de gros volume de données**
  - la taille maximale du message (en `bits`) doit être **inférieure** à la taille du module
  - l'exponentiation modulaire utilisée pour le chiffrement et le déchiffrement est une opération **coûteuse** en temps `CPU`
- c'est pourquoi il est souvent utilisé **conjointement** à des algorithmes de chiffrement symétriques
- on parle alors de **cryptographie hybride**
  - la clef publique k_{pub} permet de chiffrer la clef utilisée pour chiffrer les données
  - **Exemple :** les protocoles `TLS` (*Transport Layer Security*) pour chiffrer des connexions réseau ou des logiciels comme `PGP` ou `GnuGPG` pour chiffrer des documents

La pseudonymisation

La pseudonymisation

pseudonymisation : le traitement de données à caractère personnel de telle façon que celles-ci (RGPD art. 4 § 5) :

- **ne puissent plus être attribuées** à une personne concernée précise
- **sans avoir recours à des informations supplémentaires**, pour autant que ces informations supplémentaires **soient conservées séparément** et soumises à des mesures techniques et organisationnelles
- afin de garantir que les données à caractère personnel **ne sont pas attribuées à une personne physique identifiée ou identifiable**

Lorsque le traitement ne peut être anonymisé, le RGPD prescrit notamment le recours à la **pseudonymisation** :

- consiste à remplacer **des données directement identifiantes** (noms, lieux, codes...) par un **identifiant**
- pour qu'il soit impossible de remonter à la personne concernée, cet identifiant ne doit **avoir aucun lien** avec les caractéristiques de cette personne
- **Exemples :**
  - génération d'un nouvel identifiant
  - la CNIL recommande le hachage des données identifiantes avec une fonction cryptographique à clef secrète comme `HMAC`

La pseudonymisation

- la pseudonymisation est **réversible**, p. ex. en utilisant la mappe (table de correspondances) entre l'identifiant original et le l'identifiant public
- mais seulement par les personnes **habilitées à le faire**
- la pseudonymisation est une notion différente de **l'anonymisation** qui ne permet plus la réidentification de façon **irréversible**

  Note : du point de vue de la réglementation, la proposition « mes données sont anonymes parce que j'ai remplacé les noms par des pseudonymes » est fausse

- la pseudonymisation, telle que définie dans le RGPD, diffère aussi de la pseudonymisation telle que pratiquée, p. ex., pour **la citation d'entretiens** en sciences sociales

La pseudonymisation

- la définition de la pseudonymisation renvoie implicitement au traitement de données à caractère personnel conservées dans **des bases de données**

  *elle consiste principalement à remplacer les **clefs primaires** de la base*

- sa mise en œuvre dans d'autres contextes (entretiens, archives, . . .) est clairement **plus délicate**

  nécessite au préalable une analyse morpho-syntaxique

- la pseudonymisation n'est **pas toujours suffisante** pour prévenir la réidentification

  - la pseudonymisation **ne supprime pas** toutes les données indirectement identifiantes
  - la réidentification peut demeurer possible par **croisements**

Exemple : l'enquête MILITENS

- **MILITENS :** enquête par questionnaires en ligne sur les enseignants des premier et second degrés à partir d'un échantillon national aléatoire stratifié tiré de la base de sondage de la DEPP
- qui a d'abord fait l'objet d'**une convention** avec la DEPP
- le transfert et la conservation **des informations** de contact sur des supports chiffrés
- gestion des invitations **distinctes** de la gestion des réponses (pas d'informations de contact stockées sur le même serveur que le gestionnaire d'enquête)
- conservation **des réponses et des traitement** sur un support chiffré
- **diffusion** des données auprès des membres du projet :
  - la cryptographie asymétrique
  - agrégation des données potentiellement indirectement identifiantes issues de sources externes à l'enquête (taille de l'établissement, informations sur le quartier issues du recensement)

Pseudonymisation des questionnaires en ligne

La séparation de l'envoi des invitations et des réponses à un questionnaire en ligne :

