La protection des données à caractère personnel en sciences sociales

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

Séminaire dataSHS 2019

Lille, 9 décembre 2019

La protection des données à caractère personnel

- à partir **du début des années soixante-dix**, différents États

  RFA (1970, 1977), États-Unis (1974), France (1978), . . .

- ont adopté et mis en œuvre différents textes visant à réglementer **l'usage fait des données à caractère personnel**
- pour protéger les personnes **des conséquences** de cet usage
- jusqu'à faire de la protection des données un **droit fondamental**

  dans l'Union européenne du moins

- selon lequel les personnes ont **le droit de décider de l'utilisation des informations les concernant**

La réglementation sur les données à caractère personnel

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

- **cadre juridique**
- applicable à **l'utilisation** (« traitement ») de données à caractère personnel

  c-à-d de données permettant d'identifier des personnes physiques
- la réglementation définit **les droits des personnes** concernées par le traitement
- ainsi que les **obligations** à respecter lors du traitement de données à caractère personnel **les concernant**

Les données à caractère personnel en sciences sociales

- la réglementation revêt **un caractère volontairement très général**

  ce qui la rend parfois difficile à appréhender et à appliquer

- en conséquence de quoi, **de nombreux traitements** tombent dans le champ d'application de la réglementation
- les SHS n'y font **pas exception**

  les données à caractère personnel sont en effet depuis longtemps **le « pétrole » des** SHS

- et la réglementation s'y applique

  - même si les personnes ne sont pas **nommément citées** ou bien **pseudonymisées**
  - ou si **l'identité des personnes** n'est pas utilisée
  - ou si les données à caractère personnel collectées ne sont pas utilisées pour **(ré)identifier les personnes**

La réglementation relative aux données à caractère personnel

En France, le traitement de données à caractère personnel est, depuis 1978, 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 dernière modification est intervenue **en 2018**

En effet, le 25 mai 2018,

le règlement européen sur la protection des données est entré 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 **abroge** 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 *confirme notamment différentes interprétations de la réglementation en les explicitant*
  - marque notamment le passage d'un régime **de déclaration préalable** à un régime **de responsabilisation**

Le RGPD dans le droit français

- la LIL a été modifiée à **différentes reprises** à la suite de l'entrée en vigueur du RGPD en 2016
  - même si le règlement est d'application directe
  - le droit des États membres doit malgré tout **être adapté**
  - ces modifications sont intervenues principalement **en 2018**

    notamment avec l'adoption de la loi protection des données en mai 2018
- le (long) processus législatif arrive donc progressivement **à son terme**
- toutefois, il n'est **pas encore achevé**
  - en effet :
    - un règlement dit « ePR » (*ePrivacy Regulation*) doit encore adopté
    - pour remplacer les directives 2002/58/CE et 2009/140/CE
  - d'autre part :
    - l'entrée en application pose aussi la question de la traduction **des nouvelles dispositions** dans les faits
    - qui, pour les sciences sociales, s'ajoutent à toutes celles qui n'ont **pas été réglées depuis 1978**

La présentation

- la présentation sera **axée** sur le RGPD
- en lien avec le cadre applicable en France
  - le RGPD visait à harmoniser les législation des États membres
  - le règlement laisse malgré tout **des marges d'appréciation** aux États membres

    notamment sur les traitements à fin de recherche
  - l'adaptation est aussi rendue nécessaire lorsque le réglement ne suffit pas ou pour apporter des garanties supplémentaires

    p. ex. certaines catégories de données ou de traitement : NIR, *données de santé,. . . ou encore pour la* CNIL
  - c'est pourquoi le RGPD doit être lu de façon **combinée avec la LIL**

    gare aux interprétations ne reposant que sur le RGPD
- de plus,
  - la réglementation n'édicte que des **grands principes**
  - dont il faut trouver **la traduction pratique** pour chaque traitement
  - **la doctrine** de la CNIL (Commission nationale informatique et libertés) revêt donc une grande importance dans l'application de la réglementation

    et des autorités de contrôle pour chacun des États membres

La présentation

- la présentation sera sera plutôt axée sur les aspects
  - **généraux** (grands principes)
  - et **procéduraux**
  - de la réglementation sur les DCP
  - en les illustrant par des **cas pratiques**
- en faisant le lien avec les catégories et les pratiques des sciences sociales
  - dans les faits les raisonnements ne sont **pas si orthogonaux**
  - et l'application **pas si contraignante** qu'il y paraît au premier abord

