La réglementation relative aux données à caractère personnel en sciences sociales

Thomas Soubiran

CERAPS - UMR 8026 CNRS-Lille 2

Journées Data-SHS

11-12 décembre 2017

La réglementation sur les données à caractère personnel

La réglementation sur les données à caractère personnel (DCP) :

- ensemble de règles juridiques en vigueur relatives à **l'utilisation** (« traitement ») de DCP, c-à-d de données permettant **d'identifier des personnes physiques**
- définit **les droits des personnes** concernées par le traitement
- et les **obligations** à respecter lors du traitement de DCP les concernant

Les données à caractère personnel en sciences sociales

Le traitement de DCP est **au cœur** de l'activité des sciences sociales :

- l'utilisation de DCP peut en effet y prendre de **multiples formes** :
  - **collecte de données**
  - **les analyses** (automatisées ou non)
  - ainsi que dans les **publications**
- c'est pourquoi les traitements en sciences sociales tombent le plus souvent **dans le champ d'application** de la réglementation en vigueur
  - et ce, même si les personnes sont **nommées** ou **pseudonymisées**
  - où si **l'identité des personnes** n'est pas utilisée ou si les DCP collectées ne sont utilisées pour **(ré)identifier les personnes**

La réglementation relative aux DCP

En France, le traitement de **DCP** est jusqu'à présent encadré par **la loi informatique et libertés** (LIL) :

- loi votée le 6 janvier 1978
- elle a été modifiée par la suite à plusieurs reprises, notamment en 2004 pour transposer la **directive européenne** sur la protection des données de 1995
- la prochaine modification interviendra **l'année prochaine**

En effet, le 25 mai 2018 prochain,

le règlement européen sur la protection des données entrera en application

- le règlement général sur la protection des données (**RGPD**) est **d'application directe** dans le droit des États membres (pas de transposition)
- il **abrogera** la directive de 1995
- il **n'abroge pas** la LIL mais en rend néanmoins inapplicable les dispositions incompatibles avec le règlement

Le règlement européen sur la protection des données

Depuis sa publication au Journal officiel de l'**UE** le 24 mai 2016, le **RGPD** constitue **le nouveau texte de référence** européen en matière de protection des données à caractère personnel :

- adopté après quatre ans (d'âpres) négociations
- le **RGPD** reprend **les fondamentaux** de la directive, les grands principes restent en effet les mêmes

  *le **RGPD** explicite notamment différentes interprétations de la réglementation*

- marque notamment le passage d'un régime **de déclaration préalable** à un régime **de responsabilisation**

Le règlement européen sur la protection des données

La situation actuelle est donc **transitoire** :

- le règlement est **en vigueur** mais pas encore en application
- un certain nombre de **clarifications** doivent encore être apportées, notamment par la CNIL
- néanmoins, de par la proximité de la date d'application du règlement, la présentation portera **sur le RGPD** en mentionnant les changements avec la LIL le cas échéant

Plan

- **appréhender la réglementation sur les DCP**
  - **chronologie**
  - **remarques générales**
- **notions et agents de la protection des données** (en partant de trois notions fondamentales) :
  - **données à caractère personnel**
  - **traitement**
  - **finalité**
- **mise en œuvre de la réglementation en sciences sociales :**
  - **interprétation** (et difficultés d'interprétation) des notions dans le contexte spécifique des sciences sociales
  - **protection des données**
  - **identification**, **désidentification** et **réidentification** des personnes, particulièrement lors du traitement de données numériques

Cette présentation est partiellement issue de notices rédigées **sur la** LIL avec Émilie Masson, juriste au service du CIL du **CNRS**. Ces notices sont accessibles à cette page :

https://extra.core-cloud.net/collaborations/CIL_Extranet/partage_ESR/GuideSHS/GuideSHS.aspx

Note : l'accès nécessite de s'authentifier via la fédération d'identité de RENATER

Plan

- la présentation sera plutôt axée sur **les aspects généraux** et **procéduraux** de la réglementation sur les DCP
  - en sciences sociales, peu **de cas typiques**
  - les traitements doivent être analysés **au cas par cas**
  - la réglementation ne fournit qu'**un cadre général**
  - les interprétations spécifiques **manquent encore** pour les sciences sociales
- de plus, la présentation n'abordera (quasiment) pas la question **des données de santé** :
  - non pas que les sciences sociales ne soient pas **concernées**

    Exemple : psycholgie, STAPS, sociologie de la santé,. . .
  - mais plutôt parce que les données de santé sont considérées comme **les plus sensibles**
  - elles font donc l'objet **dispositions spécifiques**
  - l'analyse juridique n'en est que plus complexe et nécessiterait **une présentation spécifique**

Appréhender la réglementation

Chronologie

Chronologie de la réglementation sur les DCP

2018	entrée en application du règlement 2016/679 et fin du délais pour la mise en conformité pour les traitements en cours (25 mai) vote d'une nouvelle loi ?
2016	**règlement 2016/679/UE du Parlement européen et du Conseil relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données** (RGPD) *abroge la directive* 95/46/CE
	directive 2016/680/UE du Parlement européen et du Conseil relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel d'enquêtes et de poursuites en la matière ou d'exécution de sanctions pénales et à la libre circulation de ces données
2004	traduction dans le droit français de la directive 95/46/CE
1995	**directive 95/46/CE du Parlement européen et du Conseil relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données**
1981	**convention 108 pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel** *convention du Conseil de l'Europe*
1979	**résolution du Parlement européen sur la protection des droits de la personne face au développement des progrès techniques dans le domaine de l'informatique**
1978	**loi 78-17 relative à l'informatique, aux fichiers et aux libertés** (LIL)

Note : à partir du début des années 70, différents États européens ont commencé à se doter de législations sur les DCP comme le Land de Hesse en 1970 (*Hessisches Datenschutzgesetz*, la première au monde), la Suède (*Datalag*, 1973) ou la RFA (*Bundesdatenschutzgesetz*, 1977) (FUSTER GONZÁLEZ, 2014)

Autres réglementations

Autres textes traitant de la question des DCP :

2016	**loi** 2016-1321 **pour une République numérique** *succède à la* LCEN et *anticipe le* RGPD
2008	**loi** 2008-696 **du 15 juillet 2008 relative aux archives**
2004	**loi** 2004-575 **pour la confiance dans l'économie numérique** (LCEN)
2002	**directive** 2002/58 **du Parlement européen et du Conseil concernant le traitement des données à caractère personnel et la protection de la vie privée dans le secteur des communications électroniques**
1978	**loi** 78-753 **portant diverses mesures d'amélioration des relations entre l'administration et le public et diverses dispositions d'ordre administratif, social et fiscal** *création de la Commission d'accès aux documents administratifs* (CADA)
1951	**loi** 51-711 **sur l'obligation, la coordination et le secret en matière de statistiques**

ainsi que : droit à l'image, code du patrimoine,. . .

L'émergence de la réglementation relative aux DCP

- la mise en place des réglementations est liée au développement de l'informatique dans l'après-guerre
  - dans les années soixante-dix, il s'agissait principalement d'encadrer le traitement de DCP par **les États**
  - depuis s'est notamment ajouté la valorisation de DCP par **les entreprises**
- en effet, de nombreuses entreprises ont désormais un *business model* fondé sur **la marchandisation** des DCP et les sommes en jeu sont considérables
  - les négociations autour du RGPD ont ainsi généré une intense activité de **lobbying** de la part des GAFAM
- la réglementation est le produit de **rapports de force** politiques et économiques variables dans le temps qui dépassent largement la seule question des sciences sociales
- les sciences sociales **pèsent peu** et apparaissent parfois comme un dommage collatéral

Note : les sciences sociales pèsent d'autant moins que ses représentants se mobilisent peu sur le sujet

La portée du cadre applicable

Pour autant,

- pour **historicisable** qu'elle soit, la réglementation n'en est pas **contingentée** à un contexte précis, du moins dans ses principes

  Note : certaines dispositions visent malgré tout (im|ex)plicitement certains agents comme les GAFAM ou la recherche médicale

- dès le départ, les réflexions ont visé à établir un cadre **plus général** que les cas concrets qui les ont initiées comme « l'affaire » SAFARI en France

« L'affaire » SAFARI

Le contexte d'adoption la LIL :

- le développement de l'informatique avait aussi suscité des débats sur l'opportunité de légiférer à ce sujet en France

  > ***une proposition de loi*** *tendant à la création d'une commission de surveillance et de « tribunal de l'informatique » avait été formulée en 1970 par Michel Poniatowski mais n'avait pas abouti*

- la question refit surface suite au **projet** SAFARI (Système informatisé pour les fichiers administratifs et le répertoire des individus) :

  - SAFARI était un projet de base de données du Ministère de l'intérieur visant **à apparier** différentes bases administratives à partir du **NIR**
  - le projet fut révélé sous par **Le Monde** le 21 avril 1974 qui titra sur cinq colonnes : *« Safari » ou la chasse aux Français*
  - et fut **abandonné** dans la foulée

- la polémique provoquée par le projet conduisit de plus à la mise en place d'une Commission informatique et libertés dont les débats aboutirent **au vote de la loi de 1978**

Un cadre juridique général

- la première mouture de la LIL portait malgré tout la marque **du contexte de son élaboration**
- elle **contraignait** fortement les traitements du secteur publique
  - elle interdisait, p. ex., le transfert de données nominatives à la statistique publique et ce, malgré la loi sur le secret de 1951 (QUANTIN et RIANDEY, 2012)
  - il a fallu attendre la modification de la LIL par la `loi du 23 décembre 1986` pour que des informations nominatives puissent à nouveau être transmises aux services de la statistique publique

    Note : voir aussi la `norme simplifiée n° 26` du 13 novembre 1984 concernant les traitements statistiques effectués dans le cadre des travaux du Conseil national de l'information statistique (**CNIS**)
- au grès des modifications et des délibérations de la CNIL, la LIL s'est toutefois **peu à peu affranchie** de son contexte d'origine

  Note : quelques constantes demeurent, comme le strict encadrement des croisements de données entendu au sens large (cf. le principe de « minimisation » des données)

Un cadre juridique général

- les évolutions de la réglementation en vigueur ont conduit à l'élaboration **d'un cadre extrêmement général**
- qui ne se résume pas aux cas ayant conduit à légiférer
- en pratique, le problème est plutôt **inverse** :
  - dans certains cas, le caractère général du cadre est tellement **abstrait** qu'il confère même **au flou**
  - ce qui peut parfois rendre l'analyse juridique difficile, notamment pour certains traitements de DCP en sciences sociales. . .
  - . . .mais cela d'autant plus que les démarches de clarification **n'ont pas été entreprises**

Remarques préalables

Remarques préalables

- la réglementation sur les DCP est un sujet **difficile à appréhender**
- la partie qui suit vise à aborder différentes difficultés en les articulant autour **de trois points** :
  - **les données personnelles, une question juridique**
  - **un cadre juridique inapplicable ?**
  - **un cadre juridique général**

Une question juridique

Se conformer à la réglementation en vigueur est une **obligation** pour le traitement de DCP :

- le RGPD s'applique à tout traitement de DCP de personnes **résidant** sur le territoire de l'UE ou lorsque le responsable de traitement y est **établi** (RGPD art. 3)
- que les traitement soient **informatisés ou non**
- y compris pour **des fins de recherche ou d'enseignement**
- ne pas s'y conformer est une infraction **pénale**
- . . .autant d'évidences ?