- exemple d'application de **la pseudonymisation**
- différents gestionnaires de questionnaire peuvent aussi assurer **l'envoi des invitations**
- ils doivent donc avoir accès à des données à caractère personnel comme **l'adresse des répondants**
- si la sécurité de l'application (ou du serveur) est **compromise**, ces données peuvent fuiter
- pour assurer la confidentialité des données (particulièrement lors de la collecte de données sensibles), il est préférable **de séparer** l'envoi des invitations de la gestion des réponses au questionnaire
- ainsi, les données à caractère personnel peuvent être remplacées par un identifiant permettant de faire le lien entre (non-)réponses et données auxiliaires

Pseudonymisation des questionnaires en ligne

En pratique,

- il faut générer **deux clefs** :
  - une clef privée pour les données auxiliaires
  - une clef publique pour les traitements (au cas où les données auxiliaires seraient aussi compromises)

priv	pub
12144	04835
09718	02359
11259	10230
09734	11470
12162	01123
...	...

- la table permettant la mappe entre les deux doit être stockée à part

Note : par précaution, si vous attribuez un numéro pour identifier les individus, il est préférable de réaliser **une permutation** $\sigma(\#\texttt{oid})$ avant l'attribution (sinon le nombre correspondra à la ligne et l'ordre permettra la réidentification)

La « pseudonymisation » des entretiens

- l'usage de « pseudonymes » s'est progressivement répandu **pour désigner les personnes** mentionnées dans des publications
  - leur choix n'est toutefois **pas aléatoire**
  - et dépend souvent de ce que le prénom connote (par rapport au sexe, à l'âge, . . .) à propos de la personne mentionnées (Coulmont, 2017)
  - répondant ainsi à une recherche **« d'équivalence »** sur un ou plusieurs critères
- ce faisant, les « pseudonymes » contiennent des informations pouvant concourir à **la réidentification des personnes** mentionnées
- et ne sont donc **pas conformes** à la réglementation
  - d'autant plus si on utilise une `API` publique pour la construction des classes d'équivalence de prénoms
  - cette approche permet en effet de faciliter **la reconstitution de l'éventail de prénoms** dont est issu le pseudonyme

    la fonction est certes surjective mais elle est facilement invertible et la taille de l'ensemble de départ est de plus réduite
- de plus, cette approche **ne garantit en rien** la confidentialité des données

La « pseudonymisation » des entretiens

- **en soi**, les « pseudonymes » ne sont pas identifiants
- toutefois,
  - les prénoms ne sont **pas les seules informations** sur lesquelles un attaquant peut s'appuyer
  - les publications recèlent généralement de nombreuses informations relatives aux personnes

    lieux, habitudes, événements,. . .
  - l'identification peut donc se faire **par recoupements** en conjonction avec ces « pseudonymes »
- **l'incertitude** ajoutée sur le nom (et, plus généralement, sur les informations directement identifiantes) n'est **pas suffisante** en soi pour garantir la sécurité
  - la « pseudonymisation » par substitution ne garantit pas la constitution d'ensembles d'anonymat **assez larges** (quel que soit le critère de taille)
  - et ça, d'autant plus que la taille de la population étudiée est **souvent réduite**
  - ce qui ne veut évidemment **pas dire** que toutes les publications basées sur des entretiens ou des observations permettent la réidentification

La protection contre la réidentification

La protection contre la réidentification

- l'exemple de la « pseudonymisation » des entretiens montre que la protection des données **ne se limite pas** à la sécuration des données
- dans certains cas, il faut protéger **les données elles-mêmes** contre la réidentification

  et pas seulement en mettant des mesures pour en contrôler l'accès
- le problème ne se limite pas aux entretiens mais concerne aussi les `BdD`
- différents `exploits` publiés montrent que, dans les faits, **une quantité limitée** d'information est nécessaire pour réidentifier les personnes

Réidentification

Quelques exemples **de réidentifications** publiés :

- *The Massachusetts Governor* (Latanya Sweeney)

  réidentification à partir du croisement entre une base de données médicale publiée et les listes électorales

- *The AOL Search Queries* (*The New York Times*)

  réidentification à partir du croisement entre les logs de requêtes sur le moteur de recherche de `AOL` *et l'annuaire téléphonique*

- *Riding with the Stars* (Antony Tockar)

  réidentification d'une vedette américaine à partir de ses trajets en taxi

- *The Netflix Dataset* (Narayanan et Shmatikov, 2008)

  réidentification à partir du croisement entre une fichier de préférences cinématographiques et des évaluation sur `IMDB`