> *la réglementation constituant d'ailleurs une protection pour le* RdT *contre les réponses (éventuellement pénales) aux traitements qu'il met en œuvre*

Le délégué à la protection des données

- la présentation insistera notamment sur l'importance **d'associer votre** DPD (délégué à la protection des données) **à tous vos traitement de** DCP
  - et ce, **dès la conception** du traitement
  - ce qui constitue **une obligation réglementaire**
  - en plus d'une **nécessité pratique** pour la traduction dans les faits des principes la réglementation

    l'implication tardive DPD *compliquera d'autant la mise en conformité et l'enregistrement du traitement à son registre*
- elle vise avant tout **à familiariser** avec les principes et la pratique de la PdD pour **faciliter** les échanges avec le DPD lors de l'instruction des traitements

  les DPD *ne sont pas de simples chambres d'enregistrement*
- **les difficultés** rencontrées dans la mise en œuvre de la réglementation viennent plus souvent de **sa méconnaissance** que de la réglementation proprement dite

Plan

- **appréhender la réglementation sur les données à caractère personnel**
  - **chronologie**
  - **remarques générales**
- **la réglementation**
  - **notions fondamentales**
  - **modalités et agents de la protection des données**
- **mise en œuvre de la réglementation :**
  - **interprétation** (et difficultés d'interprétation) des notions dans le contexte spécifique des sciences sociales
  - **protection des données**
  - **(ré)identification** des personnes

Appréhender la réglementation

Préambule

Cas concrets

Trois **cas concrets** pour introduire le sujet :

- *Cambridge Analytica*
- Disinfolab
- Les fichiers Monsanto

Cambridge Analytica

- **collecte massive** de données à partir de FACEBOOK
  - sur plus de 30, 50 ou 87 millions d'utilisateurs
  - par la société Cambridge Analytica

    qui offrait différents services à des entreprises ou certains partis politiques pour « changer les comportements de leur publics », notamment par des un ciblage publicitaire « psychographic » (sic)
- où les utilisateurs avaient **donné leur consentement** à une utilisation à fin de recherche (**sans plus de précisions**)
- de différentes **données à caractère personnel** les concernant

  comme leur likes, localisation, contacts,. . .
- mais portant aussi **sur leurs contacts**

  sans information ni collecte du consentement

Cambridge Analytica

- soupçons **d'utilisation des données** dans le cadre de la campagne présidentielle étasunienne de 2016
- et de la campagne du « Brexit »
  - pour **profiler** des électeurs et **influencer** leur vote
  - notamment par **le placement de publicités**

  Note : *D. Trump n'a pas remporté le suffrage « populaire » mais celui des grands électeurs*

- qui a notamment abouti à la cessation d'activité de l'entreprise
- et la **la condamnation** de Facebook à une amende de **5 m^{ds} \$** par la FTC en juillet 2019

L'avocat du Diable

- à la suite de la publication de l'affaire *Cambridge Analytica*, **l'éditorialiste Meghan McCain** a noté que :

> *It happened under Obama, and it was lauded by the media as being genius. And now under the Trump campaign – it's the Cambridge Analytica scandal.*

- lors des campagnes pour les élections de 2008 et de 2012 les équipes de B. Obama ont en effet eu **un recours massif** à de nombreuses techniques
  - pour **profiler les électeurs** tout azimuth
  - *Obama for America* a nécessité un investissement technique de **100 m^{ns} \$** au dire du directeur de campagne Jeff Messina
  - ces équipes ont fait feu de tout bois, **notamment des données FACEBOOK** qui, là aussi, ont été utilisées pour **du profilage à l'insu des personnes concernées**
  - pour des finalités certes **un peu différentes**

