La réglementation relative aux données à caractère personnel en sciences sociales

Thomas Soubiran

Ceraps

UMR 8026 CNRS - université de Lille

Centre d'Études et de Recherches Administratives, Politiques et Sociales

8 février 2018

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 ne sont pas **nommément citées** ou bien **pseudonymisées**
  - où si **l'identité des personnes** n'est pas utilisée ou si les DCP collectées ne sont pas 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 **cette année**

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 **projet de loi** a été présenté au Conseil des ministres **le 13 décembre** dernier par la ministre de la justice Nicole Belloubet
- **son examen** par l'Assemblée nationale a débuté **le 6 février** (en procédure accélérée)
- le projet n'ayant pas encoré été adopté, la présentation portera **sur le RGPD**

  en mentionnant les différences avec la LIL (dans sa version actuelle) 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

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
  - la réglementation ne fournit qu'**un cadre général**
  - les traitements doivent donc être analysés **au cas par cas**
  - d'autant plus qu'en sciences sociales, la réduction à un nombre réduit **de cas typiques ou pratiques** est difficile
  - et que 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**

    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) une nouvelle loi est en cours d'examen
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**, *modifie la loi* CADA 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,. . .

Un cadre juridique européen

- les premières réglementations sont **des initiatives nationales**
- à partir de la fin des années 1970, les instances européennes ont commencé **à se saisir progressivement** de la question de la protection des données
- jusqu'à développer un cadre juridique **applicable à tous les États membres**
- le cadre européen reprend différentes notions élaborées dans **le cadre national**
- notamment, dans le cadre de la LIL
- mais aussi de la législation allemande dont elle reprend la notion **d'autodétermination informationnelle** (*informationelle Selbstbestimmung*)
  - notion proposée au début des années 1970 par deux juristes allemands (Wilhelm Steinmüller et Bernd Lutterbeck)
  - elle a été reconnue comme droit fondamental (*Grundrecht*) par le Tribunal constitutionnel fédéral de Karlsruhe par le *Volkszählungsurteil* prononcé en 1983
  - pose le principe que les personnes doivent pouvoir être en mesure **de décider de l'utilisation des DCP les concernant**
- il s'agit **d'un principe fondamental** dont découle un certain nombre d'obligations comme l'information des personnes, différents droits comme le droit d'accès, de rectification ou d'effacement ainsi que la limitation de la conservation des données

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

    voir aussi la `norme simplifiée n° 19` du 24 mars 1981 sur les traitements statitisque effectués par l'État et les établissements publics ainsi que 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**

Le contexte du règlement européen

- au fil des années, malgré **sa généralité**, le cadre réglementaire développé au niveau des États et de l'Union a montré **ses limites**
- et ce du fait de l'apparition **de nouvelles techniques** et **de nouveaux agents économiques**
- depuis les années 90, **le commerce des DCP** a ainsi été massivement développé par différents opérateurs

  massif à la fois de par la multiplication des vendeurs et des sommes engagées

- le développement de ce commerce est largement lié **au développement d'internet** mais pas uniquement

  les débuts de la marchandisation des DCP lui sont antérieurs et remontent au moins au années 1970, notamment pour les besoins du télémarketing

La commercialisation d'internet

L'ouverture d'internet est allé de pair avec sa commercialisation :

- l'usage d'internet et les différentes infrastructures qui l'ont précédé (`ARPANET`, `NSFNET`) a d'abord été **limité à quelques organisations**

  principalement l'armée des EU, des universités dans leur grande majorité étasuniennes et quelques entreprises du secteur numérique étasuniennes elles aussi

- les premiers **services commerciaux** comme `CompuServe` (mais aussi le `Minitel`) ont commencé à apparaître à la fin des années 80

  accès internet, messagerie internet mais aussi de la vente en ligne,. . .

- dès les années 90, il devînt clair que, à quelques exceptions près, les utilisateurs d'internet n'étaient **pas disposés à payer** pour accéder aux services proposés en ligne

  et ça, d'autant que plus que, de par ses origines, beaucoup de choses était déjà accessibles à titre gracieux sur le net

La commercialisation d'internet

- les prestataires se sont donc vite orientées vers **un financement par la publicité** proche de celui de **la télévision**
- pour financer des activités extrêmement coûteuses ne serait-ce qu'en **coût d'exploitation**

  Note : le numérique n'a, en effet, **rien de virtuel**
- et c'est là que les choses ont commencé à sérieusement **se gâter**
- cette « gratuité » a en effet conduit à l'éclosion d'un véritable ***business* de la surveillance** de masse (Schneier, 2015)

Le prix de la gratuité

- le soucis des annonceurs a toujours été de pouvoir caractériser **le plus précisément possible** la clientèle ciblée

  selon l'idée que, plus le profilage du client est précis, plus la publicité serait efficace

- les sites **ne se contentent donc pas** d'afficher des pages de publicité
- différentes infrastructures et techniques sont proposées par des entreprises comme `doubleclick.net` ou `tacoda.net` **pour traquer les utilisateurs** au cours de leur navigation
- et ce, **en temps réel**
- **Exemples de techniques :** (flash|zombie|...) cookies, pixels espion, web beacon, empreinte digitale d'appareil*,...