Une question juridique

En pratique, les choses paraissent **moins évidentes** :

- la question des $\mathbf{DCP}$ encore **largement négligée**, voire (sciemment) ignorée

  Note : l'intérêt pour la question varie cependant fortement en fonction des disciplines

- lorsqu'elle transparaît, la question est souvent appréhendée comme relevant de **l'éthique** (personnelle ou professionnelle) ou de la **« déontologie »**
- elle est encore rarement abordée (et enseignée) du point de vue de la réglementation
- **exemple :** les manuels d'enquêtes

Le traitements de DCP dans les manuels

La question des DCP apparaît dans l'ensemble peu abordée dans les manuels :

- éventuellement quelques références à **« la confidentialité »** ou **« l'anonymisation »** ou encore l'utilisation de pseudonymes

  Notes :
  - l'anonymisation est souvent confondue avec la pseudonymisation
  - la pseudonymisation est définie de façon précise dans le RGPD

- relève de **la relation (interpersonnelle) à l'enquêté** : la confidentialité (présumée) des informations procède de la confidentialité d'une relation privilégiée
- quelques préconisations, parfois des prescriptions, faites **sans référence** à la réglementation ou validations empiriques
- les seuls manuels qui mentionnent explicitement la réglementation sont des manuels **d'analyse de données**
- **peu de développements** (listes avec ellipses entre parenthèses), le traitement de DCP semble marqué du sceau de l'évidence

Note : la littérature reflète (et perpétue) ainsi la prénotion voulant que la réglementation ne concerne que les traitement informatisés

Un cadre juridique inapplicable ?

La réglementation est aussi parfois perçue comme :

- une **construction arbitraire**
- ou conçue à partir de situations **n'ayant rien à voir** avec les sciences sociales
- ou, pour le moins, inapplicable|inadaptée
- voire comme une **« menace »** pour les sciences sociales

Un cadre juridique inapplicable ?

Dans les faits,

- la réglementation est une protection contre des risques **effectifs** pour les personnes, p. ex. dans les relations de travail
- ces risques ont leurs pendants **dans les enquêtes en sciences sociales**
- les difficultés de l'application varient grandement selon les traitements
  - elles sont souvent liées au traitement de **données sensibles**
  - elles sont pour partie **une prophétie auto-réalisatrice**
- la réglementation crée certes **un risque juridique**
  - ne pas **exagérer** cet aléas
  - ne pas négliger que la conformité est aussi **une protection**
- surtout,
  - ce risque procède des risques induits par **les traitements de DCP** (ne pas inverser causes et conséquences)
  - ne pas **se limiter** aux seuls cas où des incidents liés au traitement de DCP se sont retournés vers les auteurs de l'enquête

La « menace »

- postulat **d'innocuité** des enquêtes pour les enquêtés
  - **corrolaire :** occultation des **« menaces »** que les traitements font courir **aux enquêtés**
  - on peut pourtant trouver des exemples du contraire, avec parfois des conséquences très graves pour des membres de la population enquêtée
  - ces incidents n'ont pas nécessairement d'effets en retour sur les auteurs de l'enquête
- peu d'enquêtes portant sur ce que **fait l'enquête aux enquêtés**
- ceci est d'autant plus problématique que le RGPD rend obligatoire **les études d'impact** dans certains cas (cf. RGPD art. 35 § 1 et *infra* p. 60)

Un cadre juridique général

Des textes comme la LIL ou le **RGPD** ne fournissent qu'un **cadre général** :

- la conformité du traitement doit être établie au regard de **principes généraux**
- l'analyse juridique du traitement doit souvent se faire **au cas par cas**, particulièrement dans les traitements en sciences sociales
  - en sciences sociales, les traitements sont **très diversifiés**
  - et ce, tant du point de vue des **données collectées** (qui peuvent aller du plus trivial au plus sensible) que **des finalités**

    Note : la finalité du traitement est tout aussi importante que les caractéristiques des données traitées
  - ou **des risques** qu'ils font courir aux personnes concernées
- or, la finalité doit être **déterminée** et **explicite**
- en conséquence de quoi, arguer d'une « finalité de recherche » n'est **pas suffisant** *en soi* pour rendre un traitement conforme

Note : les traitements à fins de recherche scientifique font toutefois l'objet des dispositions spécifiques

Un cadre juridique général

Le caractère général de la réglementation fait qu'elle ne se laisse pas facilement appréhender (et expliquée) :

- difficulté d'adopter un point de vue **synoptique**
  - il peut p. ex. paraître tentant de réduire l'application à une grille qui mapperait les situations avec un « statut » juridique
  - ou à une arborescence binaire (ou *n*-aire) qui permettrait de combiner les caractéristiques du traitement et au moins partiellement automatiser l'analyse juridique
- **la diversité des situations** rend toutefois cette approche difficilement praticable

  Exemples : appréciation de la proportionnalité et de la pertinence au regard de la finalité ou encore l'évaluation des risques
- le RGPD n'est pas une liste d'interdictions (ou d'autorisations), il énonce avant tout des principes

  Note : peu de traitements sont **interdits** par la réglementation et ces interdictions peuvent faire l'objet **d'exception**
- l'analyse juridique se fait au regard de **la finalité** et **des risques** que le traitement fait courir aux personnes concernées

Un cadre juridique général

La réduction à des situations typiques n'est pas impossible en général :

- **normes simplifiées** qui permettent p. ex. d'enregistrer un ensemble de traitements récurrents une fois pour toute

  Exemple : organisation d'événements scientifiques

- ainsi que des autorisations uniques, des méthodologies de références,...
- *le Guide informatique et libertés pour l'enseignement supérieur et la recherche* édité par l'AMUE, la **CPU** et la CNIL

  le guide couvre différentes situations comme :

  - la mise en place d'un annuaire des diplômés, d'une fédération d'identités,...
  - mais aussi les enquêtes **d'insertion professionnelle** des étudiants
  - ou encore les études sur **la diversité des origines** des étudiants et les pratiques discriminatoires

- mais la démarche est toutefois difficile **systématiser** de par l'éventail des possibilités des traitements en sciences sociales

Un cadre juridique général

Alternative (pour illustrer la mise en œuvre) : **les cas pratiques** (à défaut de concrets)

- présentent d'autres difficultés :
  - tout d'abord, ce traitement peut porter sur des infractions et des sanctions, c-à-d **des données sensibles** qui comptent parmi les plus délicates
  - de plus, ce traitement pose le problème de **la réidentification**

    Exemple : la science politique est une discipline particulièrement exposée mais qui compte un nombre relativement faible de membres (cf. *Small world* à la Watts et Strogatz)
- nécessite d'anonymiser des cas comportant un grand nombre d'informations **indirectement identifiantes** (thèmes de recherche, population, contexte, hypothèses et donc idéologie sous-jacente)
- **dilemme :** plus on supprime d'informations pouvant permettre la réidentification, plus les détails disparaissent

  cf. : difficulté algorithmique de l'anonymisation *infra* p. 126
- risque de produire des cas **trop abstraits** pour être pratiques

Note : la publication de cas pratique constitue un cas concret d'application de la réglementation qui illustre certaines difficultés de l'exercice

Un cadre juridique général

De par la généralité du cadre, **la doctrine** de la CNIL revêt une grande importance dans l'analyse juridique :

- la Commission possède un pouvoir **réglementaire**
- elle publie **des normes** (normes simplifiées, autorisations uniques, actes réglementaires uniques et méthodologies de référence,...)
- ainsi que des avis, autorisations,...
- cette doctrine sert **de référence**, notamment aux CIL

Un cadre juridique général

En pratique,

- importance d'associer **votre** CIL à vos projets de recherche

  Notes :

  - le CIL a aussi une mission de conseil et d'information
  - associer votre CIL constitue autant une **obligation réglementaire** qu'un **impératif pratique**

- importance, aussi, de se familiariser à la fois avec **les notions** et **le raisonnement** de la réglementation :

  - les distinctions usuelles qui peuvent être faites en sciences sociales n'ont pas nécessairement leurs pendants dans la réglementation

    Exemple : pas de distinction entre **collecte**, **analyse** ou encore **publication**, pas de distinction de « personnes publiques »

  - et réciproquement (notamment en fonction de la finalité)

    Exemple : la minimisation des données

  - les définitions de données identifiantes, traitement, responsable de traitement, anonymisation, pseudonymisation,. . . **ne correspondent pas forcément** à l'idée que vous vous en faites
  - et ces différences peuvent avoir des implications **très concrètes**

Définitions

Notions fondamentales

Trois notions fondamentales

Les trois notions fondamentales pour circonscrire le champ d'application de la LIL et du RGPD sont :

- **données à caractère personnel**
- **traitement**
- **finalité**

En effet, la réglementation s'applique à :

- tout **traitement** (informatique ou autre) dont la **finalité** nécessite le recueil d'informations permettant **d'identifier directement ou indirectement** les personnes physiques sur lesquelles ces informations ont été collectée
- lorsque les personnes physiques concernées **résident** ou lorsque le responsable de traitement est **établi** sur **le territoire de l'UE**

La loi impose de plus que :

- la finalité soit **déterminée**, **explicite** et **légitime**
- les données collectées soient **proportionnées** et **pertinentes** au regard de la finalité du traitement
- les données soient collectées et traitées de manière **licite**, **loyale** et **transparente**

Données à caractère personnel

Définition : toute information se rapportant à une personne physique identifiée ou identifiable (RGPD art. 4 § 1)

- il s'agit de toute donnée permettant d'identifier une **personne physique** :

  « identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale » (*ibid.*)

Deux cas de figure :

- **données directement identifiantes :** données nominatives permettant l'identification directe d'une personne comme le nom, l'adresse (postale, électronique,...), téléphone, numéro de bureau,...
- **données indirectement identifiantes :** données permettant d'identifier une personne de manière indirecte, notamment par croisement

Note : si le traitement ne nécessite pas l'utilisation de données identifiantes, le RGPD ne **s'applique pas** (RGPD art. 11 § 1)

Données indirectement identifiantes

Le RGPD porte sur les informations permettant **identifier** une personne et pas seulement la nommer :

- l'application de la réglementation ne se réduit donc pas à la seule question de **« l'anonymat »** *stricto sensu*
- l'expérience ainsi que des travaux en informatique montrent en effet que l'absence ou la suppression de données directement identifiantes (ou leur absence à la collecte) n'est **pas en soi suffisante** pour prévenir toute (ré-)identification (cf. *infra* p. 99)
- en pratique, le recoupement d'informations en apparence **anodines** (même en nombre limité) peut souvent concourir à l'identification de personnes physiques
- ainsi, **la pseudonymisation** (p. ex. de citations d'entretiens) n'est pas toujours suffisante pour empêcher la ré-identification des personnes

Traitement

Définition : toute opération ou tout ensemble d'opérations effectuées ou non à l'aide de procédés automatisés et appliquées à des données ou des ensembles de données à caractère personnel (RGPD art. 4 § 2)

- définition **très large**
- recouvre quasiment tout ce qui peut être réalisé dans le cadre **d'enquêtes de terrain** tant du point de vue de la collecte (questionnaires, *data mining* sous toutes ses formes, entretiens, observations, etc.) que de l'analyse
- mais aussi des activités relevant du **fonctionnement des équipes de recherche** comme l'organisation d'événements scientifiques

  Note : dans ce cas, il existe **une norme simplifiée**

De plus,

- pas de distinction entre **collecte**, **analyse** ou encore **publication** : toutes ces opérations font parties du traitement
- le fait que les DCP collectées ne soient **pas utilisées** du tout ou seulement dans une phase du traitement comme l'analyse ne change rien
- pas plus que le **nombre** de personnes identifiables

La finalité

Définition : ?