    > *par exemple, profiler les contacts des utilisateurs de l'application pour produire un discours à fin de les convaincre de voter pour le président sortant*

  - mais le fait que les équipes de Trump aient peut-être **fait pire** ne veut pas dire que les équipes d'Obama aient nécessairement **fait mieux. . .**

> ***Note :*** *les échecs du parti Républicain de 2008 et 2012 en la matière étant d'ailleurs à l'origine de la création de la société Cambridge Analytica*

DisinfoLab

- étude sur **« l'affaire Benalla » sur twiter**

  à la suite de soupçons d'ingérences russes

- publiée fin juillet 2018 par l'ONG belge **DisinfoLab**
- accompagnée dans la foulée de **la diffusion de fichiers** comportant notamment **des inférences**

  - sur la propension à relayer de la « désinformation russe »
  - et sur les affiliations|proximités partisanes

- la CNIL a été **saisie** par de multiples utilisateurs de twitter figurant dans le fichier

  qui a elle-même saisie son homologue belge, l'Autorité de protection des données

L'avocat du Diable

- **difficile de se prononcer** au regard des éléments disponibles
- l'analyse juridique d'un traitement :
  - est **complexe**
  - et nécessite d'avoir **tous les éléments en main**

    p. ex. lors de l'enregistrement d'un traitement au registre d'un DPD
- en attendant la décision de la CNIL, on peut toutefois noter que :
  - les informations sur lesquelles se fonde les inférences sont **rendues publiques** par **les personnes elles-même**

    cf. p. 102 pour les conditions à remplir pour traiter ce type de données
  - ce type d'inférence est tout à fait **envisageable**

    notamment sous condition d'établissement de la finalité du traitement –singulier– et de la proportionnalité et de la pertinence du traitement au regard de sa finalité
  - clairement, les réactions n'avaient **pas été anticipées** par les auteurs de l'étude

L'avocat du Diable

- **notes (suite) :** quelques questions
  - s'agit-il d'un **profilage** ?
  - à quelle **fin** répondait la diffusion des inférences ?

    question subsidiaire : la précision de l'inférence (modularité)
  - le traitement ressort-il de **l'intérêt légitime** du responsable de traitement ?

    *« Une ONG qui se consacre à **la transparence** se sert de **données publiquement** disponibles concernant **des élus** (promesses faites à l'époque de leur élection et participation effective aux scrutins de l'assemblée où ils siègent) pour les classer selon le respect de leurs engagements.*
    *Même si l'incidence sur les personnalités politiques concernées peut être **considérable**, le fait que le traitement se fonde sur des informations publiques et se rapporte à leurs responsabilités publiques, ajouté à une finalité évidente de **renforcement de la transparence** et de la responsabilité, **fait pencher la balance en faveur** de l'intérêt du responsable du traitement »* Avis 06/2014 du G29
- ce cas illustre le problème de la rediffusion des données pour **la réplication**

  la réglementation pousse vers l'anonymisation complète dans ce cas, y compris pour les fins de recherche

Les fichiers Monsanto

- article publié par **le quotidien *Le Monde*** en janvier 2019
- sur la réalisation par Fleishman-Hillard et Publicis pour le compte **de Monsanto** :
  - d'un **« fichage illégal »** de personnalités selon leur position sur le glyphosate
  - d'une **« cartographie »** des acteurs du dossier

  selon les points de vue

  Note :
  - du point de vue de la réglementation, il s'apparente plutôt à **un profilage**
  - le traitement ne semble toutefois pas avoir été automatisé ?
- qui a conduit à différents dépôts de **plaintes**

  dont *Le Monde* auprès du parquet de Paris

L'avocat du Diable

En attendant la décision des autorités compétentes, on peut noter que :