  Certaines de ces techniques sont même l'objet de spécifications par le `w3c` *comme les* web beacons

---

* pour plus de détails sur cette technique, voir la **présentation** faite à data-shs (Soubiran, 2017a) en décembre dernier

Enchères

- les utilisateurs sont ensuite **mis aux enchères** (*real time biddings*) (Olejnik, Minh-Dung et Castelluccia, 2013) par des régies publicitaires pendant le chargement de la page
  - les enchères sont réalisées en fonction **du profil** établi à partir de l'activité de utilisateurs

    Exemples : historique des pages visitées, recherches soumises (moteur de recherche, recherche de produits sur un site de vente,. . .)
  - ce profil (et donc des DCP) **sont transmis aux annonceurs** qui renchérissent si le profil les intéressent
  - comme sur les marchés de transactions à haute fréquence, ces enchères sont réalisées par **des machines**

    ces transactions prennent en général moins de 100 ms
- la « gratuité » n'est donc **qu'apparente**

  et on attache sans doute une trop grande valeur à la gratuité
- d'où le dicton : sur internet, quand c'est gratuit, **c'est toi le produit**

Enchères

- l'utilisation des DCP collectées pour ces enchères **ne se limitent pas** au temps immédiat
- les données des utilisateurs peuvent être **cumulées dans le temps** :
  - par les collecteurs
  - mais aussi par les annonceurs
  - les enchères nécessitent l'envoi de DCP pour déterminer la valeur du profil de l'utilisateur par l'annonceur
  - les annonceurs peuvent être en mesure de relier les propositions qui leur sont faites
  - ces données peuvent aussi être vendues sur des bourses de données (*Data Exchange*) via des plateformes spécifiques (*Data Management Platform*)

    Note : ces plateformes proposent généralement leurs services en mode SaaS (Sofware as a Service)
- les enchères servent donc **au profilage des utilisateurs**
- car le profilage augmente la valeur de la proposition **de placement de publicité**
- au de-là, les collectes de DCP sur internet ont plus généralement conduit à la mise en place **d'un profilage massif** des populations
- **Exemple :** Google

Big Googler is watching you

- Google propose **de nombreux services** « gratuits » ou payant :
  - messagerie (`gmail.com`), stockage (`googledrive`), collaboration (`googledoc`), streaming (`youtube.com` acquis en 2006), cartographie de la terre (`googlemap`) ou de Mars, réseautage social (`google+`), `OS` (`android`),...
  - ainsi que des api pour les développeurs web : `googlefont`, `googleapis`,...
  - mais aussi des services de traque : `doubleclick.net` (acquis en 2007), `googleanalytics`

    *ces deux sites comptent parmi **les principaux traqueurs** du web*
  - en tout près de 150 services de natures diverses
- l'objet de ces services est de collecter de DCP pour les croiser et ensuite **profiler les utilisateurs**
- pour exploiter **commercialement** ces profils via `AdWord`, `AdSense`,...

  et, plus accessoirement, améliorer « l'expérience utilisateur »

Gafam & co.

- google **affirme** pouvoir diffuser diffuser de la publicité sur plus de **2 millions de sites** et **650 000 applications**
- faisant qu'une entreprise de $\simeq$ 70 000 salariés a **un chiffre d'affaire** supérieur au 2/3 des PIB des pays de la planète (66 m^{ds} \$)
- à partir de recettes **essentiellement publicitaires**
- Google n'est bien évidemment pas la seule entreprise a avoir adopté ces pratiques
- c'est aussi le cas de nombreux sites comme Facebook, Amazon, LinkedIn,...

Le marché des données

- les données personnelles ont donc **une valeur**

  Exemple : un site permet de calculer votre valeur pour des annonceurs sur internet.
  Il en ressort notamment que des reseignements come l'âge, le sexe ou le lieu de résidence valent environ 0.0005 $ par personne.

