La mise en conformité des traitements de données personnelles en SHS : bases épistémologiques pour la négociation

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

Colloque inaugural de la PUD-GA

Grenoble, 13 septembre

Introduction

- la mise en œuvre du **cadre juridique applicable au traitement de données à caractère personnel** (DCP)
- implique **la mise en conformité** des traitements
- ainsi que l'obligation, pour le responsable de traitement (RdT), **d'enregistrer** ses traitements **au registre de son délégué à la protection des données** (DPD) désigné
- la présentation portera sur **la mise en conformité** des traitements en SHS
  - en référence aux traitements numériques
  - et, plus particulièrement, l'enquête par questionnaire
- dans la perspective de **leur enregistrement**

Note :
La présentations est en partie issue d'une série de présentations disponibles qui développent, entre autres, les aspects réglementaires et les mesures à adopter : https://pro.univ-lille.fr/thomas-soubiran/publications/#dcp

Introduction

- la présentation se focalisera sur la mise en œuvre **de trois des grands principes** de la réglementation

  information, finalité, proportionnalité et de pertinence

- selon **deux axes** :
  - d'une part, ce que vous pourrez faire dépendra de votre capacité à **le motiver**
  - d'autre part, les principes de la réglementation et les démarches qu'elle implique
    - ne sont pas nécessairement **orthogonaux** à la préparation des enquêtes
    - et que les principes de la réglementation peuvent aussi renvoyer à **certains débats** récurrents dans différentes disciplines
  - autrement dit, on peut dessiner **une « épistémologie »** (implicite) de la réglementation
    - **∼la personnalité des objets**

      nommé ainsi faute d'avoir trouvé un terme générique dans la littérature
    - **la définition de l'objet :** le raisonnement hypothético-déductif
    - **l'opérationnalisation des hypothèses :** le principe de parcimonie
  - l'adéquation avec la réglementation dépendant toutefois des **parti-pris adoptés**

    `TIMTWOTDO`

Plan

Les grands principes de la réglementation

Récap : les grands principes de la réglementation

Les **grands principes** de la réglementation sont :

- **information :**

  *les personnes doivent être en mesure de **décider de l'utilisation** des informations les concernant*

- **limitation de la finalité :**

  *les données doivent être traitées de façon **compatible** avec une finalité **précise***

- **minimisation des données :**

  *seules les informations **strictement nécessaires à la réalisation** de la finalité doivent être traités*

- **limitation de la conservation :**

  *une fois la finalité réalisée, les informations doivent être **détruites** ou **anonymisées***

- **protection dès la conception (*privacy by design*) :**

  *la protection des personnes et la sécurité des données doit être intégrée **dès la conception** du traitement*

La mise en conformité du traitement

- **la mise en conformité du traitement** consiste trouver **la traduction de ces principes** pour chaque traitement
- les principes sont volontairement **très abstraits**
  - si la réglementation n'apparaît pas comme pensée pour les sciences sociales
  - c'est qu'elle n'a été pensée **pour aucune application en particulier**
  - ou presque. . .

  la réglementation est **parcimonieuse** *dans ses principes (cf. p. 27)*
- **toute la difficulté** de la mise en conformité réside donc dans **la traduction pratique de ces principes**
  - la généralité des principes permet une certaine **souplesse** dans l'application de la réglementation

    afin de pouvoir s'adapter à la multiplicité des traitements
  - mais la généralité confère parfois au **flou**. . .

- le cadre applicable aux DCP est un exemple **de droit « souple »**

  ∼règles de droit qui n'ont pas de caractère obligatoire (oxymore ?)