Unique in the Crowd

- étude publiée par (Montjoye et al., 2013)
- portant sur **la mobilité** à partir de données d'un opérateur téléphonique

  *quinze mois de déplacements d'*1/2 *millions d'utilisateurs*

- et ne cherchant pas tant à **réidentifier** les personnes
- qu'à démontrer **l'unicité** des traces laissées par les utilisateurs du réseau

  - l'étude montre ainsi que **quatre points** choisis **au hasard** suffisent pour identifier 95% des personnes ($\epsilon > .95$)

    et que deux points suffisent pour identifier 50% des personnes ($\epsilon > .5$), *avec* ϵ *la fraction de traces uniques*

  - et qui montre aussi **les difficultés à désidentifier** ce type de données en réduisant la résolution des données

    à partir d'un modèle utilisant une loi de puissance liant l'unicité ϵ, *la résolution temporelle* h *(durée d'observation), résolution spatiale* v *(nombre d'antennes) et le nombre de points disponibles pour un attaquant :*

    $$\epsilon = \alpha - (vh)^{\beta}$$

Réidentification

- à partir de ces quelques exemple, on peut notamment remarquer :
  - la pseudonymisation **n'empêche pas** la réidentification
  - les trois premiers exemples et le dernier montrent le caractère identifiant **des données géographiques**
  - le troisième montre, lui, que **le caractère épars** des données ne constitue pas une protection mais peut, au contraire, faciliter la réidentification
- les études adoptent des point de vue **différents** (diversité et entropie des populations, unicité des individus)
  - mais elles ont ceci de commun de montrer l'hétérogénéité des populations est **le moteur de la réidentification**
  - en pratique, c'est cette hétérogénéité qui est **exploitée** pour le profilage des personnes à des fins commerciales, de surveillance, criminelles,. . .
  - soient les traitements de DCP qui ont motivés la mise en place **d'une réglementation spécifique**

La protection contre la réidentification

- différentes techniques de **protection contre la réidentification** ont été proposées :

  k-anonymity, *l-diversity*, *t-closeness*, *differential privacy*, ...

- ainsi que **des attaques** contre (ou des failles de) chacune d'entre elles
- ce qui ne veut **pas dire** qu'elles sont systématiquement défaillantes
- la difficulté est plutôt **d'estimer** le niveau de protection contre la réidentification qu'elles proposent (à la différence de la cryptographie)
  - pas de consensus pour définir et mesurer la privauté
  - difficultés à modéliser l'information auxiliaire détenue par un attaquant

    touche aux limites de la théorie de l'information ?

Libéralisation des données

- plusieurs des études reposent sur des données **rediffusées** par leurs producteurs
- ce qui illustre **les tensions** (voire les)injonctions contradictoires) actuelles entre
  - protection des données
  - et diffusion des données
- et là, la question concerne aussi les publications de documents reposant sur des **données qualitatives**

Conclusion

Conclusion

- l'application de la réglementation sur les DCP ne se résume pas à **des formalités**
- mais peut avoir des **conséquences** sur le traitement de données
- et implique de prendre **les mesures adéquates**
- ce qui nécessite de s'y préparer **à l'avance**
- p. ex., en l'intégrant dans **un plan de gestion** de données

Merci pour votre attention

Bibliographie

Bibliographie I

Coulmont, Baptiste (2017), « Le petit peuple des sociologues. Anonymes et pseudonymes dans la sociologie française »,. *Genèses*, 107, 2, p. 153–175.

Montjoye, Y.-A. de, C. Hidalgo, M. Verleysen et V. Blondel (2013), « Unique in the Crowd : The privacy bounds of human mobility »,. *Nature srep*, 1376, 3.

Narayanan, Arvind et Vitaly Shmatikov (2008), « Robust De-anonymization of Large Sparse Datasets »,. *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, SP '08, Washington, IEEE Computer Society, p. 111–125.

Soubiran, Thomas (2017), *Protection des données à caractère personnel et qualité des enquêtes statistiques*. journée d'étude APPEL Le cadre juridique applicable aux traitements de données à caractère personne. URL : `https://hal.archives-ouvertes.fr/hal-01589980`.