- et **un marché** toujours plus structuré

  le commerce des DCP aurait ainsi généré plus de 150 m^{ds} \$ en 2012

- la marchandisation des DCP a depuis déjà longtemps conduit à l'apparition d'une profession spécialisée dans la collecte et la vente de DCP : **les courtiers de données** (*data brokers*)

  les courtiers de données achètent des informations provenant de sources diverses pour ensuite les revendre à d'autres compagnies

- **Exemple :** Acxiom

  - société fondée en 1969 aux États-Unis
  - spécialisée dans « la donnée client, l'analytique et les services marketing »
  - avec aujourd'hui des filiales dans différents pays dont la France
  - et des informations sur près de 700 m^{ns} de personnes (Federal Trade Commission, 2014)
  - chiffre d'affaire : 1,15 m^{ds}\$/an

Le marché des données

- la mercantilisation des DCP a aussi plus récemment conduit à l'éclosion d'un ***business* de la protection** des DCP
- des entreprises, des cabinets d'avocat se spécialisent dans la consultation, suppression,... de DCP
- cette surveillance en masse a aussi conduit au développement **d'une offre logicielle**

  Exemples : `plugins` (extensions) pour navigateurs, `proxy`, réseaux superposés (`Tor`),...

- qui n'est parfois **pas sans ambivalences**
- **Exemple :** `Ghostery`

  - extension propriétaire pour navigateur web chargée de bloquer les mouchards et les cookies des pages web que l'internaute visite
  - développée par une société de...marketing
  - elle récupère notamment (sur la base du volontariat) des données sur les publicités bloquées pour les envoyer aux annonceurs pour leur permettre « d'améliorer » leur publicité

L'extension de la surveillance

- internet n'est **pas le seul vecteur de l'extension** continue de la portée de la surveillance
- plus généralement,
  - **les moyens de collecte** ne cessent d'augmenter

    vidéosurveillance (publique ou privée), mobilité (téléphones, wifi,. . .), accès, transactions, self-tracking. . .
  - ainsi que **les capacité de stockage**

    750 € pour stoker toute la musique jamais enregistrée
  - **la puissance de calcul**

    la loi de Moore se tasse mais les cœurs se multiplient et se distribuent
  - **l'efficacité des algorithmes**

    les algorithmes de reconnaissance faciale sont aujourd'hui plus performants que des humains dans certaines conditions

  Bref : les techniques de surveillance sont toujours plus efficaces pour un coût toujours moindre
- la multiplication de ces traces offre des possibilités **de croisements inédites**

  Exemple : identification de personne à partir d'images de vidéosurveillance et de photographies glanées sur internet

Le RGPD

- ce qui précède ne donne qu'un très **bref aperçu** de l'ampleur de ce qui est à l'œuvre

  *porosité avec la surveillance par les États, volontairement (PRISME) ou involontairement (*hack *par une agence de renseignement), cybercriminalité,. . .*

- et de **l'opacité** qui entourent ces traitements

  opacité qui contraste avec l'idéologie de la transparence utilisée pour justifier les collectes. « Si vous n'avez rien à vous reprocher. . . »

- et donc de ce qui a conduit à la rédaction **d'un nouvel acte législatif** européen
- qui :
  - renforce les droits des personnes et les obligations des responsable de traitement
  - ainsi que les sanctions
  - mais supprime partiellement les contrôles préalables

Le RGPD

- à l'instar de la LIL, les circonstances de l'élaboration du **RGPD** peuvent sembler **très éloignée** des sciences sociales
- mais, comme pour la loi de 1978 ou la directive 95/46/CE de 1995, la rédaction du nouvel acte **n'a pas donné lieu** à une mobilisation autour de ces questions

  alors que les négociations ont duré quatre ans et ont été largement publicisées

- toutefois,
  - les sciences sociales procèdent aussi à **des exploitations** importantes de **DCP**
  - une utilisation non-marchande **ne dissout pas** les risques attenants au traitement de **DCP**
  - les traitements **« à fins scientifiques »** n'ont pour autant pas été oubliés

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 traitements 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

Note : autant d'assertions qui sont d'ailleurs utilisées pour justifier le désintérêt pour la réglementation et son application

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. 75)

Un cadre juridique général

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

- si la réglementation n'apparaît pas comme pensée pour les sciences sociales, c'est qu'elle n'a été pensée **pour aucune application en particulier**

  cf. abstraction progressive de la réglementation du contexte de son élaboration

- 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

Un cadre juridique général

- l'analyse juridique des 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**

    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 : une finalité recherche permet toutefois d'établir **la licéité du traitement**
- les traitements à fins de recherche scientifique font toutefois l'objet **de 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. 129
- 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 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

    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**
- importance, aussi, **d'associer votre** CIL à vos projets de recherche
  - le CIL ne veille pas seulement à **la conformité des traitements** de DCP réalisés par le responsable de traitement (cf. *infra* p. 84)
  - il a aussi une mission **de conseil et d'information**