- en effet, son application repose notamment **sur la CNIL**
  - qui dispose (**droit « dur »**) de pouvoirs
    - de **contrôle**
    - de **sanction**
    - qui ont de plus été renforcés par le RGPD et la loi protection des données
  - mais repose aussi (**droit « souple »**)
    - sur **les préconisations** de la CNIL
    - ainsi que **ses certification**

      des personnes, des produits et des systèmes de données ou de procédures

    - et encore **ses délibérations**
  - qui forment **la doctrine** de la CNIL
  - et n'ont **aucun caractère obligatoire**

La doctrine de la CNIL

- ce qui ne veut **pas dire** qu'il ne **faut pas** prêter attention à la doctrine de la CNIL
  - de par **la généralité** du cadre applicable aux DCP
  - cette doctrine est **fondamentale** dans son application
  - dans les faits, la mise en conformité nécessite de **se référer** à la doctrine de la commission
- mais, les SHS se distinguent surtout par **leur (quasi-)absence** dans la doctrine de la CNIL
  - de par **le faible intérêt** suscité par la protection des données depuis 1978

    avec des nuances, particulièrement disciplinaire (-cf. démographie du fait de l'INED)
  - comme le montre **l'analyse thématique** des délibérations de la CNIL
- ce qui **complique** d'autant plus la mise en conformité
  - car les délibérations, préconisations,. . . portent le plus souvent sur des traitements **très éloignés** des SHS
  - cf. SOUBIRAN (2019b) pour plus de détail sur l'analyse

Principe d'information des personnes

Le sujet de l'objet

- en SHS, **les enquêtés** peuvent être des :
  - « individus »
  - « agents »
  - « acteurs »
  - « actants »
  - ou encore, un « terrain », un « matériau »
  - *ad lib*. . .

  enquêtés est ici entendu au sens large : ceux sur qui les informations sont collectées
- du point de vue de la réglementation, les enquêtés sont **des personnes** :
  - ils sont dotés **d'une personnalité juridique**
  - et sont titulaires **de droits** et **de devoirs**

Le sujet de l'objet

- la particularité de la réglementation sur les DCP étant que :
  - ce sont **les personnes concernées** qui ont **des droits**

    explicitement listés dans le RGPD
  - et c'est **le responsable de traitement** qui a **des obligations**

    tout aussi explicitement listées dans le RGPD ainsi que les sanctions encourues pour leur non-respect
- donc, pas de **rupture épistémologique** possible ici (« malédiction »[2], Bourdieu, Chamboredon et Passeron (1968))
  - non seulement, **les objets parlent**
  - mais, en plus, **ils ont des droits**

  mais ceci ne concerne bien évidemment que le traitement et ne contraint pas les interprétations

Informationelle Selbstbestimmung

- les personnes ont, en premier lieu, droit à **l'autodétermination informationnelle**

> LIL art. 1 : Les droits des personnes **de décider et de contrôler les usages qui sont faits des données à caractère personnel les concernant** et les obligations incombant aux personnes qui traitent ces données s'exercent dans le cadre du règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 [. . .] et de la présente loi.

- de ce principe découle tous **les autres droits**
- de ce point de vue, la question de **l'information des personnes** sur les traitements les concernant est **primordiale**
  - l'application de la réglementation **ne s'identifie pas à l'obtention consentement**

    > on peut d'ailleurs s'en passer dans certains cas (d'autres fondements juridiques sont envisageables)

  - il est par contre **beaucoup plus difficile de s'affranchir** plus ou moins partiellement de l'obligation d'information des personnes
  - c'est toutefois **envisageable** pour les traitements à des fins de recherche scientifique

    > il peut même être –théoriquement du moins– envisageable de ne pas décrire précisément la recherche

  - mais sous conditions

Information des personnes

Exemple : les collectes de **données manifestement rendues publiques** par la personne concernée

- le responsable de traitement est soumis à une obligation **d'information** des personnes (RGPD **art. 14 § 1**)
- mais **pas besoin de consentement**, même pour **les données sensibles**

  sous conditions

- ces informations doivent être fournies dans **un délai raisonnable** après avoir obtenu les données à caractère personnel, mais ne dépassant pas un mois RGPD **art. 14 § 3 (a)**

Information des personnes

Exemple : les collectes de **données manifestement rendues publiques** par la personne concernée

néanmoins, ces obligations **ne s'appliquent pas** dans les cas suivants (RGPD **art. 14 § 5**) :

- lorsque l'information est impossible ou exige des efforts **disproportionnés**

  *en particulier pour les traitements à des fins **archivistiques dans l'intérêt public**, à des fins de **recherche scientifique ou historique** ou à des fins **statistiques***