- là encore, les informations sur lesquelles se fonde le profilage semblent avoir été **rendues publiques par les personnes elles-même**
- **le ∼profilage** n'est pas interdit (cf. p. 82)
- pas plus que **la représentation d'intérêt**

  question de l'intérêt légitime du responsable de traitement dans ce cas de figure

- les ONG semblent réaliser des **opérations similaires**

  sans parler des journalistes et des largesses que le cadre applicable leur procure

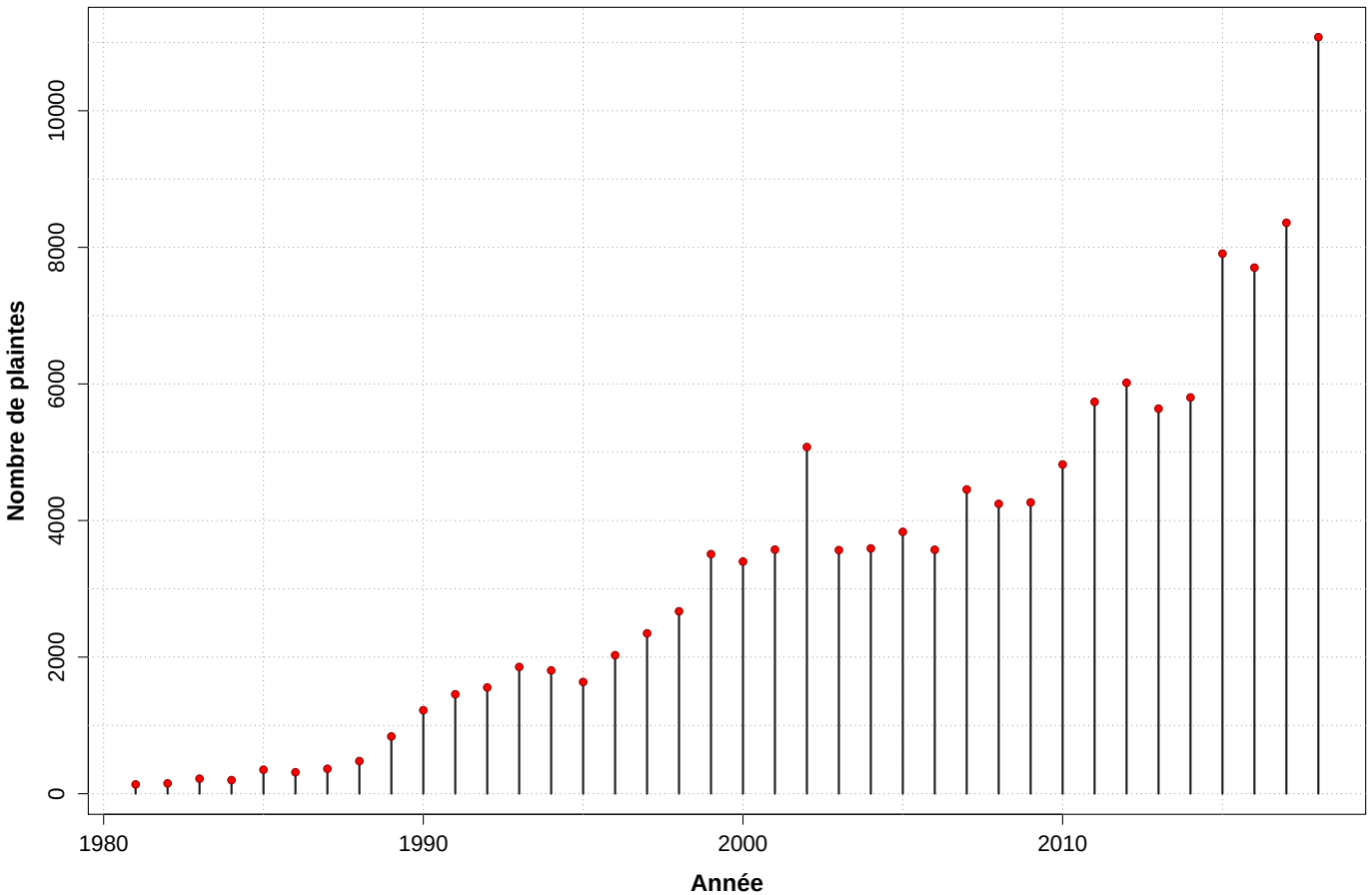
- l'émoi suscité provient sans doute plus de **la personnalité du** RdT que du traitement proprement dit