- la notion de finalité ne semble pas avoir de définition explicite
- la notion est toutefois **caractérisée** dans les textes

La finalité se doit en effet d'être (RGPD **art. 5 § 1 (a)**) :

- **déterminée** : la finalité du traitement doit avoir été clairement définie avant la collecte
- **explicite** : la finalité doit être transparente
- **légitime** : la finalité du traitement doit être liée à l'activité du responsable de traitement (p. ex. : réaliser des enquêtes quand on est membre d'une UMR de sociologie)

La finalité

Du point de vue de la réglementation,

- 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*

- pour les sciences sociales, la finalité correspond plutôt à la **problématique** de la recherche
  - l'utilisation de chaque données traitée doit en effet être **motivée**
  - les traitements doivent respecter différentes **règles et principes**

    cf. infra : *proportionnalité et de pertinence, autodétermination informationnelle,. . .*
  - **les risques** pour les personnes concernées doivent aussi être évalués
- c'est pourquoi **chaque traitement** doit faire l'objet d'un examen

La finalité

De plus,

- les données ne peuvent pas être traitées ultérieurement **d'une manière incompatible** avec les finalités du traitement
  - les données ne peuvent être traitées **que pour la réalisation** de la finalité pour laquelle elles ont été collectées
  - le détournement de finalité constitue une **infraction pénale** (art. 226 § 21 (c) du code pénal)
  - la finalité peut néanmoins être **redéfinie** en cours de traitement sous conditions
- **exceptions :** les traitements à fins d'archivage publique, à fins de recherche et à fins de statistique
  - ces traitement ne sont **« pas considérés comme incompatibles »** avec les finalités initiales du traitement
  - des données collectées pour une autre finalité peuvent donc être utilisées pour la recherche (cf. *infra* p. 53)

La finalité

La notion de finalité est la **pierre angulaire** du RGPD :

- la question n'est pas seulement ce qui va être **collecté** mais aussi ce qui va en être **fait**
- l'important est d'établir quelle sera **l'utilisation** des données au regard de la finalité
- dans certains cas, la finalité peut même **complètement changer** l'analyse juridique d'un même type de données

  Exemple : le profilage (RGPD art. 4)
  - fait l'objet d'un encadrement juridique **plus strict** que d'autres traitements
  - notamment parce que le profilage peut servir de fondement à **une décision** (automatisée) sur la personne concernée ou l'affecter de manière significative
  - obligations relatives à l'information des personnes physiques, l'étude d'impact à réaliser par le responsable de traitement,. . .

- **exception** : **les données sensibles** qui constituent des catégories spécifiques quelle que soit la finalité de leur utilisation

Les données sensibles

Le RGPD distingue des catégories particulières de DCP : **les données sensibles**

En effet, les traitements de DCP qui révèlent :

- **l'origine raciale ou ethnique** (« étant entendu que l'utilisation de l'expression " origine raciale " dans le présent règlement n'implique que l'Union adhère à des théories tendant à établir l'existence de races humaines distinctes » (c51))
- **les opinions politiques**, les convictions **religieuses** ou **philosophiques** ou **l'appartenance syndicale**

ainsi que le traitement :

- des données **génétiques**, des données **biométriques** aux fins d'**identifier** une personne physique de manière unique, des données concernant **la santé**
- des données concernant la **vie sexuelle** ou l'**orientation sexuelle** d'une personne physique

sont **interdits** (RGPD **art. 9 § 1**) .

À cela s'ajoute le traitement des données à caractère personnel relatives (RGPD **art. 10**) :

- aux **condamnations pénales** et aux **infractions**
- aux **mesures de sûreté connexes** (mise en détention, peines de prison,. . .)

Dérogations à l'interdiction de collecte des données sensibles

Cette interdiction peut néanmoins faire l'objet **d'exceptions** (RGPD art. 9 § 2), sauf pour les deux derniers cas :

- la personne concernée a donné son **consentement** explicite au traitement (sauf si le droit national ou de l'UE en vigueur prévoit une interdiction qui ne peut pas être levée)
- le traitement porte sur des données à caractère personnel qui sont manifestement **rendues publiques** par la personne concernée

  Note : cette exception doit être interprétée de façon restrictive, cf. p. ex. l'avis 5/2009 du 12/6/2009 du G29 sur les réseaux sociaux en ligne

- le traitement est nécessaire à des fins archivistiques dans l'intérêt public, à des fins de **recherche scientifique ou historique** ou à des fins **statistiques** mais sur le **fondement du droit** de l'UE ou des États membres (c10,c52) entre autres conditions comme la **proportionnalité** à la finalité

Et lorsque : l'exécution des obligations et de l'exercice des droits propres au responsable du traitement ou à la personne concernée ; la sauvegarde des intérêts vitaux de la personne concernée ou d'une autre personne physique ; association ou tout autre organisme à but non lucratif et poursuivant une finalité politique, philosophique, religieuse ou syndicale (. . .)

Consentement de la personne concernée

Définition : toute manifestation de volonté, libre, spécifique, éclairée et univoque par laquelle la personne concernée accepte, par une déclaration ou par un acte positif clair, que des données à caractère personnel la concernant fassent l'objet d'un traitement (RGPD art. 4 § 11)

- **« manifestation » :** pas de consentement tacite, le responsable de traitement doit pouvoir **démontrer** que la personne a donné son consentement (RGPD art. 7 § 1)

  Exemple : le fait qu'une personne ait répondu à un entretien ou à un questionnaire **ne suffit pas** pour attester du consentement (c32, c42)

- le consentement doit en effet être **éclairé** :

  le responsable de traitement doit pouvoir attester qu'un certain nombre **d'informations** ont été fournies à la personne comme la finalité du traitement, identité du responsable de traitement,. . . (cf. information des personnes)

- **Exemples :**

  - questionnaire : formulaire de consentement (bloquant) avant le questionnaire
  - entretien : selon les cas enregistrement oral ou signature

Consentement de la personne concernée

De plus,

- avec le **RGPD**, le consentement doit être **distinct** des autres questions (p. ex. **CGU**)
- il ne peut y avoir de **consentement global**, la personne doit consentir explicitement à chaque traitement s'il y a plusieurs (c32)
- la personne concernée peut **retirer** sont consentement **à tout moment**

> *toutefois, le retrait du consentement « ne compromet pas la licéité du traitement avant retrait »* (RGPD art. 7 § 3)

Note : les traitement concernant **les enfants** font l'objet de dispositions spécifiques (RGPD art. 8) et requièrent notamment le consentement du tuteur légal

Principes de proportionnalité et de pertinence

RGPD art. 5 § 1 (c) : 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 (minimisation des données)

- seules les données **directement en lien** et **strictement nécessaires** à la réalisation finalité du traitement peuvent être recueillies
- le type de données à caractère personnel qui va être collecté doit donc être **motivé** et justifié au regard des objectifs poursuivis

Ces deux principes sont généralement interprétés d'une façon très **restrictive** :

- on parle alors de **minimisation** des données
- en pratique, c'est un des aspects les plus **délicats** de l'application de la réglementation aux sciences sociales (cf. *infra* p. 80)

Licéité, loyauté et transparence de la collecte des traitements

RGPD art. 5 § 1 (a) : Les données à caractère personnel doivent être [. . .] traitées de manière licite, loyale et transparente au regard de la personne concernée (licéité, loyauté, transparence)

- conditions de **licéité** du traitement (RGPD art. 6) :
  - le traitement est nécessaire à l'exécution d'**une mission d'intérêt public**, comme la recherche ou l'enseignement
  - **autres conditions :** consentement, exécution d'un contrat, obligation légale, sauvegarde des intérêts vitaux de la personne, . . .

  Note :
  - la licéité est une condition nécessaire mais **non suffisante**
  - une fin de recherche **ne suffit pas** *en soi* à rendre un traitement conforme
- **loyauté et la transparence** : la personne concernée doit être informée de l'existence du traitement et de ses finalités (c60) ainsi que de ces droits

Information des personnes

La loyauté et la transparence du traitement impliquent notamment **l'information des personnes** (c39) :

- les personnes doivent en effet être en mesure de décider de l'utilisation de leurs données (principe **d'autodétermination informationnelle**)
- le responsable de traitement doit donc fournir **différentes informations** aux personnes concernées (**RGPD art. 13 § 1**) :
  - l'identité du responsable de traitement, des destinataires de données
  - la finalité du traitement
  - la durée de conservation
  - la liste de ses droits (cf. droits des personnes)

Note : il peut être envisageable **de ne pas décrire précisément** la recherche dans le cas de traitements des données à caractère personnel à des fins de recherche scientifique (c33)

Droit des personnes

Les personnes concernées ont un droit :

- **d'accès** (RGPD art. 15)
- **de rectification** (RGPD art. 16)
- **d'effacement** (RGPD art. 17)
- **de limitation** (RGPD art. 18)
- **d'opposition** (RGPD art. 21)
- d'introduire **une réclamation** auprès d'une autorité de contrôle (RGPD art. 77)
- ainsi que la notification en cas de modification (RGPD art. 19) et le droit à la portabilité des données (RGPD art. 20)

Notes :

- en cas de traitements à des fins de recherche scientifique ou historique ou à des fins statistiques, **l'UE ou les États** peuvent peuvent prévoir **des dérogations** aux droits d'accès (art. 15), de rectification (art. 16), à la limitation du traitement (art. 18), de de modification (art. 19), de portabilité (art. 20) et au droit d'opposition (art. 21) (RGPD art. 89 § 2)
- le droit à l'effacement ne s'applique pas si la mesure est susceptible de compromettre gravement la réalisation des finalités (RGPD art. 17)

Collectes indirectes

La collecte n'est pas toujours réalisée **directement** auprès de la personne :

- **Exemples :** fouille (archives, internet, base de données,...), entretiens,...

Note : tout ce que est en **libre accès** n'est pas nécessairement **libre de droits** :

cf : *CGU, licences, droit des base de données,...*

Dans ce cas,

- le responsable de traitement est là aussi soumis à une obligations **d'information** des personnes (RGPD art. 14 § 1)
- de plus, les 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)

Note : la réglementation **ne distingue pas** des « personalités publiques »

Collectes indirectes

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

- information impossible ou exigeant 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

Note : ceci ne constitue pas **un blanc-seing**, il faut bien évidemment **motiver** l'application de ces exceptions

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

La conservation et la réutilisation des DCP

Rappel : les données ne doivent être collectées que pour des finalités déterminées, explicites et légitimes, et ne pas être traitées ultérieurement d'une manière incompatible avec ces finalités **(limitation des finalités)**

De façon corrélative,

RGPD art. 5 § 1 (e) **:** la conservation est limitée à la durée nécessaire à la réalisation des finalités du traitement

- à l'issue de cette période le responsable de traitement doit, soit **détruire** l'ensemble des données, soit les rendre complément **anonymes**

  Notes :

  - la destruction doit être être **autorisée** par les archives nationales ou départementales
  - attention aux données **indirectement identifiantes** qui peuvent se révéler très difficiles à anonymiser

- la conservation **au-delà** de cette durée est néanmoins possible pour les fins de recherches scientifiques et historiques ou à des fins statistiques (RGPD art. 5 § 1 (e))
- pour autant que **les mesures techniques et organisationnelles** appropriées soient prises pour respecter le principe de minimisation des données
- la conservation est toutefois **distincte** de la réutilisation

La réutilisation des DCP à des fins scientifiques

La LIL prévoit qu'il peut être procédé à des traitements poursuivant une autre finalité :

- si la personne y a **consenti**
- après **autorisation** de la CNIL

Le **RGPD** prévoit que :

- un traitement ultérieur à des fins historiques, statistiques ou scientifiques **« n'est pas réputé incompatible »** (RGPD art. 5 § 1 (b))
- pour autant que, là aussi, **les mesures techniques et organisationnelles** appropriées soient prises pour respecter le principe de minimisation des données

  > le responsable de traitement doit ainsi évaluer s'il est possible d'atteindre ces finalités grâce à un traitement de données qui ne permettent pas ou plus d'identifier les personnes concernées (c156))