- si l'information des personnes est susceptible de **compromettre gravement** la réalisation de la finalité du traitement

Information des personnes

Exemple : les collectes de **données manifestement rendues publiques** par la personne concernée

dans ces cas de figure,

- le responsable de traitement doit prendre **les mesures appropriées** pour protéger les droits et libertés ainsi que les intérêts légitimes de la personne concernée
- lorsque l'information des personnes est impraticable, la CNIL recommande de fournir **une information générale**, par exemple sous forme de mention sur le site

Information des personnes

Exemple : les collectes de **données manifestement rendues publiques** par la personne concernée

dans ces cas de figure,

- le responsable de traitement doit prendre **les mesures appropriées** pour protéger les droits et libertés ainsi que les intérêts légitimes de la personne concernée
- lorsque l'information des personnes est impraticable, la CNIL recommande de fournir **une information générale**, par exemple sous forme de mention sur le site

Questions : pour les traitements en SHS, qu'est-ce qui constitue des efforts disproportionnés, qu'est-ce qui est reconnu comme pouvant compromettre gravement la réalisation de la finalité du traitement,. . . ?

conjecture :

- l'autodétermination informationnelle est une déclinaison **du choix rationnel**
  - maximiser sa fonction d'utilité nécessite (notamment) **une information parfaite** des individus
  - or, le traitement crée une situation **d'asymétrie d'information** entre le responsable de traitement et les personnes concernées
  - la réglementation vise **à rétablir la symétrie de l'information** entre la personne concernée et le responsable de traitement
  - pour qu'ils puissent décider de façon à maximiser **leur fonction de confidentialité**
- la discussion de ce point dépasse le propos de cette présentation

  mais ***« privacy paradox »***,. . .

Principes de finalité des traitements et de minimisation des données

Fin de traitements

Pour être conforme, un traitement se doit de répondre **à une fin** (et une seule) :

- Les données à caractère personnel doivent être [. . .] collectées pour des finalités déterminées, explicites et légitimes, et ne pas être traitées ultérieurement d'une manière incompatible avec ces finalités (RGPD **art. 5 § 1 (a)**, LIL **art. 4 § 2**)
- principes **de finalité** et **de limitation de la finalité**

De plus,

- Les données à caractère personnel doivent être [. . .] adéquates, pertinentes et limitées à ce qui est nécessaire au regard des finalités pour lesquelles elles sont traitées (RGPD **art. 5 § 1 (c)**, LIL **art. 4 § 3**)
- principe **de minimisation des données**

  ou encore principes de proportionnalité et de pertinence au regard de la finalité

Principe de finalité

- tout traitement de données personnelles doit répondre **un but précis** (« déterminé »)

  la question n'est pas seulement ce qui va être **collecté** mais aussi ce qui va en être **fait** et dans quel **but**

- le principe de finalité est **la pierre angulaire** de la mise en œuvre de la réglementation
- car de la finalité découle :
  - **ce qui peut être collecté**
    - seules les données **directement en lien** et **strictement nécessaires** à la réalisation finalité du traitement peuvent être recueillies
    - **chaque information** qui va être collecté doit donc être **motivé** et **justifié** au regard des objectifs poursuivis
  - et **les opérations qui peuvent être réalisées** sur les données collectées
  - **l'information des personnes**
  - ainsi que **la durée de conservation**,. . .

Note : une « finalité recherche » n'est **pas** une finalité **suffisamment déterminée et explicite** pour rendre un traitement conforme

*les données collectées en sciences sociales et leur utilisation sont, dans les faits, **trop diversifiées** pour être considérées comme déterminées et explicites*

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 operandi*)
- **pendant combien de temps :** limitation de **la durée de conservation** des données

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

L'enquête en questions

Autrement dit, pour enregistrer un traitement, 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**

  en n'oubliant pas les publications, rediffusion de base de données,. . .