- **autres cas :** SNCF, FO, Ikea,. . .
- ces différents cas permettent de souligner combien :
  - la protection des données crée **un « aléas juridique »** (litote)

    sanctions par la CNIL, suites judiciaires,. . .
  - car la réglementation figure désormais en place de choix dans l'arsenal des guerres de positions juridiques
  - ces aléas peuvent frapper des RdT aux activités très **diverses**

    entreprises, ONG, syndicats,. . .
  - et peuvent avoir **leurs pendants en** SHS. . .

Évolution du nombre de plaintes déposées à la CNIL (1981-2018)



Source : Open CNIL

Préambule

- l'utilisation grandissante des données à caractère personnel depuis plusieurs décennies
- maintenant a conduit à la mise en place puis l'adaptation **d'un cadre réglementaire spécifique**
- principalement (mais pas seulement) du fait de **développements techniques** et, plus particulièrement,informatiques

  *et surtout du fait **des risques** qui découlent de leur utilisation*

- et **aux réactions** que leur réalisation (ou leur anticipation) peuvent susciter chez **les personnes concernées**

Préambule

- le volume croissant du traitement de données à caractère personnel est lié mais ne se **réduit toutefois pas**
  - à des développements **strictement techniques**

    comme la diffusion d'internet, de la téléphonie mobile ou la croissance de la puissance de calcul et des performances des algorithmes de classification,. . .
  - ou à des injonctions **juridiques**

    ou « bureaucratiques »
- mais renvoie à des questions :
  - **politiques** (surveillance par les États)
  - **économiques** (marchandisation toujours plus poussée dans le cadre d'une économie de marché mondialisée)

*les développements techniques sont le plus souvent **une réponse à une demande** (ou son anticipation)*

Préambule

- il ne faut jamais perdre de vue que la mise en conformité des traitements vise avant tout à

  la protection des personnes contre les aléas du traitement de données à caractère personnel

- et, une fois de plus, ne pas la réduire à une **injonction bureaucratique** des tutelles ou des financeurs
- il s'agit d'appréhender **les risques** pour les personnes
- en intégrant une démarche **de sécurisation** des données
  - protection des systèmes d'information
  - protection contre la (ré-)identification
- et ce, **dès la conception** du traitement (*Privacy by design*)

Préambule

- la question cruciale du traitement de données à caractère personnel est celle de
  - **la (ré-)identification des personnes**

    le terme de ***données (personnellement) identifiantes*** *est d'ailleurs sans doute plus parlant que celui de* DCP *mais sa portée est moindre*
  - et les **risques** attenants
- hors, la (ré-)identification ne nécessite le plus souvent **qu'une quantité d'information** étonnamment faible

  différentes études montrent que, même sans disposer d'information nominatives, quelques octets *sont souvent suffisants*
- du fait de l'extrême **variabilité** des populations et des milieux dans lesquelles ils évoluent
- qui rend **(quasi-)unique** les individus qui les composent (et donc aisément identifiables)

Préambule

- l'unicité peut prendre de **très nombreuses formes** :
  - p. ex., d'un point de vue **biologique**

    crêtes et plis papillaires, iris, réseaux veineux,. . .
  - qui a conduit au développement de la **biométrie**
  - mais aussi du point de vue **des comportements** individuels
  - qui a conduit aux développement de méthodes répondant aux mêmes fins que la biométrie mais en se fondant sur **les propriétés des individus**
- **la « masse »|généralité** des données collectées ne protège pas *en soi* de la réidentification

  question cruciale a prendre en considération lors de la rediffusion des données
- ces aspects seront abordés **en fin de présentation**

  dans la sous-partie La protection contre la réidentification, *p. 227*

La mise en place de réglementation

Chronologie de la réglementation sur les DCP

2018	entrée en application du règlement 2016/679 et fin du délais pour la mise en conformité pour les traitements en cours (25 mai) vote de la loi **protection des données**
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]