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 **d'identifier** une personne et pas seulement de la nommer :

- l'application de la réglementation ne se réduit donc pas à la seule question de **« l'anonymat »** *stricto sensu*
- Autrement dit, elle ne se limite pas à la seule question de savoir si des renseignements comme des noms figurent dans les informations détenues :
  - 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. 129)
  - 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 (cf. *infra* p. 121)
- plutôt que d'anonymat, il est donc préférable de parler **de possibilité de réidentification** 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. 68)

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**

  *voire même ce qui **pourrait** en être fait, indépendamment de la finalité affichée*

- 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. 96)

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, . . .
  - 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 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 recommandations** du **G29**
- manque **d'un référentiel** propre aux sciences sociales (cf. *infra* p. 101)

Destinataires de données

Définition : la personne physique ou morale, l'autorité publique, le service ou tout autre organisme qui reçoit communication de données à caractère personnel, qu'il s'agisse ou non d'un tiers (RGPD art. 4 § 9)

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

  Note : destinataires de données est une notion distincte **de tiers autorisé**

  le tiers autorisé, comme les autorités publiques, bénéficie d'une habilitation lui permettant d'obtenir la communication des données

- **Exempe :** les membres d'un projet de recherche, sans être limité, p. ex., aux membres des UMR du ou des responsables de traitement

Note : tiers désigne une personne physique ou morale, une autorité publique, un service ou un organisme autre que la personne concernée, le responsable du traitement, le sous-traitant et les personnes qui, placées sous l'autorité directe du responsable du traitement ou du sous-traitant, sont autorisées à traiter les données à caractère personnel (RGPD art. 4 § 10)

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 remises** 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 **10 ou 20 millions €**
- ou **2 ou 4 % du chiffre d'affaires** annuel mondial de l'exercice précédent
- le plus élevé de ces deux montants est retenu
- les montants maximums concernent notament les violations des principes fondamentaux d'un traitement (licitéité, transparence, finalité déterminée, proportionalité, données sensibles,. . .), du droit des personnes, du non-respect d'une injonction,. . .

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 XXI^{e} siècle 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. 73 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 prosopographique 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 (Soubiran, 2017b)

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
- **autre cas :** les indicateurs

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

La minimisation des données

Cas pratique (plus délicat) : la religion

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

La minimisation des données

Problèmes :

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

Toutefois,

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

Fazit

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

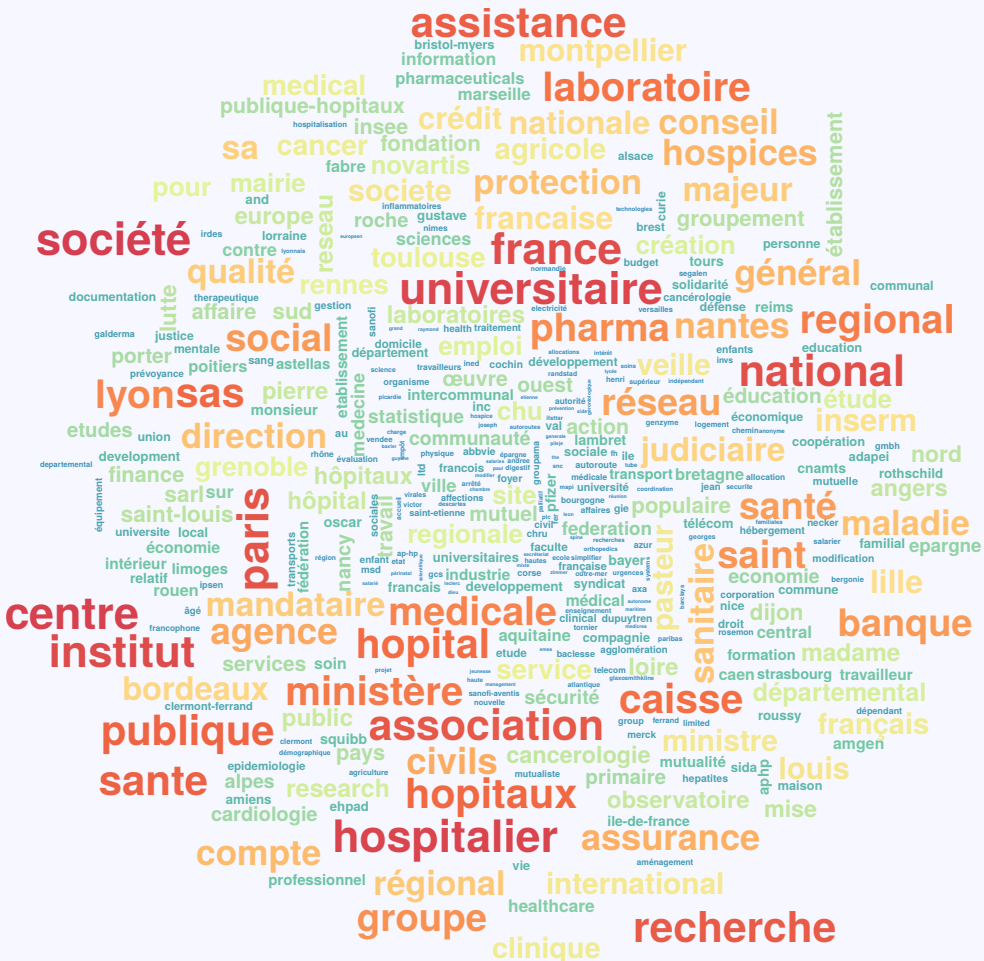
Note : il est important de souligner que ce n'est pas toujours le cas et que la proportionnalité et la pertinence des traitements peuvent être établis dans de très nombreuses situations