- et si et seulement si le traitement sert **uniquement** une finalité de recherche (RGPD art. 89 § 4)
- pour autant, les personnes concernées ont toujours **des droits**

Modalités et agents de la protection des données

Le responsable de traitement

Définition : la personne physique ou morale, l'autorité publique, le service ou un autre organisme qui, seul ou conjointement avec d'autres, détermine les finalités et les moyens du traitement (RGPD art. 4 § 7)

- le responsable de traitement n'est pas nécessairement une personne physique
- le responsable de traitementest soumis à différentes obligations :
  - le responsable du traitement met en œuvre des **mesures techniques et organisationnelles appropriées** pour s'assurer et être en mesure de démontrer que le traitement est effectué conformément au RGPD (RGPD art. 24 § 1)
- le responsable de traitement est de plus **responsable pénalement**

Note : le fait que le responsable de traitement soit responsable pénalement ne signifie pas que la responsabilité des différentes catégories de personnels ne puisse pas être engagée à un titre ou un autre

Le responsable de traitement dans l'ESR

Dans le cadre de l'ESR, le responsable de traitement d'un traitement n'est (généralement) **pas** le ou les **(enseignants-)chercheurs** :

- en pratique, les responsables de traitement peuvent varier selon les activités
- **enseignements :** chef d'établissement (p. ex. le président de l'université)
- **recherche :** le directeur de l'entité dont dépend le chercheur (UMR)

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

Formalités préalables

Parmi les obligations du responsable de traitement, la LIL impose que :

- si le traitement comporte des **DCP**, il doit faire l'objet de **formalités** (déclarations, autorisations) **avant** la la mise en œuvre du traitement
- les formalités doivent être réalisées auprès de la CNIL ou d'un CIL pour une large partie d'entre elles

Le **RGPD** **supprime (partiellement) cette obligation** :

- le **RGPD** considère en effet que :

  « cette obligation [générale de notifier les traitements de données à caractère personnel aux autorités de contrôle] génère une charge administrative et financière, **sans pour autant avoir** systématiquement **contribué à améliorer la protection des données à caractère personnel** » (c89)

- cependant, toutes les formalités préalables **ne seront pas amenées à disparaître** (p. ex. pour les données relatives aux infractions et aux mesures de sûreté)
- en partie laissé à l'appréciation des États

La responsabilisation

- la contrepartie de la suppression des formalités préalables est **l'inversion de la charge de la preuve** :

  *désormais, il incombera donc au **responsable de traitement** de démontrer qu'il est en conformité avec le règlement (RGPD art. 24 § 1)*

- le responsable de traitement doit tenir **un registre** actualisé de traitement des données (RGPD art. 30 § 1)

  ce registre comporte les informations suivantes : nom et les coordonnées du ou des responsables du traitement, les finalités, description des catégories de personnes concernées et des catégories de données à caractère personnel, catégories de destinataires, délais de conservation, description des mesures de sécurité

- ce registre peut être tenu par son représentant, **le** CIL

Protection des données dès la conception et par défaut

Parmi les (nouvelles ?) obligations du responsable de traitement figurent aussi :

- **la protection des données dès la conception (RGPD art. 25 § 1) :** le responsable de traitement doit mettre en œuvre toutes les mesures techniques et organisationnelles nécessaires au respect de la protection des données personnelles **dès la conception** du traitement
- **la protection des données par défaut (RGPD art. 25 § 2) :**
  - cf. finalité : le responsable de traitement doit mettre en œuvre toutes les mesures pour que seules les données **strictement nécessaires** à la réalisation de la finalité soient traitées **par défaut**, -ie : sans intervention de la personne concernée
  - ces mesures doivent garantir que seules **les personnes habilitées** accèdent aux données

Note : au delà des obligations réglementaires, l'expérience montre que la mise en conformité en cours de route est souvent impraticable (ex : collecte directe de données sensibles sans demande du consentement)

Analyse d'impact relative à la protection des données

RGPD art. 35 § 1 : lorsqu'un type de traitement, en particulier par le recours à de nouvelles technologies [...] est susceptible d'engendrer un risque élevé pour les droits et libertés des personnes physiques, le responsable du traitement effectue, avant le traitement, **une analyse de l'impact** des opérations de traitement envisagées sur la protection des données à caractère personnel

- disposition introduite par le **RGPD**
- requise « particulièrement » pour :
  - le traitements de **données sensibles** (RGPD art. 35 § 3 (b))
  - les traitements **« à grande échelle »** (p. ex. sur les réseaux sociaux)
  - ou le traitements de données se rapportant à **des condamnations ou des infractions**
- **des listes** rendant obligatoire ou dispensant de l'analyse doivent être être dressées par les autorités de contrôle (RGPD art. 35 § 4 et art. 35 § 5)

  *si l'analyse révèle un risque particulièrement élevé, l'autorité de contrôle doit être **consultée***
- la CNIL et le G29 ont publié des guides pour réaliser ce type d'études

Remarques

- d'un certain point de vue, la protection des données dès la conception et les études d'impact ne sont pas des nouveautés
- ces mesures étaient en quelque sorte **implicites** dans la LIL

  en pratique, la réalisation des formalités préalables implique d'anticiper les éventuels risques pour les personnes concernées par le traitement

- les études d'impact illustrent de plus la spécificité de la réglementation sur les DCP :
  - la réglementation édicte **des grands principes**
  - **les modalités de son application** telles que la minimisation des données et, plus généralement, les mesures de protection à adopter doivent être déterminées **au regard du traitement**
  - en s'appuyant notamment sur **la doctrine** de la CNIL et **les recommendations** du G29
- manque **d'un réferentiel** propre aux sciences sociales

Le sous-traitant

Définition : la personne physique ou morale, l'autorité publique, le service ou un autre organisme qui traite des données à caractère personnel pour le compte du responsable du traitement (RGPD art. 4 § 8)

- **définition très large :** entreprise à qui la réalisation d'une enquête est sous-traitée mais aussi vacations pour des transcriptions d'entretiens ou encore la prestation de service en ligne

RGPD art. 28 :

- le prestataire doit présenter **des garanties suffisantes**
- le traitement par un sous-traitant est régi par **un contrat ou un autre acte juridique** de l'UE
- l'autorisation de la CNIL est nécessaire si le sous-traitant est établie **en dehors de l'UE**
- le RGPD s'applique même si le sous-traitant n'est **pas établi** sur le territoire de l'UE
- formalités ?

Les obligations du sous-traitant

Le contrat de sous-traitance devra contenir un certain nombre **de dispositions impératives** :

- le sous-traitant ne traite des données personnelles que **sur instruction documentée** du responsable de traitement
- les données ne doivent être traitées **que pour la réalisation de la finalité**
- le sous-traitant doit prendre toutes les mesures appropriées **pour assurer la confidentialité et la sécurité** des données

  Définition : les données contenues dans ces supports et documents sont strictement couvertes par **le secret professionnel** (article 226-13 du code pénal)

- les données doivent **être détruites ou remise** une fois la finalité réalisée (sans conservation de copies)
- le sous-traitant met à la disposition du responsable du traitement toutes les informations nécessaires **pour démontrer le respect des obligations** prévues au présent article et pour permettre la réalisation d'audits, y compris des inspections, par le responsable du traitement ou un autre auditeur qu'il a mandaté, et contribuer à ces audits
- ces obligations doivent **se répercuter** à ses sous-traitants (*ad lib*)

Champ d'application territorial

Le RGPD s'applique si (RGPD art. 3) :

- **le responsable de traitement** -ou son sous-traitant- est établi sur **le territoire de l'UE** (même si les personnes concernées n'y résident pas)
- **les personnes concernées** résident sur **le territoire de l'UE** (même si le responsable de traitement -ou son sous-traitant- n'y est pas établi)

Note :

- le second cas n'était **pas prévu** dans la LIL

  > *la définition par rapport au seul pays du responsable de traitement a parfois pu conduire à des situations. . .cocasses*

- il vise clairement **les** GAFAM ***et. al.***

La Commission nationale informatique et libertés (CNIL)

La CNIL est une **autorité administrative indépendante** crée par la loi de 1978 :

- elle est composée de **18 membres** élus ou nommés principalement issus de différentes instances publiques (Parlement, hautes juridictions de l'État,...) qui sont assistés par près de 200 agents
- la commission dispose d'un pouvoir de **contrôle** et de **sanction** (renforcé par le **RGPD**) mais aussi des missions d'**avis**, de **conseil** et **labellisation**
- elle dispose de plus d'un pouvoir **réglementaire** : la CNIL édicte des normes (normes simplifiées, autorisations uniques, actes réglementaires uniques et méthodologies de référence,...)

Au niveau de l'Union,

- la CNIL est membre du **G29** (Groupe de travail de l'article 29 de la directive 95/46/CE) qui est un organe consultatif de l'UE composé des différentes autorités de protection des données des membres de l'Union
- le **G29** publie régulièrement des avis ainsi que des lignes directrices sur des points précis de l'application de la réglementation

Sanctions

Les infractions à la LIL sont des infractions **pénales** :

- jusqu'à **300 000 d'amendes**
- jusqu'à **5 ans d'emprisonnement**

Note : personne n'est jamais allé en prison sur le fondement de la LIL

Le RGPD **augmente considérablement** le niveau des sanctions financières encourues en cas d'infraction (RGPD art. 83 § 1) :

- jusqu'à **20 millions €**
- ou **4 % du chiffre d'affaires** annuel mondial
- le plus élevé de ces deux montants est retenu

Note : la loi pour une République numérique a déjà porté le plafond à 3 millions €

Sanctions

Le **niveau de sanction** dépend notamment :

- de la nature, gravité et durée de la violation
- du nombre de personnes concernées, du dommage subi, des catégories de DCP concernées
- des violations commises précédemment, des mesures techniques et organisationnelles mises en œuvre,. . .

De plus,

- le RGPD introduit aussi la possibilité d'engager **des actions de groupe** ($\simeq$ *class actions*) en matière de DCP
- la LIL a **déjà été modifiée** en ce sens par la loi de modernisation de la justice du 16 novembre 2016

Le correspondant informatique et libertés (CIL) (futur DPO)

- le CIL a été crée par la modification de 2004 de la LIL en application de la directive européen de 1995 pour prendre en charge une partie des formalités préalables
- le CIL sera remplacé par le **délégué à la protection des données** (**DPO**) à l'entrée en vigueur du **RGPD**
- les fonctions du **DPO** (**RGPD art. 39**) :
  - **informer** et **conseiller** le responsable de traitement
  - **contrôler** le respect du règlement
  - **coopérer** avec l'autorité de contrôle et faire office de point de contact pour l'autorité de contrôle sur les questions relatives au traitement

    Note : le **DPO** n'en est pas pour autant une émanation de la CNIL
- le **DPO**, représentant du responsable de traitement, tient à jour **un registre des traitements** (**RGPD art. 30 § 1**)

Note : cf. *supra* **responsabilisation** et **inversion de la charge de la preuve** p. 58 et suivantes

La désignation du DPO

La désignation du **DPO** est obligatoire dans les cas suivant (**RGPD art. 37 § 1**) :

- 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)
- les activités de base du responsable du traitement ou du sous-traitant consistent en des opérations de traitement qui [...] exigent **un suivi régulier et systématique** à grande échelle des personnes concernées
- les activités de base du responsable du traitement ou du sous-traitant consistent en un traitement à grande échelle de **données sensibles**

Pour **les UMR CNRS-université**, la désignation du CIL doit se faire en fonction de **l'employeur** du **DU** (cf. courrier du 4 septembre dernier de la **CPU** et du **CNRS**) :

- si le **DU** est personnel université, il faut désigner le CIL de l'université
- si le **DU** est personnel **CNRS**, il faut désigner le CIL du **CNRS**

Note : pour les **DU** non-**CNRS**, si le CIL de l'employeur ne peut exercer cette mission, le CIL du **CNRS** peut être nommé à sa place

Le CIL dans vos projets de recherche

Le CIL dans vos projets de recherche :

- l'application de la réglementation peut impacter **ce que vous pouvez collecter** et **la façon** dont vous pouvez le collecter et le traiter
- le **RGPD** renforce de plus considérablement les obligations du responsable de traitement
- l'association de votre CIL à vos projet de recherche est plus que jamais **cruciale**
- et ce, dès **la conception du projet**

La mise en œuvre de la réglementation dans les traitements en sciences sociales

Les grands principes

- **limitation de la finalité :** les données doivent être traitées de façon **compatible** avec une finalité **précise**
- **minimisation des données :** seuls 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**
- **information :** les personnes doivent être en mesure de **décider** de l'utilisation des informations les concernant
- **protection dès la conception (*privacy by design*) :** la protection des personnes et la sécurité des données doit être intégrée **dès la conception** du traitement

Remarques préalables

Quelques remarques préalables sur la démarche à adopter :

- d'abord, **désigner le CIL** si ce n'est pas déjà fait
- ensuite, déterminer si le traitement nécessite **de collecter des DCP**

  Note :

  - considérer les données comme non identifiantes n'est généralement **pas la meilleure des stratégies** (cf. *infra*)

    il vaut mieux partir sur l'idée que les données sont identifiantes et établir par la suite qu'elles ne le sont pas ***que l'inverse***

  - il est préférable consulter son CIL de toute façon, les données indirectement identifiantes rendant très souvent les traitements **nominatifs**

- être en mesure décrire **le plus précisement possible** le projet sous tous ces aspects

  le diable est toujours dans les détails. . .