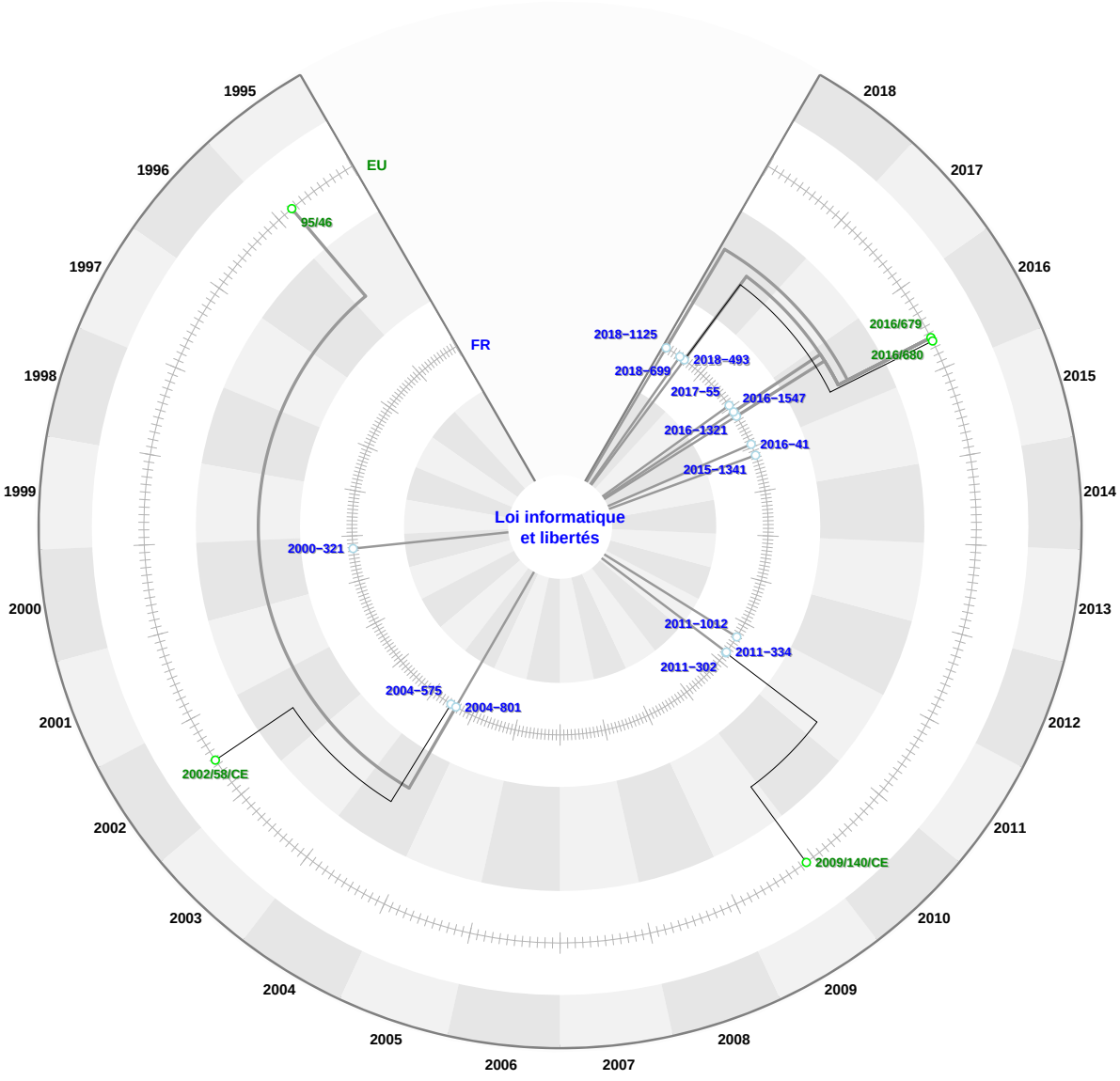
Autres réglementations

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,. . .