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 principalement **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

Les responsables de traitement dans les délibérations de la CNIL



Note : les couleurs on été calculées en utilisant la fonction de répartition empirique $\hat{F}_n(t) = 1/n \sum_{i=1}^{n} \mathbb{1}_{x_i \leq t}$.
Pour améliorer la lisibilité, la taille des mots a été calculée avec la transformation
$[x - min(x)/(max(x) - min(x))]^{\alpha}$. La taille n'est donc pas linéairement proportionnelle à la fréquence.

Classification des délibérations de la CNIL

- *topic model* : modèle probabiliste pour faire ressortir des thèmes appelé LDA (*Latent Dirichlet Allocation*) (BLEI, NG et JORDAN, 2003)
- approche différente des l'analyse des données textuelles *à la française* :
  - il s'agit de regrouper les documents autours d'un ou plusieurs thèmes
  - et non de dégager des champs, « mondes »,. . .lexicaux
- le modèle LDA est un modèle génératif : la fréquence d'apparition des mots dans un document est modélisée en fonction de l'appartenance du document à une classe latente représentant un thème (*topic*) ou, en l'occurrence, un type de traitement
- qui est une variante bayesienne du modèle de mélange fini :

$$f(y|x,\Theta) = \sum_{k=1}^{K} \pi_k f_k(y|x,\theta_k) \quad \text{avec} \quad \sum_{k=1}^{K} \pi_k = 1 \quad \text{et } \pi_k > 0 \quad \forall k$$

Note : où y désigne la variable dépendante univariée ou multivariée suivant une densité conditionnelle f, f_k désigne la distribution de y pour la classe k avec $y \sim f_k(y, \theta_k)$, x est un vecteur de variables indépendantes, π_k la probabilité (inconnue) d'observer la classe k, θ_k le vecteur de paramètres spécifiques à la distribution k et $\Theta = (\pi_1, \ldots, \pi_k, \theta_k^\top, \ldots, \theta_k^\top)$ est le vecteur regroupant tous les paramètres

Latent Dirichlet Allocation

Le modèle génératif est le suivant :

- soient :
  - w, un mot d'un vocabulaire de taille V
  - un documents est un N-tuple de mots $\mathbf{w} = \{w_1, \ldots, w_N\}$ du corpus $\mathcal{D} = \{\mathbf{w}_1, \ldots, \mathbf{w}_D\}$
- pour chaque document $\mathbf{w}$:
  - $\theta \sim \mathrm{Dirichlet}(\alpha)$, $\alpha < 1$ est la distributions des thèmes (probabilité d'occurrence d'un thème)
  - $\phi \sim \mathrm{Dirichlet}(\beta)$ est la distribution des mots (probabilité d'apparition d'un mot conditionnellement à un thème)
  - puis pour chacun des N (indices de) mots $i = \{1, \ldots, N\}$ du document $\mathbf{w}$
    - choisir un thème $z_i \sim \mathrm{Multinomiale}(\theta)$
    - choisir un mot selon une distribution multinomial conditionnellement au thème $p(w_i|_i z, \phi)$

Notes :

- il s'agit du modèle génératif, pas de son estimation
- un document peut être rattaché à plusieurs thèmes

Latent Dirichlet Allocation

- on utilise une loi de Dirichlet

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{k}\Gamma(\alpha_k)}\theta^{\alpha_1-1}\dots\theta^{\alpha_K-1} \quad p(\phi|\beta) = \frac{\Gamma(\sum_{k=1}^{K}\beta_k)}{\prod_{k=1}^{k}\Gamma(\beta_k)}\phi^{\alpha_1-1}\dots\phi^{\beta_K-1}$$

  comme a priori sur θ et ϕ car elle est conjuguée à la loi multinomiale
- étant donnés α et θ, la probabilité jointe du mélange de thèmes θ, des K thèmes $\mathbf{z}$ et des N mots $\mathbf{z}$ est donnée par :