- déterminer **le ou les responsables de traitement**

  Exemple : lorsque le traitement implique différents partenaires, académiques ou non

La licéité du traitement

Le traitement doit avant tout répondre à différents **grands principes** comme, en premier lieu, **la licéité** :

- **condition de licéité :** le traitement est nécessaire à l'exécution d'**une mission d'intérêt public**

  Exemple : l'enseignement, la recherche

- il s'agit toutefois d'une condition nécessaire mais **non suffisante**

D'autres obligations doivent être respectées :

- la finalité du traitement doit être **déterminée**, **explicite** et **légitime**

  -ie : la problématique de la recherche doit être clairement définie

- les données traitées doivent être **proportionnées** et **pertinentes** au regard de la finalité du traitement
- les données doivent être collectées et traitées de manière **loyale** et **transparente**
- ainsi que d'autres obligations dans le cas du traitement de **données sensibles**
- ...

Les DCP dans les enseignements de méthodes

Les DCP dans les enseignements de méthodes :

- l'enseignement est une mission de **service publique**, la collecte de DCP dans le cadre d'enseignements est donc **licite**
- il s'agit toutefois d'une condition nécessaire mais **non suffisante**
- du fait qu'il s'agit d'un apprentissage à la recherche, l'analyse est **la même** que pour la recherche :
  - une finalité « enseignement » n'est **pas suffisamment précise** pour décrire le traitement
  - comme dans le cadre de la recherche, les enquêtes peuvent être là aussi **très diverses**
  - donc pas de possibilité d'enregistrement unique
- en pratique, il faut donc enregistrer **toutes les enquêtes** réalisées dans le cadre d'enseignements

Exemple : si les étudiants d'un TD se réunissent en sous-groupes et choisissent un thème, le traitement de chacun des groupes devra faire l'objet d'un enregistrement

La finalité

- la finalité correspond à **la problématique** de la recherche (et pas la thématique ou la question de recherche)
- la finalité du traitement (et donc la problématique doit être **déterminée**, **explicite** et **légitime**
- vous devez déterminer à l'avance ce que vous voulez démontrer et comment, c-à-d quelles DCP sont nécessaires à la démonstration et pourquoi elles sont nécessaires
- il faut donc **formuler** toutes vos hypothèses *a priori*

  Note : traitement prosopographiquee est un oxymore

- **une finalité par traitement**, l'utilisation de données à d'autres fins que celles prévues est une infraction

  Note : toutefois, une exception est prévue pour les traitements ultérieurs à fin de recherche

Fins statistiques et fins de recherche scientifique ou historique

Pour les traitement à fins statistiques et fins de recherche scientifique ou historique, différentes exceptions sont prévues dans le RGPD :

- l'information des personnes peut éventuellement **ne pas être complète** lors de collectes directes
- l'obligation d'information peut même être éventuellement partiellement **allégée** dans le cas de collectes indirectes
- **des données sensibles** peuvent être collectées moyennant, p. ex., le consentement
- les données collectées peuvent être **archivées**
- les données peuvent être **réutilisées** et ce, même si elles n'ont pas été collectées pour une finalité scientifique
  - pour autant que soient mises en œuvre les mesures techniques et organisationnelles appropriées requises par le règlement afin de garantir les droits et libertés de la personne concernée
  - et que les personnes en soit informées

Fins statistiques et fins de recherche scientifique ou historique

Toutefois,

- ces dispositions ne consituent pas un **blanc-seing**
- elles doivent être **motivées**

Surtout,

- ces dispositions doivent pour partie encore faire l'objet de **clarifications** dans la perspective de l'entrée en application du RGPD

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 DCP 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 DCP 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

Principes de proportionnalité et pertinence

RGPD art. 5 § 1 (c) : 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 (minimisation des données)

- 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

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 DCP :

- 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

Principes de proportionnalité et pertinence

À mon avis,

- il manque encore **un étalonnage** spécifique pour l'appréciation de la proportionnalité et de la pertinence des traitements en sciences sociales
- les termes (plus ou moins explicites) de l'appréciation reposent actuellement sur des cas souvent très éloignés des sciences sociales

  Exemple : les délibérations de la CNIL
  - les délibérations portent essentiellement sur des traitements réalisés par **des entreprises** ou par **le public** (gouvernement, État, administrations, collectivités,. . .)
  - les délibérations concernant la recherche relèvent quasi exclusivement **la recherche médicale**
- les sciences sociales sont en effet quasi **absentes** des délibérations de la CNIL
  - quatre délibérations $(3 + 1)$ concernant les traitements de deux **UMR**
  - cf. prophétie auto-réalisatrice
- cette rareté impacte **l'analyse juridique** des traitements

  ce qui peut notamment avoir pour effet des mesures pouvant parfois paraître disproportionnées

Les responsables de traitement dans les délibérations de la CNIL

Les finalités des traitements en sciences sociales

À mon avis,

- l'appréciation de la proportionnalité et de pertinence des traitements pose, plus généralement, la question **des finalités** des recherches en sciences sociales
- les finalités en sciences sociales **diffèrent** des finalités des entités (entreprises, administrations, associations...) qui constituent le gros des délibérations de la CNIL :
  - les sciences sociales n'ont pas directement à faire à des administrés, des assurés sociaux, des usagers, des employés, des clients,... mais bien à **des enquêtés**
  - généralement, les traitements n'utilisent pas les DCP collectées pour prendre **une décision** sur ces personnes concernées
  - ou les **affecter** de façon significative
- les traitements en sciences sociales ne visent souvent des personnes physiques que pour mieux **s'en abstraire**
- et produire au final des discours **de portée générale** non contingentés à un échantillon ou un autre

La protection des données

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 **DCP** ») 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 responsable de traitement

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 (cf. *infra* p. 96)

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 :
  - répond à une recherche **« d'équivalence »** (par rapport au sexe, à l'âge, . . .) (Coulmont, 2017)
  - les pseudonymes contiennent donc des informations pouvant concourir à **la réidentification des personnes**

La pseudonymisation

- la définition de la pseudonymisation renvoie implicitement au traitement de DCP 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** (cf. *infra*)

La protection des données en pratique

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 **DCP** ou de s'en débarrasser (moins de **DCP**, moins de contraintes)

La protection des données en pratique

Sécurité au niveau applicatif :

- chiffrement **des connexions** (p. ex. à des serveurs `http`, `ftp`, de données,...) : `TLS`, `VPN`,...
- certaines données ne devraient être accessible que depuis **un réseau local**, voire **pas accessibles du tout**...
- pseudonymisation des données des base de données :
  - pseudonymisation des clefs primaires et secondaires si elles contiennent des **DCP**
  - stockages séparés des **DCP**

    Note : gestion des invitations à un questionnaire en ligne distincte de la gestion des réponses
  - cf. l'avis `0829/14` du **G29** du 04/05/2014 sur les techniques d'anonymisation
- et aussi renoncer **aux services « gratuits »** pour y substituer les services recommandés par vos institutions

Remarque : algorithmes et systèmes cryptographiques

Il faut bien distinguer **les systèmes** (protocoles,...) utilisant la cryptographie des **algorithmes cryptographiques** proprement dits :

- un même protocole peut utiliser **plusieurs algorithmes** en **les combinant** ou **en proposant plusieurs choix**

  Note : cette distinction est avant tout **heuristique**, l'articulation entre les différents éléments constitutifs de la sécurisation informatique étant beaucoup plus complexe

- **Exemples d'algorithmes :** `DES` (obsolète), `MD5` (obsolète), `SHA-1` (obsolète), `SHA-2`, `RSA`, `AES`, `A5/1`,...

  Note : les algorithmes reposent eux-mêmes sur des « primitives », cryptographiques ou non :

  exponentiation modulaire dans un corps fini $\mathbb{F}_p$ avec p prime, fonctions de hachage, générateurs de nombres aléatoires de qualité cryptographique,...

Remarque : algorithmes et systèmes cryptographiques

- **Exemple de système : `HMAC` (*keyed-Hash Message Authentication Code*)**
  - fonction de hachage cryptographique à clef secrète utilisée pour garantir l'intégrité des données et authentifier un message
  - repose sur une fonction de hachage cryptographique au choix, y compris `MD5` ou `SHA-1` :

$$HMAC(K, texte) = H(\ (K \oplus opad) \ ||\ H((K \oplus ipad) \ ||\ texte)\ )$$

    avec H une fonction de hachage itérative, K une clef secrète

- **Exemple de système : `TLS` (*Transport Layer Security*)**
  - `TLS` combine cryptographie asymétrique et cryptographie symétrique
  - la cryptographie asymétrique permet de transférer les clefs qui serviront à chiffrer les échanges entre le client et le serveur
  - aux différentes étapes de l'établissement de la connexion, différents types d'algorithmes peuvent être proposés par le serveur au client

- ainsi que `PGP`, `FTPS`, `blockchain`,...