Chronologie des réglementations française et de l'UE (1995-2018)



EU

95/46 : Directive 95/46 du Parlement et du Conseil du 24 octobre 1995 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

2002/58/CE : Directive 2002/58/CE du Parlement européen et du Conseil du 12 juillet 2002 concernant le traitement des données à caractère personnel et la protection de la vie privée dans le secteur des communications électroniques

2009/140/CE : Directive 2009/140/CE du Parlement européen et du Conseil du 25 novembre 2009 modifiant les directives 2002/21/CE relative à un cadre réglementaire commun pour les réseaux et services de communications électroniques, 2002/19/CE relative à l'accès aux réseaux de communications électroniques et aux ressources associées, ainsi qu'à leur interconnexion, et 2002/20/CE relative à l'autorisation des réseaux et services de communications électroniques (Texte présentant de l'intérêt pour l'EEE)

2016/679 : Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 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, et abrogeant la directive 95/46/CE

2016/680 : Directive (UE) 2016/680 du Parlement européen et du Conseil du 27 avril 2016 relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel par les autorités compétentes à des fins de prévention et de détection des infractions pénales, d'enquêtes et de poursuites en la matière ou d'exécution de sanctions pénales, et à la libre circulation de ces données, et abrogeant la décision-cadre 2008/977/JAI du Conseil

FR

2000-321 : Loi n° 2000-321 du 12 avril 2000 relative aux droits des citoyens dans leurs relations avec les administrations

2004-575 : Loi pour la confiance dans l'économie numérique

2004-801 : Loi n° 2004-801 du 6 août 2004 relative à la protection des personnes physiques à l'égard des traitements de données à caractère personnel et modifiant la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés (1)

2011-302 : Loi portant diverses dispositions d'adaptation de la législation au droit de l'Union européenne en matière de santé, de travail et de communications électroniques

2011-334 : Loi n° 2011-334 du 29 mars 2011 relative au Défenseur des droits

2011-1012 : Ordonnance n° 2011-1012 du 24 août 2011 relative aux communications électroniques

2015-1341 : Ordonnance n° 2015-1341 du 23 octobre 2015 relative aux dispositions législatives du code des relations entre le public et l'administration

2016-41 : Loi n° 2016-41 du 26 janvier 2016 de modernisation de notre système de santé

2016-1321 : Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique

2016-1547 : Loi n° 2016-1547 du 18 novembre 2016 de modernisation de la justice du XXIe siècle

2017-55 : Loi n° 2017-55 du 20 janvier 2017 portant statut général des autorités administratives indépendantes et des autorités publiques indépendantes

2018-493 : Loi n° 2018-493 du 20 juin 2018 relative à la protection des données personnelles

2018-699 : Loi n° 2018-699 du 3 août 2018 visant à garantir la présence des parlementaires dans certains organismes extérieurs au Parlement et à simplifier les modalités de leur nomination

2018-1125 : Ordonnance n° 2018-1125 du 12 décembre 2018 prise en application de l'article 32 de la loi n° 2018-493 du 20 juin 2018

L'adaptation du droit français

- l'adaptation du droit français et, plus particulièrement, de la LIL a commencé dès 2016
- mais a surtout lieu en 2018
  - avec l'adoption en procédure accélérée de **la loi protection des données** en mai 2018
  - puis aussi par **un décret** du Conseil d'Etat le 1er août 2018
  - ainsi qu'une **ordonnance** du gouvernement le 12 décembre 2018
- le décret du Conseil d'Etat porte sur différentes **mesures d'application**
  - notamment **« en matière de traitements à des fins de recherche scientifique ou historique ou à des fins statistiques »**
  - en vertu de l'article de l'**art. 89** du RGPD (cf. p 101)
  - la question a donc été tranchée par **le juge administratif**