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha)\prod_{i=1}^{N} p(z_i|\theta)p(w_i|z_i, \beta)$$

- la loi marginale d'un document (pour un document)

$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{i=1}^{N}\sum_{z_n} p(z_i|\theta)p(w_i|z_i, \beta)\right)\mathrm{d}\theta$$

  permet de calculer les paramètres (par une approximation, la somme de toutes les combinaisons de N mots dans K thèmes étant impraticable)

Résultats de la classification

- présentation des résultats provisoires d'un modèle **à 15 classes**
- provisoire, car **la préparation du corpus** n'est pas terminée
- de plus, le nombre de classes apparaît **insuffisant** :
  - certaines classes demeurent encore trop hétérogènes relativement aux traitements qu'elles agrègent
  - un modèle avec 20 classes améliore significativement la classification
- le modèle à 15 classes permet toutefois de dégager **différentes catégories de traitements**
- les classes les plus nettement définies sont au nombre de sept et représentent 39 % des délibérations

Résultats de la classification

- les traitements **statistiques** (au sens large) du secteur public ($p = 3,3$ %)
  - **recherche médicale :** enquêtes réalisées INSERM, **INVS**, Institut Pasteur, **ARS**,... souvent liées à de la mise en œuvre de politiques publiques
  - **la statistique publique :** et, plus particulièrement, les **SSM** avec les enquêtes (ou les systèmes d'information) mis en place par l'INSEE, la DARES (travail), la DREES (santé), la DEPP (éducation),... ainsi que l'INED
- les traitements de données **médicales** ($p = 6,2$ %)

  CHR ?U, laboratoires privés,...
- les traitements relevant du secteur **médico-social** ($p = 3,8$ %)

  suivis et accompagnements, notamment de personnes en situation de handicap
- les traitements **biométriques** ($p = 7,5$ %)

Résultats de la classification

- les traitements dans l'exercice **des fonctions régaliennes** ($p = 5,1$ %)
  - fichiers de police, de gendarmerie, des douanes,...
  - principalement des demandes d'avis
- **les sanctions** et mises en demeure (anonymisées...) prononcées par la CNIL ($p = 4$ %)
- **les transferts** de données en dehors de l'UE ($p = 9$ %)
- **ainsi que :** les traitements mis en œuvre par les collectivités, administrations, les dispositifs d'alerte professionnelle, la vérification des identités, surveillance sur internet, banque-assurance (fraude), entreprises,...

Les sciences sociales dans les délibérations de la CNIL

- **diversité** des traitements dans les délibérations
- importance **des traitements médicaux** dans le corpus
- les traitements les plus similaires à ceux des sciences sociales sont les traitements de **la statistique publique**

  *les très (très) rares traitements de sciences sociales émanant de l'**ESR** y sont d'ailleurs agrégés par le modèle*

- les traitements des **SSM** ont toutefois **des finalités différentes**, avant tout administratives

  avec le cas particulier de l'INED

- ces traitements sont aussi régis par **des règles spécifiques** :
  - la loi de 1951 modifée
  - de plus, une partie des traitements des **SSM** tombent dans le champ d'application de la `norme simplifiée n° 26` qui dispense de nombreux traitements effectués dans le cadre du CNIS d'un contrôle préalable de la CNIL
  - pour les traitements de certaines catégories de données

    identité, situation familiale, diplômes, logement, vie professionnelle, situation économique et financière, déplacement, habitudes et conditions de vie ,. . .

  - et dans le respect de l'**art. 27** de la LIL
- **les entreprises d'enquêtes** privées sont quasiment absentes des délibérations

Les finalités des traitements en sciences sociales

À mon avis,

- cette rareté impacte **l'analyse juridique** des traitements

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

- 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 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 accessibles 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

  Note : Condoleezza Rice a été nommée membre du conseil d'administration de `Dropbox` en avril 2014

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`,...