La fin justifie les moyens

- au regarde de ce qui se précède, les démarches IL ne **se rajoutent pas** à l'enquête
  - dans le sens où la préparation de l'enregistrement procède **d'une démarche homologue** à la préparation de l'enquête
  - par contre, la mise en œuvre d'autres principes nécessitent **des actions spécifiques**

    particulièrement pour ce qui est de la sécurité des données et des systèmes d'information
- d'autre part, la question n'est **pas tant de savoir ce qu'il est possible (ou pas) de faire**
  - mais plutôt de **votre capacité à motiver** ce que vous voulez faire et comment
  - **sous différentes conditions** (licéité, proportionnalité et pertinence,. . .)

    bien évidemment, tout n'est pas possible
  - de ce fait, lors de l'enregistrement, il convient de
    - ne **pas se brider** *a priori*

      tout en étant transparent
    - à partir du moment où l'on a défini **des objectifs clairs**
    - et identifié **les moyens pour y parvenir**

Remarques

- l'application de ces différents principes fait pencher la conception et la réalisation du traitement du côté d'une forme faible de l'approche **∼hypothético-déductive**
- plutôt que **l'induction**
  - il faut d'abord déterminer clairement l'hypothèse que l'on souhaite éprouver

    ou réfuter – au sens poppérien –
  - pour déterminer ce qui est nécessaire à sa mise à l'épreuve
  - et cela vaut pour un questionnaire comme une grille d'entretien
- l'application stricte de ce principe peut, **parfois**, se révéler **problématique** dans certains cas
  - une finalité monographique (ou prosopographique) est, de ce point de vue, **un oxymore**
  - le RGPD ménage toutefois **quelques marges** pour les traitements à fin de recherche (c33)

    « il n'est pas possible de cerner entièrement la finalité du traitement des données à caractère personnel à des fins de recherche scientifique au moment de la collecte des données »
  - mais sous quelles conditions ?

Le principe de parcimonie

Le principe de parcimonie

- l'approche ∼**hypothético-déductive** permet d'esquisser une convergence entre la préparation de l'enquête et l'enregistrement du traitement
- les principes de finalité et de proportionnalité + pertinence peuvent aussi être reliés au principe **de parcimonie**
- ce principe est souvent associé au **« rasoir » d'Ockham** et à la formule :

  « *pluralitas non est ponenda sine necessitate* » (les multiples ne doivent pas être utilisés sans nécessité)

- il s'agit d'un principe d'abord **heuristique**
  - qui pose qu'entre **plusieurs hypothèses**
  - l'hypothèse **la plus simple** est préférable

    simple et non simpliste

  - si elle apparaît **suffisante**
- le principe n'est **pas qu'heuristique**, la parcimonie permettant d'analyser certains processus (physiques ou biologiques)

Le principe de parcimonie

Ce principe a de nombreuses expression dans différentes disciplines comme, p. ex., dans les sciences de la nature (Sober (2015)) :

- **Newton**
  - *Mathematical Principles of Natural Philosophy*

    « we are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances »
  - ce n'est pas la peine **d'en rajouter**
- **Einstein**
  - *On the Method of Theoretical Physics*

    « it can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation »
  - apocryphe ?

    « everything should be made as simple as possible, but not simpler »
  - l'hypothèse la plus parcimonieuse peut être **complexe**

    la simplicité est *relative*. . .

Le principe de parcimonie

- ces raisonnements ont leur pendant en **statistique**
- notamment dans la question de **la sélection des modèles**
- un modèle **trop complexe** fait en effet courir le risque **de surajustement|surapprentissage**
  - le modèle « colle » alors trop bien aux données
  - et masque la structure des données
  - en modélisant du bruit
  - et s'ajustera donc mal à des données similaires conduisant à des inférences erronées
  - qui trop s'ajuste mal prédit

Critères d'information

- la parcimonie s'incarne dans différentes mesures d'ajustement (Burnham et Anderson (2002))
- telles que
  - **le critère d'information d'Akaike** (`AIC`)

$$\text{AIC} = -2 \ln \hat{L}(\theta) + 2\,p \tag{1}$$

    avec $\hat{L}$ le maximum de la fonction de vraisemblance du modèle et p, le nombre de paramètres du modèle
  - **le critère d'information bayésien** (`BIC`)

$$\text{BIC} = -2 \ln \hat{L}(\theta) + \ln(n)\,p \tag{2}$$

    avec n, la taille de l'échantillon
- **la complexité** est mesurée par **le nombre de paramètres nécessaires** au test de l'hypothèse
- dans le cas où deux hypothèses différentes ont la même **(log-)vraisemblance**
- ces mesures favoriseront le modèle consommant le moins de paramètres -ie : l'hypothèse la plus simple