Exemple de sécurisation des DCP

Exemple : enquête par **questionnaires en ligne**

- différents gestionnaires de questionnaire peuvent aussi assurer **l'envoi des invitations**
- ils doivent donc avoir accès à des DCP comme **l'adresse des répondants**
- si la sécurité 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 DCP peuvent être remplacées par un identifiant permettant de faire le lien entre (non-)réponses et données auxiliaires

En pratique,

- il faut générer **deux clefs** : une clef privée pour les données auxiliaire et une clef publique pour les traitement (au cas où les données auxiliaires seraient aussi compromises)
- 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** avant l'attribution (sinon le nombre correspondra à la ligne et l'ordre permettra la réidentification)

Identification, désidentification et réidentification des personnes

Identification des personnes

- l'application de la réglementation ne se limite pas à la stricte question de **anonymat** (le mot est d'ailleurs quasi absent du RGPD)
- le RGPD traite de la question plus large de **l'identification** des personnes :
  - la protection des personnes n'est pas liée à **un état** : être ou ne pas être anonyme
  - mais à **des actions** et à **des situations** : pouvoir réidentifier une personne à des degrés divers à un moment donné

    *les possibilités de réidentification **sont variables** selon les situations et dans le temps (p. ex., les informations dont dispose a priori un attaquant peuvent varier)*
- là comme ailleurs, **le point de vue crée l'objet** :
  - la protection des données nécessite de modéliser les relations entre un attaquant et les personnes concernées
  - d'où **la contingence** de l'analyse juridique
  - mais aussi des **mesures de sécurité** qui se limitent pas à la pseudonymisation ou au chiffrement

Identification des personnes

- le développement **des techniques d'identification** des personnes est un phénomène ancien
- elles constituent **un champ de recherche** toujours plus actif

Exemples :

- la biométrie : empreintes digitales, visage, rétine, réseaux veineux de la main, `IRM` (imagerie par résonance magnétique), . . .)
- la sociométrie (?) : caractéristiques et pratiques sociales à partir de `BdD`

- **les techniques de réidentification** des données anonymisées ou pseudonymisées et **la protection contre la réidentification** constituent elles aussi un champ de recherche actif
- de tous ces travaux, il ressort notamment que :

  - **tout laisse une empreinte** (ou, plus exactement, tout peut être **utilisé** comme empreinte)
  - **la quantité d'information** nécessaire à la (ré)identification des personnes n'est souvent pas **très élevée**
  - l'identification des personnes est en effet facilitée par **l'hétérogénéité** (biologique, sociale. . .) des populations
  - la désidentification est **difficile** sans porter préjudice à la rediffusion (cf. *Open Data*)

Empreinte digitale d'appareil

Exemple : la prise **d'empreinte digitale d'appareil** (*device fingerprint*) (Nikiforakis et al., 2013)

- par analogie à la biométrie, désigne un ensemble de techniques permettant de **pister** p. ex. la navigation d'une personne sur internet (connexions multiples à un même site mais aussi entre sites) mais sans laisser de traces sur la machine de l'utilisateur (`cookies`)
- cette technique est lié au développement d'un web toujours plus **interactif et dynamique**
- lorsque `JavaScript` est activé sur un navigateur, la page chargée peut récupérer **un très grand nombre d'informations** de façon passive ou active :
  - matérielles : taile et résolution de l'écran, caractéristiques de la carte graphique via `WebGL`,
  - logicielles : `OS`, navigateur, protocoles supportés, modules installés
  - autres : fuseau horaire, langue, polices, exécution cachée de code (canva -rendu 2d ou 3d-)

Note : dans de nombreux cas, l'`IP` n'est pas suffisante pour identifier et pister des internautes. Elle est néanmoins considérée comme une données à caractère personnel par la Cnil car elle peut aussi permettre d'identifier des personnes par recoupement

Empreinte digitale d'appareil

- prises **séparément**, ces informations ne semblent pas identifiantes car elles peuvent correspondre à des millions d'utilisateurs
- combinées, elles peuvent pourtant identifier un appareil avec **une forte probabilité**
- une étude de l'**EFF** (Eckersley, 2010) a par exemple montré que sur 1 millions de visites sur une page dédiée de leur site, **83.6 % des navigateurs étaient uniques**
- plusieurs sites proposent de calculer l'empreinte de votre navigateur comme **panopticlick** (dont est issu l'étude de l'**EEF**) ou celui du projet **AmIUnique** de l'Inria Rennes - Bretagne Atlantique (Laperdrix, Rudametkin et Baudry, 2016)

Note : le test enregistre des informations collectées via votre navigateur pour alimenter la base données du projet

- les techniques d'empreintes digitale d'appareil sont de plus en plus utilisées pour pister les internautes et plusieurs entreprises proposent ce type de service
- elles sont beaucoup **plus difficiles à contrer** que, p. ex., les `cookies`
- l'accès à ces informations se fait généralement **à l'insu des utilisateurs** et donc en infraction avec la directive 2002/58/EC
- le **G29** a publié un avis qui qualifie les empreintes digitales d'appareil comme des traitement de **DCP** (avis `9/2014` du 25/11/2014) + wp247

De plus,

- les techniques d'empreintes digitale d'appareil montrent comment la disposition d'informations *a priori* **peu discriminantes** prises singulièrement peuvent identifier des personnes physiques par leur combinaison
- certaines informations, sans être directement identifiantes, présentent, dans les faits, **une forte entropie**
- surtout, la **combinaison** de ces différentes informations présente souvent une **entropie suffisante** pour ce condensat soit unique et permette donc d'identifier un utilisateur

Entropie

Notion fondamentale en **sécurité des systèmes d'information** (chiffrement, génération de nombres (pseudo) aléatoires,...) mais aussi pour les questions de **ré-identification** à partir de données **non directement identifiantes**

- l'entropie est une mesure proposée par Claude Shannon dans le cadre de **la théorie mathématique de l'information** qu'il a contribué à fonder
- C. Shannon travaillait dans le domaine des télécommunications
- dans sa thèse, il s'est notamment intéressé à la transmission de codes via un canal perturbé de façon à ce que cela ne conduise pas à une perte d'informations

> *The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.* (Shannon, 2001)

- information n'est pas ici entendue au sens **sémantique** du terme
- la théorie mathématique de l'information ne s'intéresse qu'au contenant du signal lui-même, **pas ce qu'il contient ou signifie**

> *Frequently the messages have meaning ; that is, they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages.* (Shannon, 2001)

Note : l'entropie des caractéristiques sociales est ici différente de l'entropie sociale de T. Parsons qui est utilisée par analogie à l'entropie en thermodynamique

Entropie de l'information de Shannon

L'entropie de l'information de Shannon :

- l'information est conçue comme étant stockée ou transmise par une **variable aléatoire** qui peut prendre différentes valeurs comme les lettres d'un alphabet
- intuitivement, l'entropie sert à mesurer **la quantité d'information** que contient cette variable

Interprétation de l'entropie de Shannon :

- l'information (« surprise ») moyenne de la variable
- plus petit nombre de `bits` nécessaires en moyenne pour coder un message m (ou nombre de question oui-non pour déterminer l'information complète)
- mesure de redondance ou d'imprévisibilité
- autant d'interprétations que de domaines d'application

Exemples :

- plus l'entropie d'un fichier sera faible, plus il sera facile à compresser
- plus l'entropie d'un mot de passe sera forte, plus il sera robuste

Note :

- ce n'est pas la longueur en soi d'un mot de passe qui compte mais son **entropie**
- ce critère n'est toutfois pas suffisant pour garantir la sécurité du mot de passe (l'entropie ne prend pas en compte l'aspect **sémantique** de l'information)

Entropie de l'information de Shannon

- plus formellement, mettons qu'un message soit encodé avec un alphabet comportant n symboles
- l'entropie de la variable aléatoire discrète correspondante $X = x_1, \ldots, x_n$ peut être définie comme l'espérance de l'information contenue par X, $\mathbb{I}(X)$:

$$H(X) = \mathbb{E}[\mathbb{I}(X)] = \mathbb{E}[-log_b(p(x)]$$

avec $p(x_i) = Pr(X = x_i)$ la densité de X

- elle a pour forme (en base $b = 2$) :

$$H(X) = \sum_{i=1}^{n} p(x_i)\mathbb{I}(X) = \sum_{i=1}^{n} p(x_i) log_2 \left(\frac{1}{p(x_i)}\right) = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i) \quad (1)$$

- l'entropie est donc calculée par la somme de l'information contenue par chaque symbole pondéré par la probabilité d'apparition des symboles, soit **l'information moyenne** de la source

Note : le terme d'entropie a été suggéré par J. von Neuman à C. Shannon de par la relation de (1) avec l'entropie de Boltzmann et parce que « personne n'y comprenait rien »

Entropie de l'information de Shannon

- $\mathbb{I}(X) = log_b(1/p_i)$ sert de mesure du **contenu** de l'information d'un symbole
- $\mathbb{I}(X) = f(1/p_i)$: l'idée sous-jacente de l'utilisation de l'inverse de $p(x)$ comme mesure de l'information contenue est que,
  - **plus la probabilité d'un événement est faible, plus il est intéressant**
  - et plus la probabilité de son occurrence **contient d'information**
  - on parle aussi de « surprise » car plus un événement est rare, plus il est surprenant

Exemple : pour l'empreinte digitale d'appareil, un OS de type GNU-LINUX est plus rare et contribue plus à l'unicité du profil que, p. ex. WINDOWS 7. Cette modalité contient donc plus d'information que les autres modalités de la variable pour, p. ex. identifier une personne.

- $\mathbb{I}(X) = f(1/p_i)$: la fonction logarithmique est utilisée car elle confère à la mesure un certain nombre **de propriétés intéressantes**

$\mathbb{I}(X) = f(1/p_i)$

Analogie entre théorie de l'information et **théorie des sondages** :

- dans les deux cas, on cherche **à minimiser le coût** et **maximiser la qualité des données**
- **information :** minimiser la perte d'information
- **sondage :** minimiser la variance d'échantillonnage, c-à-d la variance due au plan de sondage $p(s)\ \forall\ s \in U$

Lors du tirage d'un échantillon,

- chaque unité similaire à celles précédemment tirées au hasard apporte **moins d'information** (redondance) alors que son coût, lui, ne décroît généralement pas

  Note : similarité du point de vus de $p(s)$

- le gain d'information à chaque tirage est donc fonction **du nombre d'unités similaires** ayant déjà eu lieu
- il est donc **inversement proportionnel** à la probabilité d'occurence

$\mathbb{I}(X) = f(1/p_i)$

- le gain d'information à chaque tirage est aussi fonction de **la probabilité d'occurrence** de ce type d'unité dans la population échantillonnée
- les unités sont plus ou moins rares et les unités les plus rares peuvent être plus intéressantes car **plus atypiques**
- plus généralement, on souhaite donc orienter le tirage en fonction de l'information qu'ils contiennent (p. ex. en fonction d'une mesure de taille)

  Exemples : le nombre de salariés d'une entreprise, le revenu d'un ménage ou la proportion de chômeurs à l'îlot

  Note : toutes ces variables sont asymétriques, la taille est donc liée à la rareté

- pour cela, la théorie des sondage permet d'intégrer des informations auxiliaires sous la forme **de poids de sondages** π_k qui permettent de faire varier la probabilité de sélection des unités correspondantes

Il peut être montré que les plans de sondage à probabilités inégales suivants maximisent l'entropie :

- le plan de Poisson (tillé2001)

$$p(s) = \left\{ \prod_{k \in s} \pi_k \right\} \left\{ \prod_{k \in s} (1 - \pi_k) \right\}, \quad s \in U$$

avec les contraintes $\sum_{\substack{s \subset U \\ s \ni k}} p(s) =_p i_k$ et $\sum_{s \subset U} p(s) = 1$

- le plan à taille fixe et à entropie maximale (tillé2001)

$$p(s) = exp \left\{ \sum_{k \in \mathcal{S}} \lambda_k - \mu(n) \right\}, s \in \mathscr{S}_n$$

avec $\mathscr{S}_n = \{s | \#s = n\}$ et $\mu(n) = \sum_{s \in \mathscr{S}_n} exp \sum_{l \in s} \lambda_k$

- ainsi que la plan de Pareto, de Sampford,...