La pseudonymisation

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

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

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

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

La pseudonymisation

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

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

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

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**

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 traitements (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)

Un contre exemple : la « pseudonymisation » des entretiens

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

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

Ensembles d'anonymat

- la modélisation de la protection des données contre la réidentification repose notamment sur **la notion d'ensembles d'anonymat** (*anonymity sets*)
- notion proposée par D. Chaum pour modéliser la sécurité d'un réseau appelé **le réseau du dîner de cryptographes** (DC-nets *Dining Cryptographers Networks*)

  Note : à ne pas confondre avec le dîner des philosophes de E. Dijkstra

- dans ce cas particulier, la notion désigne **le nombre de personnes** membres d'un réseau qui auraient pu envoyer un message
- D. Chaum l'a utilisé pour développer **un protocole de sécurité** pour prouver qu'une personne avait réalisé une action sans révéler son identité
- il est illustré par l'exemple suivante :

  - des cryptographes participent à un repas organisé par la NSA
  - **problème :** comment savoir si la NSA ou un des cryptographes a réglé l'addition sans révéler l'identité du cryptographe en question ?

Le protocole du dîner de cryptographes

- consiste à partitionner le graphe des participants en relations deux-à-deux
- repose sur le partage de clefs secrètes d'une longueur de 1 `bit`
- chaque paire ($P_i = P_j$) choisit au hasard une clef partagée secrète ($k_{ij} = k_{ji}$) d'une longueur de 1 `bit`

  chacune des clefs est donc partagée qu'avec un seul autre membre

- le message $m_i = \{0, 1\}$ ($m_i = 1$ si P_i a réalisé l'action, 0 sinon) de chaque participant est chiffré au moyen de ces deux clefs

$$b_i = k_{ij} \oplus k_{ik} \oplus m_i = \begin{cases} k_{ij} \oplus k_{ik} \text{ si } m_i = 0 \\ \overline{k_{ij} \oplus k_{ik}} \text{ si } m_i = 1 \end{cases} \qquad \text{(addition modulo 2)}$$

- le résultat est calculé par :

$$\begin{aligned} b_1 \oplus \ldots \oplus b_i \oplus \ldots \oplus b_n &= (k_{1,n} \oplus k_{1,2} \oplus m_1) \oplus \ldots \oplus (k_{i,j-1} \oplus k_{i,j+1} \oplus m_i) \oplus \ldots \\ &\quad \oplus (k_{n,n-1} \oplus k_{n,1} \oplus m_n) \\ &= m_1 \oplus \ldots \oplus m_i \oplus \ldots \oplus m_n \end{aligned}$$

Le résultat vaut 1 si l'action a été réalisée, 0 sinon (tous les messages s'annulent sauf si un des m_i vaut 1)

Ensembles d'anonymat

- **Généralisation :** propriété de ne pas être identifiable dans un ensemble (groupe) $\mathcal{E}$ de taille n
- consiste à créer des **classes d'équivalence** dont tous les membres ont les mêmes caractéristiques
- selon l'observation que plus le nombre d'individus correspondant est **grand**, plus la réidentification est **difficile**
- la notion d'ensemble d'anonymat fournit donc **une mesure** de l'anonymat
- **Exemple :** ***k-anonymity***

  un fichier est dit k-anonyme si chaque individu est indiscernable de $k - 1$ autres individus du fichier

Ensembles d'anonymat

- la taille de l'ensemble est souvent insuffisante pour garantir l'anonymat
- l'anonymat ne se conçoit pas dans l'absolu mais **relativement à une situation**
- il faut aussi prendre en compte :
  - de l'informations auxiliaires dont peut disposer un attaquant
  - **Exemple : DC-nets**

    la connaissance d'une partie des clefs partagées réduit la taille l'ensemble d'anonymat, puisqu'on peut ainsi déterminer si une ou plusieurs personne ont réalisé l'action ou ne l'ont pas réalisée
  - de la distribution des données
  - d'autres mesures d'anonymat ont ainsi été proposées fondées sur l'entropie (1) ou l'entropie relative (divergence de Kullback-Leibler) (2)

$$H(X) = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i) \qquad (1) \qquad D_{KL} = -\sum_{x \in \chi} P(x) log \frac{P(x)}{Q(x)} \qquad (2)$$

La « pseudonymisation » des entretiens

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

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

Protection contre la réidentification

Pour modéliser **la probabilité de désidentification**, il faudrait notamment intégrer :

- la taille de la population dont sont issus les répondants
- mais aussi (et surtout) **sa diversité**

  Exemples :
  - dans une organisation hiérarchique, plus on remonte l'arborescence, plus le nombre de personnes occupant les postes tend à se réduire
  - le problème se pose aussi de façon transversale (cf. la psychologue de l'établissement)

- dans les faits, l'entropie (diversité) peut contribuer **à la protection contre la réidentification**
- mais plus l'entropie est forte, **plus la probabilité de réidentification** est importante quelle que soit la taille

  Exemple : peut ainsi discerner toutes les personnes parmi des utilisateurs de téléphones portables à partir d'un nombre limité de coordonnées GPS transmis par leur équipement (Montjoye et al., 2013)