L'information des critères

- ces deux formules permettent de faire aussi le lien entre **complexité** et **information disponible**

  information est ici entendu au sens que la théorie mathématique de l'information lui donne

- par des voies **très différentes**

  - l'`AIC` minimise **la perte d'information**
    La perte d'information est ici mesurée par le critère d'information de Kullback-Leibler entre le modèle $f()$ et $g()$ le processus ayant généré les données y

    $$KLIC(f, g) = \int g(y) \log\left(\frac{g(y)}{f(y)}\right) \quad (3)$$
    $$= E_g[g(y)] - E_g[f(y)] \quad (4)$$

    $g()$ est évidemment inconnue. Toutefois, le premier terme de (4) est constant pour tous les modèles. L'`AIC` peut alors être obtenu en réalisant une expansion de taylor du second terme de (4).
  - le `BIC` minimise **la longueur de la description**
    Du point de vue de la `MDL`, le meilleur modèle est en effet celui qui permet l'encodage le plus court des données. Le `BIC` a toutefois été dérivé par d'un point de vue bayésien qui consiste à maximiser la probabilité postérieure d'un modèle conditionnellement aux données

    $$P(M_k|\mathbf{y}) = \frac{P(\mathbf{y}|M_k)P(M_k)}{P(\mathbf{y})}$$

    en mettant un *a priori* non informatif sur $P(M_k)$.

- $-2ln\, g(y|\hat{\theta})$ sert **d'estimateur biaisé** à

$$d(\hat{\theta}) = E_g[-2ln\, f(y|\theta)]|_{\theta=\hat{\theta}}$$

qui mesure la séparation entre le modèle génératif $g(y)$ et le modèle estimé $f(y|\hat{\theta})$

- et **le biais**

$$E_g[d(\hat{\theta})] - E_g[-2ln\, f(y|\hat{\theta})]$$

peut être estimé par $2\,p$

- la correction peut être approchée de la formule du calcul de **la variance (non-biaisée) de l'échantillon**

$$s^2 = \frac{n}{n-1}\sigma^2$$

L'information des critères

- `AIC` et `BIC` peuvent être reliées à **une autre définition de l'information**
- en l'effet, leur dérivation font intervenir **l'information de Fisher** $\mathcal{I}(\theta)$ (Cavanaugh et Neath (2011), Neath et Cavanaugh (2012))

  - $\mathcal{I}(\theta)$ mesure **la quantité d'information** qu'une variable X apporte à un paramètre θ
  - elle est définie comme la variance de la fonction de score $s(\theta) = \frac{\partial ln\, L(\theta)}{\partial \theta}$

$$\mathcal{I}(\theta) = E\left[\left(\frac{\partial}{\partial \theta} ln\, f(X\theta)\right)^2 \middle| \theta\right] \tag{5}$$

$$= E\left[\frac{\partial^2}{\partial \theta} ln\, f(X\theta) \middle| \theta\right] \quad \text{si } f() \text{ est deux fois diffentiable} \tag{6}$$

Note : $V(\theta) \geq \mathcal{I}(\theta)^{-1}$ avec une égalité lorsque $f()$ est la fonction ayant généré les données

L'information des critères

- dans (1) et (2), $\mathcal{I}(\theta)$ **disparaît**
- mais demeure dans **d'autres versions** de p. ex. l'`AIC`
- telle que **le `GAIC`** (*Generalized Akaike Information Criterion*)

$$\text{GAIC} = -2 \ln \hat{L}(\theta) + 2 \operatorname{tr}(Q(\hat{\theta})^{-1} \mathcal{I}(\hat{\theta}))$$

avec

$$Q(\hat{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^\top} \ln f(y_i; \hat{\theta}) \quad \mathcal{I}(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial}{\partial\theta} \ln f(y_i; \hat{\theta}) \frac{\partial}{\partial\theta} \ln f(y_i; \hat{\theta})^\top \right)$$

ce critère est dit robuste dans le sens où, lorsque le modèle est correctement spécifié, $Q(\theta) = \mathcal{I}(\hat{\theta})$ et $\operatorname{tr}(Q(\theta)^{-1}\mathcal{I}(\theta)) = p$. Le `GAIC` est alors égal à l'`AIC`.
Il est de plus identique à l'`AIC` en population finie (-ie : données de sondages probabilistes) dans le cas du tirage aléatoire simple (Lumley et Scott (2015)).