$\mathbb{I}(X) = log_b(1/p_i)$

- $log_b(x)$ donne le nombre **de `bits` nécessaires** pour encoder x en base b
- **Exemples :**

Encodage ascii :

- les caractères ascii sont encodés sur 7 `bits`, soit

$$2^7 + 2^6 + 2^5 + 2^4 + 2^3 + 2^2 + 2^1 + 2^0 = 1111111_2 \text{ (base 2)}$$
$$= 255_{10} \text{ (base 10)}$$

- la valeur maximale vaut donc $255_{10} = 2^8 - 1$
- opération inverse : $log_2(256) - 1 = 7$ donne donc le nombres de `bits` nécessaire pour encoder 255 symboles

Entropie d'un tirage à pile ou face :

- on tire à pile ou face (avec une pièce équilibrée)
- $p(x) = 1/2$
- l'information contenue vaut donc $log_2(1/.5) = log_2(2)$ soit (intuitivement) **1 `bit` d'entropie**
- elle est de plus égale à l'entropie de la variable :

$$H(X) = (1/2)log_2(1/.5) + (1/2)log_2(1/.5) = log_2(1/.5)$$

qui se trouve elle-même égale **à l'entropie maximale** de cette variable ($log_2(2) = 1$)

$\mathbb{I}(X) = log_b(1/p_i)$

- **positivité :** $H \geq 0$, du fait que, par définition, $0 \leq p(x)) \leq 1$ et donc $-log_b(x) \geq 0$
- **additivité :** on **souhaite** que la mesure de l'information sur un événement ω_i soit additive

$$\mathbb{I}_{tot} = \mathbb{I}(x_1) + \mathbb{I}(x_2) + \ldots + \mathbb{I}(x_n)$$

  - on cherche donc une fonction f de la probabilité de ω_i qui satisfasse cette contrainte : $\mathbb{I}(\omega_i) = f(P[\omega_i])$
  - soit un message C composé de deux événements indépendants A et B

$$\mathbb{I}(C) = \mathbb{I}(\mathbb{I}(A) \cap \mathbb{I}(B)) = \mathbb{I}(A) + \mathbb{I}(B)$$

  - or, la probabilité d'observer C vaut

$$P(C) = P(A \cap B) = P(A)P(B)$$

  - on souhaite donc trouver une fonction telle que

$$\begin{aligned} \mathbb{I}(C) &= \mathbb{I}(A) + \mathbb{I}(B) \\ f(P(C)) &= f(P(A)) + f(P(B)) \\ &= f(P(A)P(B)) \end{aligned}$$

  - la fonction logarithmique a pour propriété que $log(xy) = log(x) + log(y)$
  - on obtient donc $\mathbb{I}(\omega_i) = Klog_b(P[\omega_i])$, avec $K = -1$ pour s'assurer que H soit toujours positive

Autres propriétés

- l'entropie est **bornée** par $0 \leq H(X) \leq log_b(n)$:
  - **l'entropie minimale** est atteinte lorsque une probabilité vaut 1 et les autres 0 ($|supP(X)| = 1$) :
    $$S(p_1 = 0, \ldots, p_i = 1, \ldots, p_n = 0)$$
  - **l'entropie maximale** est atteinte lorsque toutes les probabilités valent $1/n$ ($p(x) = 1/n \; \forall \, x$) :
    $$S(p_1 = 1/n, \ldots, p_i = 1/n, \ldots, p_n = 1/n)$$

  l'entropie peut donc être vue comme une façon de mesurer **l'incertitude** sur un système (au sens physique du terme)
- si un message est **découpé en morceaux**, H est égale à la somme pondérée des des valeurs des parties
  $$H(1/2, 1/3, 1/6) = H(1/2, 1/2) + 1/2H(2/3, 1/3)$$
- l'entropie est aussi **continue et monotone** : un léger changement de probabilité résulte dans un léger changement d'information

Entropie de plusieurs variables

- **entropie jointe** de deux variables :

$$H(X, Y) = \mathbb{E}\left[\frac{1}{P(X, Y)}\right] = -\sum_{x,y} p(x, y) log_b p(x, y)$$

- généralisation à plus de deux variables :

$$H(X_1, \ldots, X_n) = \mathbb{E}\left[\frac{1}{p(x_1, \ldots, x_n)}\right] = -\sum_{x_1,\ldots,x_n} p(x_1, \ldots, x_n) log_b p(x_1, \ldots, x_n)$$

- **entropie conditionnelle** de deux variables :

$$H(X, Y) = \mathbb{E}\left[\frac{1}{P(X|Y)}\right] = -\sum_{x,y} p(x, y) log_b p(x|y)$$

- généralisation à plus de deux variables :

$$H(X, Y, Z) = H(Z|X, Y) + H(Y|X) + H(X)$$

$$H(X_{1:n}) = \sum_{i=1}^{n} H(X_i|X_{1:i-1})$$

Note : la fonction log_b transforme les produits de probabilités $P(X, Y, Z) = P(Z|X, Y)P(Y|X)P(X)$ en sommes d'entropies

Anonymity sets

- notion fondamentale des études sur la réidentification
- elle peut être définie comme **le nombre de personnes** correspondant à un profil (ou **classe d'équivalence** $\mathcal{E}$)
- plus le nombre d'individus correspondant est **grand**, plus la réidentification est **difficile**
- ce **degré d'anonymat** n'est toutefois **pas suffisant** pour prévenir la réidentification
- d'autres facteurs rentrent en ligne de compte, comme **la discernabilité** d'une personne dans sa classe d'équivalence

Entropie des empreintes digitales d'appareil

caractéristique	entropie (bits)
plugins	15.4
fonts	13.9
user agent	10.0
http accept	6.09
video	4.83
timezone	3.04
supercookies	2.12
cookies enabled	0.35

Source : eckersley2010

Note : comme les variables ne sont pas indépendantes, l'étude utilise **le contenu conditionnel** de l'information dans les calculs : $\mathbb{I}_{s+t}(x_{i,s}, x_{i,t}) = -log_2(P[x_{i,s}|x_{i,t}])$

- les variables plugins et fonts présentent **une forte entropie**
- la distribution des empreintes est **extrêmement asymétrique** (83.6 % des empreintes étant uniques et la plus part des empreintes ont un effectif correspondant très limité)
- les appareils mobiles sont **moins facilement identifiables**
- l'étude montre de plus que même lorsque **certaines caractéristiques changent** entre deux visites sur le site, la réidentification reste possible (validation par cookies installés sur les navigateurs)

Entropie des empreintes digitales d'appareil

caractéristique	Panopticlick	AmIUnique
`list of plugins`	0.817	0.578
`list of fonts`	0.738	0.446
`user agent`	0.531	0.570
`screen resolution`	0.256	0.277
`timezone`	0.161	0.201
`cookies enabled`	0.019	0.042

Source : laperdrix2016

Note : les deux études comportant un nombre différent d'enregistrements, l'entropie des variables a été **normalisée** par les auteurs au moyen de la formule $H(x)/H_m$ où $H_m = log_2(n)$ désigne l'entropie maximale de la variable

- les résultats des deux études sont **proches**
- la variable `plugins` présente notamment une entropie moindre mais qui reste forte

  Note : les auteurs expliquent cette différence par la progression de la part des téléphones dans les connexions internet

- limites des deux études :
  - de par leur mode de recrutement, elle renseigne les propriétés d'un **public averti**
  - la faible entropie de la variable `timezone` montre que le recrutement des participants s'est fait dans **une zone géographique restreinte**
  - en conséquence de quoi, pas de possibilité **d'inférence** (comme c'est souvent le cas sur les données issue d'internet)

33 Bits of Entropy

L'entropie permet de mesurer globalement la quantité d'information nécessaire pour **identifier n'importe qui sur la planète** :

- si on souhaite identifier une personne prise au hasard, combien de `bits` sont-ils nécessaires ?

$$log_2(N) = log_2(7,55 \text{ milliards}) \simeq 33 \texttt{ bits}$$

33 `bits` d'information sont donc nécessaire pour identifier une personne (et $log_2(67 \text{ millions}) = 26$ `bits` si on se restreint à la France)

> **Note :** à titre de comparaison, les architectures courantes travaillent sur des mots d'une longueur de 64 `bits` et les architectures 128 `bits` seront sans doute amenées à se répandre dans les années à venir

- cette notion a été popularisée par **Arvind Narayanan**, alors chercheur en informatique à l'université du Texas à Austin

The Netflix Prize

A. Narayanan (avec son collègue V. Shmatikov) (2008) s'est aussi distingué en publiant un article montrant les possibilités de réidentification à partir d'un jeu de données **pourtant « anonymisé »** par son diffuseur :

- en 2006, `Netflix` a organisé **un concours** pour trouver un meilleur algorithme de recommandation que celui utilisé par la plateforme
- pour cela, `Netflix` a **diffusé** une base renseignant plus de 100 millions évaluations de films par près de 500 000 utilisateurs du site ($\simeq 1/8$ de l'ensemble)
- A. Narayanan et V. Shmatikov ont trouvé plus intéressant de trouver un moyen **de désidentifier les données**
- ils ont ainsi réussi à prouver qu'il était possible réidentifier une partie des individus de la base
- et ce, alors que les données ne comportaient aucunes données **directement identifiantes** et avaient été pseudonymisées

Note : plusieurs utilisateurs du service lancèrent par la suite une *class action* contre la plateforme pour infraction au *Video Privacy Protection Act*. La publication de l'article n'avait en effet pas arrêté le concours ni empêché `Netflix` de continuer à publier des données toujours plus identifiantes dans le cadre du concours.

The Netflix Prize

- pour réidentifier les personnes, A. Narayanan et V. Shmatikov ont **comparé** les évaluations de la base `Netflix` avec celles réalisées sur le site `IMDb`

  Note : comme les **CGU** d'`IMDb` interdisent la récupération massives d'information sur le site, les auteurs se sont contentés de réidentifier un nombre limité personnes

- les données sont pourtant **éparses** (chaque individu n'a évalué qu'une infime portion de l'ensemble des films)
- l'entropie des données est donc **faible** (la k-anonymisation est ici impraticable)
- mais c'est pourtant l'éparpillement des données qui va servir de fondement à la réidentification
- en utilisant la mesure de **similarité**

$$Sim(r_1, r_2) = \frac{\sum r_{1,i} r_{2,i}}{|sup(r_1) \cup sup(r_1)|}$$

  la plus part des enregistrements s'avèrent **différents**

- de plus, l'étude propose **un modèle probabiliste** de réidentification pour l'appliquer à ces données

The Netflix Prize

Il ressort de l'étude que seul un volume **relativement limité** d'information auxiliaire est nécessaire pour réidentifier les abonnés de `Netflix`

- avec seulement 8 évaluations (et leurs dates), 99 % des abonnés peuvent être réidentifiés
- avec deux évaluations, 68 %
- et seulement 3 *bits* d'entropie supplémentaires sont nécessaires pour réidentifier les autres
- sans les dates, six à huit évaluations de films hors des 500 films les plus évalués sont nécessaires pour identifier 84 % des abonnés

The Netflix Prize

- la réidentification nécessite **des données auxiliaires** mais comme dans la plus part des scenarii d'attaques
  - mais ce type de données peut être facile à obtenir (cf. `IMDb` ou *infra*)
- la réidentification de ce type de données peut paraître **vénielle** parce qu'elles ne renseignent pas des données sensibles
  - mais l'évaluations peut aussi être utilisée pour **inférer** l'orientation politique ou sexuelle des personnes
  - et donc déterminer les orientations de personnes si elles peuvent être réidentifiées
- de plus, l'algorithme n'est **pas spécifique** à `Netflix` ou `IMDb`
- A. Narayanan et V. Shmatikov ont appliqué une démarche similaire **aux réseaux sociaux** pour la réidentification d'utilisateurs de `Twiter` en les croisant avec des profils `Flickr` (Narayanan et Shmatikov, 2009)

Réidentification à partir d'attributs

La réidentification peut aussi être réalisée à partir **de caractéristiques sociales élémentaires** :

- **Latanya Sweeney** (2000) a ainsi montré à partir du recensement de 1990 que le code postal (*ZIP code*) à cinq chiffres, le sexe et la date de naissance identifiaient **87 % de la population des États-Unis** de façon unique

  Note : une autre étude (Golle, 2006) estime le nombre à 63 % à partir des mêmes données (ainsi que le recensement de 2000). L'auteur indique toutefois ne pas être en mesure d'expliquer la différence entre les deux études.