  > *dans la majorité des cas, seules deux coordonnées suffisent (p. ex., domicile et lieu de travail)*

- l'information auxiliaire dont peut disposer un attaquant, particulièrement **s'il appartient à la population**

La proportionnalité des traitements qualitatifs

- **Rappel :** la publication **fait partie du traitement**
- les possibilités de recoupements offertes par les publications qualitatives pose la question de l'application des principes **de proportionnalité et de pertinence** de ce type de traitement :
  - la question est de savoir si le luxe de détails divulgués est **toujours nécessaire** à la démonstration
  - la divulgation répond-t-elle seulement aux nécessités de **la démonstration**
  - ou répond-t-elle à **d'autres fins** comme la production d'un effet de réel ?
- ce qui illustre la nécessité de **la réalisation d'études d'impact**

  et de la mise en place d'un cadre de référence pour leur mise en œuvre
- cette question est rendue d'autant plus pressante par **la diffusion accrue** des publications via internet

  portails de revue avec barrière mobile, *archives institutionnelles*, `google.books`,...

Réidentification

- la réidentification des personnes ne nécessite souvent **qu'une quantité d'information limitée**
  - la population mondiale est $\simeq 6,5$ milliards d'individus, soit $log_2(6.5e^9) = 33$ `bits` of entropy...
  - seuls 33 `bits` sont donc nécessaires pour identifier toutes les personnes vivantes sur terre
- **Exemple : la biométrie**

  seuls 12 points suffisent pour identifier une personne à partir de ses empreintes digitales
- **Exemple : traque sur internet**
  - **la prise d'empreintes digitales d'appareil** est de plus en plus utilisée pour traquer les utilisateurs sur la toile
  - elle consiste à utiliser `JavaScript` pour recueillir un ensemble de renseignements sur le navigateur et son environnement

    version du navigateur, protocoles supportés, extensions, polices installées, `OS`, taille et résolution de l'écran, rendu `2d` ou `3d`,...
  - pris séparément, ces renseignements peuvent paraître **anodins**
  - combinés, ils peuvent **presque sûrement** identifier la quasi intégralité des utilisateurs et les suivre dans le temps
  - et ça, malgré des changements de configuration (Eckersley, 2010)

Réidentification

- plusieurs études ont démontré les possibilités de réidentification à partir de données **indirectement identifiantes**
- quelques **exemples** de réidentifications publiés :
  - *The Massachusetts Governor* (Latanya Sweeney)

    réidentification à partir du croisement entre une base de données médicale publiée et les listes électorales
  - *The AOL Search Queries* (*The New York Times*)

    réidentification à partir du croisement entre les logs de requêtes sur le moteur de recherche de `AOL` *et l'annuaire téléphonique*
  - *The Netflix Dataset* (Arvind Narayanan et Vitaly Shmatikov)

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

Réidentification

- toutes ces données avaient été **pseudonymisées**, ce qui n'a pourtant pas empêché la réidentification
  - les deux premiers exemples montrent le caractère identifiant **des données géographiques**
  - le troisième montre, lui, que **le caractère épars** des données ne constitue pas une protection mais peut, au contraire, faciliter la réidentification
- pour une revue de ce type d'exploits,
  - voir (Ohm, 2010)
  - ainsi que la **présentation** aux journées journées Data-SHS (Soubiran, 2017a) pour plus de développements sur l'algorithmique de la réidentification et les difficultés pratiques de la désidentification

Conclusion

Conclusion

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

Blei, David M., Andrew Y. Ng et Michael I. Jordan (2003), « Latent Dirichlet Allocation »,. *J. Mach. Learn. Res.* 3, p. 993–1022.

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

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.

Federal Trade Commission (2014), *Data Brokers. A Call for Transparency and Accountability*. 110 p. url : https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf.

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

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

Ohm, Paul (2010), « Broken Promises of Privacy : Responding to the Surprising Failure of Anonymization »,. *UCLA Law Review*, 57.

Olejnik, Lukasz, Tran Minh-Dung et Claude Castelluccia (2013), "Selling Off Privacy at Auction". working paper or preprint. url : https://hal.inria.fr/hal-00915249.

Bibliographie II

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.

Schneier, Bruce (2015), *Data and Goliath : The Hidden Battles to Capture Your Data and Control Your World,*. New York, NY, USA, W. W. Norton & Company. 448 p.

Soubiran, Thomas (2017a), *La réglementation relative aux données à caractère personnel en sciences sociales*. journées Data-SHS. URL : http://ceraps.univ-lille2.fr/fileadmin/user_upload/enseignants/Soubiran/doc/data-shs--dcp.pdf.

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

Merci pour votre attention