- ou du `BIC`

L'information des critères

- ces critères permettent de relier **complexité des hypothèses** et **quantité d'information** disponible pour les tester
- et illustrent le principe (simple...) que, **plus une hypothèse est complexe, plus elle demande de l'information pour la tester**
  - on ne peut tester que des hypothèses aussi complexes que les données le permettent
  - le rejet d'une hypothèse plus complexe se fait donc au regard des données disponibles – cf. réfutabilité
  - la complexification du modèle rajoute de l'incertitude

    arbitrage biais–variance
- on peut certes adapter **la taille de l'échantillon à la complexité des hypothèses**

  en lien avec la complexité des hypothèses et la variance des variables mobilisées pour la tester (puissance des tests)
- mais on est toujours **contraint à la parcimonie** au regard de l'information disponible
  - car le recueil de l'information a **un coût**

    pas seulement monétaire
  - qui **limite** la quantité d'information disponible globalement
  - cf. Soubiran (2017) dans le cas des données auxiliaires

Parcimonie, proportionnalité et pertinence

- le principe de minimisation des données n'est donc peut-être **pas si contraignant** qu'il n'y paraît
  - le principe de parcimonie permet d'envisager des critères **d'adéquation et de pertinence** au regard de la finalité
  - en reliant **les hypothèses et l'information** nécessaire pour les tester

  du moins sur le principe, la pratique peut se révéler plus ardue – cf. après
- la référence au principe de parcimonie permet aussi de **modérer** une application **trop drastique** du principe de minimisation
  - on observe en effet **rarement directement** ce que l'on souhaite mesurer

    profession, attitudes,. . .
  - ce qui peut impliquer de collecter **un nombre conséquent de renseignements** sur les personnes interrogées
  - de plus, les analyses portant sur des données **non-expérimentales**
  - il est nécessaires d'introduire **des variables de contrôle**

    + questions de l'endogénéité des variables, . . .
  - qui ne se **pas liées directement** aux hypothèses

Les concupiscences

- en résumé, **la mise en conformité** des traitement nécessite en premier lieu
  - de se concentrer sur **un nombre limité** d'hypothèse précises
  - et de déterminer précisément **ce qui est nécessaire** pour les tester
- **à mon avis**, de ce point de vue,
  - l'application de la réglementation permet d'aller contre certains penchants lors des collectes de données
  - qui conduisent notamment à multiplier les thèmes abordés dans les questionnaires
  - parfois même sans les problématiser *a minima*
  - et donc sans pouvoir déterminer ce qui est précisément nécessaire pour répondre à la question
  - ne collectant donc pas assez d'information pour y répondre
  - ce qui conduit à des questionnaires trop longs et trop imprécis
  - ne prenant pas en compte les conditions de passation des questionnaires
  - et où, à force de vouloir dresser des grands tableaux,
  - on ne peut distinguer ni le détail, ni l'ensemble

Application du principe de minimisation des données

Principes de proportionnalité et pertinence

- ***The Name of the Game :*** vous faire collecter **le moins d'informations possible** (minimisation des données)

  avoir de (bonnes) raisons (clairement définies) de collecter des données **ne suffit pas**