- en utilisant **les listes électorales**, elle a aussi réussi à identifier **William Weld**, l'ancien gouverneur du Massachusetts dans une base médicale de séjours à l'hôpital des agents publics de l'État
- six personnes à Cambridge avait la même date de naissance, trois étaient des hommes et une seule correspondant à son *ZIP code*

  Note : les codes postaux étasuniens correspondent à un découpage infra communal et permettent donc un géoréférencement plus précis

- l'ironie de l'histoire est que W. Weld avait **approuvé la publication** des données en assurant que la confidentialité des données était garantie par les mesures d'anonymisation prises

Modèles de désidentification

- la pseudonymisation est **clairement insuffisante** pour empêcher toute réidentification
- dans le but de pouvoir diffuser des données tout **en protégeant la privauté des personnes** de la réidentification par croisements, différents modèles et algorithmes de désidentification ont été proposé
- ils reposent généralement sur des scenarii où un attaquant dispose d'informations auxiliaires qu'il peut relier aux **données indirectement identifiantes** divulguées $\mathcal{QI}$
- $\mathcal{QI}$ est défini comme l'ensemble des attributs non sensibles $\{Q_1, \ldots, Q_{|\mathcal{QI}|}\}$ de la table T peuvant être liés à d'autres informations pour identifier une personne de façon unique

  Exemple : sexe, âge, code postal

- $\mathcal{S} = \{s_1, \ldots, s_{|\mathcal{S}|}\}$ désigne l'ensemble des données sensibles
- $\mathcal{A} = \mathcal{QI} \cup \mathcal{S}$ désigne l'ensemble des attributs
- $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$, un sous-ensemble d'attributs

le k-anonymat

Définition : une table T satisfait le k-anonymat si pour tout tuple $t \in T$ il existe $k-1$ autres tuples $t_{i,1}, t_{i,2}, \ldots, t_{i,k-1} \in T$ tels que $t[\mathcal{C}] = t_{i,1}[\mathcal{C}] = t_{i,2}[\mathcal{C}] \ldots, t_{i,k-1}[\mathcal{C}]$

- le **k-anonymat** a été des premier modèle de désidentification publié
- il a été proposé par L. Sweeney pour tout $\mathcal{C} \in \mathcal{QI}$
- le k-anonymat est obtenu par la combinaison de **deux opérations** :
  - **généralisation :** la valeur est remplacée par une tranche (ex. : âge : 26 ans $\Rightarrow [20, 29]$)
  - **suppression :** la valeur est supprimée

  Note : la suppression est en fait un cas particulier de généralisation où il ne reste qu'une seule catégorie
- ces opérations créent **des classes d'équivalence** où tous les individus ont les mêmes attributs
- les attributs identifiants sont soit **supprimés** ou **généralisés** de façon à ce que chaque groupes correspondent **à $k-1$ personnes** (chaque individus est indiscernable de $(k-1)$ autres individus)

le k-anonymat

En pratique,

- différents algorithmes ont été proposés pour implémenter le k-anonymat
- mais le k-anonymat est un problème de **difficulté NP** même pour $k = 3$
- au-delà de la complexité de l'opération, plusieurs **attaques** sont possibles :
  - attaque par **homogénéité** :

    Exemple : à un profil de $\mathbf{DCP}$ ne correspont qu'une seule donnée sensible

    chaque groupe de $\mathbf{DCP}$ identiques ne présente pas assez d'entropie
  - attaque par **connaissance** *a priori* (approche bayesienne) :

    croyance *a priori* :

    $$\alpha_{q,s} = P_f[t[S] = s | t[Q] = q]$$

    croyance *a posteriori* :

    $$\beta_{q,s,T^*} = P_f[t[S] = s | t[Q] = q \wedge \exists t^* \in T^*, t \to t^*]$$
- de plus, le le k-anonymat n'est pas efficace pour les données **éparses** ou **de grandes dimensions** (malédiction des dimensions)

la ℓ-diversité

D'autres approches ont été proposées comme **ℓ-diversité** (Machanavajjhala et al., 2007) :

- pour prévenir les deux attaques précitées, la ℓ-diversité repose sur un principe de **diversité des attributs sensibles** des classes d'équivalence
- ℓ désigne **le degré de protection** contre la connaissance *a priori* de l'attaquant (le nombre d'enregistrements que l'attaquant doit être en mesure d'écarter pour trouver sa cible)
- une table T est (entropiquement) **ℓ-diverse** si pour chaque q^*-bloc (classe d'équivalence) :

$$-\sum_{s \in S} p_{q^*,s} log(p_{q^*,s}) \geq log(l)$$

  où $p_{q^*,s}$ désigne la fraction des tuples dans le q^*-bloc dont la valeur vaut s
- les attributs sensibles de chaque q^*-bloc prennent au moins ℓ **valeurs différentes**
- une version moins stricte : la (c, l)-diversité récursive

la ℓ-diversité

Différentes études ont toutefois souligné **les limites** de cette autre approche

- la ℓ-diversité nécessite que les données présentent **une entropie suffisante**

  ℓ doit être déterminé en fonction de l'entropie de T

- elle peut conduire à une généralisation **trop drastique**

De plus, plusieurs attaques sont envisageables :

- **attaque par similarité sémantique :** les différentes modalités de la variable sensible peuvent être regroupées en une catégorie suffisamment signifiante (cf. *supra* attaque par homogénéité)

  Exemple : un ulcère et une gastrite sont deux infections stomacales

- ***Skewness attack* :** utilisation de la différence entre la distribution de la variable sensible dans la base et sa distribution dans la population

Autres approches

Autres approches dans la même veine :

- *t*-proximité, *p*-sensibilité,. . .
- *differential privacy* (pour les requêtes sur des bases de données),. . .
- différentes études ont, là aussi, souligné **leurs limites** respectives

Autre type d'approches :

- les approches évoquées précédemment sont issues **de l'informatique**
- l'intérêt pour ces questions **est relativement récent** dans cette discipline (principalement depuis la fin des années 90)

  auparavant, la confidentialité des données se confondait avec la sécurité des données
- la question a été abordée de façon systématique beaucoup plus tôt **en statistique** dans le cadre de la diffusion de données

  la diffusion de données, particulièrement de statistiques administratives, est en effet ***une pratique ancienne***
- ce qui a conduit au développement **d'autres approches** (Duncan, Elliot et Salazar Gonzalez, 2011) :
  - perturbations des données : PRAM (*Post-Randomization Method*), micro-agrégation, *shuffling*
  - simulation de données
  - accès sécurisés (**CASD** en France)
  - . . .

Fazit

En résumé,

- il ne faut jamais sous-estimer **les possibilités de réidentification** offertes par des données

  *quand bien bien même celles-ci paraissent **éparses**, **imprécises** et **lacunaires***

- en conséquence de quoi,
  - il est préférable de considérer les données comme **de toute façon identifiantes** *a priori*
  - et de prendre **les mesures appropriées** (*a minima*, sécuriser les données)

Surtout,

- il ne faut pas non plus sous-estimer **les effets d'une éventuelle réidentification** sur les personnes concernées
- et ça, d'autant plus que le RGPD rend obligatoire **la réalisation d'études impact**
- et renforce **les sanctions**

Identification et diffusion des données

Les différents exemple évoqués montrent la forte tension qui existe entre **la diffusion des données** en libre-accès et **la protection des personnes** :

- supprimer de l'information permet de mieux **se protéger de la réidentification**
  - mais sans pour autant pouvoir l'interdire complètement
  - et peut contribuer à diminuer l'information contenue par des données déjà parfois pauvres
- des solutions algorithmiques existent
  - mais il semble difficile d'arriver **à une solution optimale** tant elles dépendent de la distribution et de l'entropie des données
  - le résultat risque de se faire **au détriment** d'un ou l'autre des objectifs
  - des solutions strictement algorithmiques semblent donc **impraticables**
- ce qui constitue un frein aux politiques de **libéralisation** de l'accès aux documents administratifs (*Open Data*)
  - la LRN soumet ainsi la publication **aux autres réglementations** en vigueur comme la réglementation relative aux DCP
  - de même, tout projet de diffusion de données collectées dans le cadre d'une enquête doit **évaluer au préalable** les possibilités de réidentification et les risques attenants

Conclusion

Conclusion

- la réglementation **encadre** la collecte de DCP et **parfois** la limite
- l'application de la règlementation peut **impacter** ce que vous pouver collecter et la façon dont vous pouvez le traiter
  - implications **pratiques** et même **épistémologiques** (parcimonie, rapport à la population enquêtée,. . .)
  - mais l'impact **varie** considèrablement en fonction du traitement
  - elle affecte avant tout **les modalités** de la collecte et de l'analyse (consentement, sécurisation,. . .)
  - difficultés pratique de l'analyse juridique **dans certains cas**

Note : ces difficultés sont aussi le résultat du peu d'intérêt suscité par la question depuis 1978

- toutefois, l'encadrement et les éventuelles contraintes qui en découlent ont pour objet la **protection des personnes** concernées
- en protégeant les personnes, la réglementation certes crée **un aléas juridique**
  - mais cet aléas procède des risques que les traitements font courir aux personnes concernées
  - de plus, la conformité est une protection contre cet aléas

bibliographie

Bibliographie I

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

Duncan, George T., Mark Elliot et Juan Jose Salazar Gonzalez (2011), *Statistical Confidentiality : Principles and Practice,* Statistics for Social and Behavioral Sciences, Springer. 200 p.

Eckersley, Peter (2010), « How Unique is Your Web Browser ? », *Proceedings of the 10th International Conference on Privacy Enhancing Technologies*, PETS'10, Berlin, Springer, p. 1–18.

Fuster González, Gloria (2014), *The Emergence of Personal Data Protection As a Fundamental Right of the EU,* Springer. 274 p.

Golle, Philippe (2006), « Revisiting the Uniqueness of Simple Demographics in the US Population », *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society*, WPES '06, Alexandria, ACM, p. 77–80.

Laperdrix, Pierre, Walter Rudametkin et Benoit Baudry (2016), « Beauty and the Beast : Diverting modern web browsers to build unique browser fingerprints », *37th IEEE Symposium on Security and Privacy (S&P 2016)*, San Jose, IEEE Computer Society. URL : https://hal.inria.fr/hal-01285470.

Machanavajjhala, Ashwin, Daniel Kifer, Johannes Gehrke et Muthuramakrishnan Venkitasubramaniam (2007), « L-diversity : Privacy Beyond K-anonymity », *ACM Trans. Knowl. Discov. Data*, 1, 1.

Bibliographie II

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.

– (2009), « De-anonymizing Social Networks », *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy*, SP '09, Washington, IEEE Computer Society, p. 173–187.

Nikiforakis, Nick, Alexandros Kapravelos, Wouter Joosen, Christopher Kruegel, Frank Piessens et Giovanni Vigna (2013), « Cookieless Monster : Exploring the Ecosystem of Web-Based Device Fingerprinting », *Proceedings of the 2013 IEEE Symposium on Security and Privacy*, SP '13, Washington, IEEE Computer Society, p. 541–555.

Quantin, Catherine et Benoît Riandey (2012), « Les techniques d'appariements sécurisés. De l'épidémiologie à la démographie », *Les systèmes d'information en démographie et en sciences sociales. Nouvelles questions, nouveaux outils ? : Actes de la Chaire Quetelet 2006*, Chaire Quetelet, Louvain, Presses univ. de Louvain, p. 483–498.

Shannon, Claude E. (2001), « A Mathematical Theory of Communication », *SIGMOBILE Mob. Comput. Commun. Rev.* 1, 5, p. 3–55.

Sweeney, Latanya (2000), *Uniqueness of Simple Demographics in the U.S. Population*. Rapp. tech. 3. Carnegie Mellon University.

Merci pour votre attention