- en pratique, c'est un des aspects **les plus délicat** de l'application de la réglementation en sciences sociales :
  - dépasse l'aspect **procédural**
  - peut toucher **au contenu** des recherches elle-mêmes
  - particulièrement lors de la collecte de **données sensibles**
- car, même si on peut trouver **des convergences de principe** (cf. *supra*)
  - le principe est généralement de façon **très restrictive**
  - la difficulté étant alors d'établir ce qui peut être considéré comme **proportionné et pertinent** pour une recherche
  - en **l'absence des** SHS dans la doctrine de la CNIL
  - qui peut conduire les DPD à adopter **des positions défensives**

    la première mission du DPD est **la protection juridique de son établissement**

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) : cas où la réalisation de la finalité nécessite de croiser pratiques religieuses et pratiques politiques (**autres données sensibles**)

  Exemple : analyse du vote

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

  avec des contre-exemples comme « l'origine » des personnes (cf. Soubiran (2019a)), mais la question dépasse la réglementation sur les DCP

- 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

L'impact sur les personnes

- comme le montre **l'analyse thématique des délibérations** de la CNIL,
  - la réglementation vise **à protéger** les personnes
  - contre **l'impact** sur elles de traitements (automatisés ou non) se fondant sur leur DCP
  - comme des **prises de décisions**

    cf. l'attention portée au profilage
- les finalités en SHS **diffèrent** des finalités des entités (entreprises, administrations, associations...) qui constituent le gros des délibérations de la CNIL

  les SHS n'ont généralement pas directement à faire à des administrés, des assurés sociaux, des usagers, des employés, des clients,... mais bien à **des enquêtés**
- cet aspect ne doit pas être négligé lors des démarche IL
  - pour motiver une application **moins drastique** du principe de minimisation
  - tout en adoptant **les mesures de sécurité** adéquates
  - et en gardant à l'esprit que l'absence de prises de décisions **ne signifie pas** que le traitement soit sans conséquences sur les personnes

    particulièrement par le biais des publications

Conclusion

Conclusion

- **la mise en conformité** des traitements nécessite en premier lieu
  - de déterminer **ce que l'on souhaite montrer**
  - et de déterminer **ce qui est nécessaire** pour en faire la démonstration
- différentes **convergences** peuvent être notées entre la préparation d'une enquête et la mise en conformité du traitement

  mais en se gardant d'une application trop drastique du principe de minimisation
- une partie **des difficultés** dans la mise en conformité proviennent de **la (quasi-)absence des** SHS de la doctrine de la CNIL

  mais il est possible d'y remédier

Merci pour votre attention

Bibliographie

Bibliographie I

Bourdieu, Pierre, Jean-Claude Chamboredon et Jean-Claude Passeron (1968), *Le métier de sociologue,* Les textes sociologiques, Mouton, 431 p.

Burnham, Kenneth P. et David R. Anderson (2002), *Model Selection and Multimodel Inference,* Statistical Theory and Methods, Springer-Verlag New York, xxvi + 488 p.

Cavanaugh, Joseph E. et Andrew A. Neath (2011), « Akaike's Information Criterion : Background, Derivation, Properties, and Refinements », sous la dir. de Miodrag Lovric, Springer Berlin Heidelberg, p. 26-29.

Lumley, Thomas et Alastair Scott (2015), « AIC and BIC for modeling with complex survey data », *Journal of Survey Statistics and Methodology*, 1, 3, p. 1-18.

Neath, Andrew A. et Joseph E. Cavanaugh (mar. 2012), « The Bayesian Information Criterion : Background, Derivation, and Applications », *WIREs Comput. Stat.* 2, 4, p. 199-203.

Sober, Elliott (2015), *Ockham's Razors : A User's Manual,* Cambridge University Press, x + 314 p.

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 personnel*, Lille, 28 avr. 2017, `https://hal.archives-ouvertes.fr/hal-01589980`.

Bibliographie II

– (juil. 2019a), « La réglementation applicable aux données personnelles en SHS », école d'été Quantille 2019, Lille, 24-27 juin 2019, https://pro.univ-lille.fr/fileadmin/user_upload/pages_pros/thomas_soubiran/dcp/qtll2019-dcp.pdf.

– (jan. 2019b), « Quarante ans de délibérations de la CNIL », colloque du PIREH *Histoire, langues et textométrie*, Paris, 1er-2 jan. 2019, https://pro.univ-lille.fr/fileadmin/user_upload/pages_pros/thomas_soubiran/dcp/pireh2019--doctrine-presentation.pdf.