Les fins d'archives

- la situation contraste ici avec celle **des archivistes**
  - le projet de réglement a en effet conduit à une intense activité **de représentation d'intérêts**

    dont une pétition européenne (cf. après)
  - suite à des craintes pour leur activité du fait de certains droits des personnes (cf. p. 101)

    dans la première mouture du RGPD
  - et visant à préserver des dérogations déjà présentes dans la LIL
- à la suite de leur mobilisation, les archivistes ont obtenu **gain de cause**
- et les dérogations traitements à fin d'archives font l'objet d'un article **LIL art. 78**

  et c'est ce même article renvoie à **un décrêt du Conseil** d'État pour les fins de recherche scientifique ou historique, ou et (sic) à des fins statistiques

La pétition de l'Association des achivistes français

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**
- et faire de la protection des données un **droit fondamental** de l'UE

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**
- et faire de la protection des données un **droit fondamental** de l'UE :
  - art. 8 de la Charte des droits fondamentaux :
  - art. 16 du traité sur l'Union européenne (traités de Maastricht, Nice et Lisbonne)

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**
- et faire de la protection des données un **droit fondamental** de l'UE :
  - art. 8 de la Charte des droits fondamentaux :
    1. Toute personne a **droit à la protection des données à caractère personnel la concernant**.
    2. Ces données doivent être traitées **loyalement, à des fins déterminées et sur la base du consentement de la personne concernée** ou en vertu d'un autre fondement légitime prévu par la loi. Toute personne a le droit **d'accéder** aux données collectées la concernant et d'en obtenir **la rectification**.
    3. Le respect de ces règles est soumis au contrôle **d'une autorité indépendante**.
  - art. 16 du traité sur l'Union européenne (traités de Maastricht, Nice et Lisbonne)

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**
- et faire de la protection des données un **droit fondamental** de l'UE :
  - art. 8 de la Charte des droits fondamentaux :
  - art. 16 du traité sur l'Union européenne (traités de Maastricht, Nice et Lisbonne)

> 1. Toute personne a **droit à la protection des données à caractère personnel la concernant**.
> 2. Le Parlement européen et le Conseil, statuant conformément à la procédure législative ordinaire, **fixent les règles relatives à la protection des personnes physiques à l'égard du traitement des données à caractère personnel** par les institutions, organes et organismes de l'Union, ainsi que par les États membres dans l'exercice d'activités qui relèvent du champ d'application du droit de l'Union, et à la libre circulation de ces données. Le respect de ces règles est soumis au contrôle d'autorités indépendantes.
> 3. Les règles adoptées sur la base du présent article sont sans préjudice des règles spécifiques prévues à l'article 39 du traité sur l'Union européenne.

Une construction européenne

- le cadre européen reprend différentes notions élaborées dans **un cadre national**
- notamment, dans le cadre de la LIL
- mais aussi de la législation allemande
- dont elle reprend le principe **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**

Note : Cette notion a aussi été développée aux États-Unis par le juriste et politiste Alan Westin quelque années auparavant

Le principe d'autodétermination informationnelle

- le principe d'autodétermination informationnelle est **un principe fondamental** dont découle
  - certaines **obligations**

    comme l'information des personnes
  - ainsi que différents **droits**

    comme le droit d'accès, de rectification ou d'effacement ainsi que la limitation de la conservation des données
- à ce titre, il est ici important de souligner que :
  - lorsque la question de **la mise en conformité** des traitement est envisagée
  - l'attention se focalise (trop) souvent sur **le recueil du consentement**

    qui n'est nécessaire que dans un nombre limité de cas, les fins de recherche pouvant bénéficier de dispositions spécifique en la matière
  - alors que la question de l'information des personnes est (litt.) **primordiale**

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
  - mais aussi des archivistes. . .
- 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 :

- les services de l'État ont commencé à s'informatiser dans les années soixante

  Exemple : l'INSEE qui, dès lors, avait la possibilité technique de conserver des fichiers de données individuelles comme les bulletins de recensement et non plus des statistiques

- ces développements de l'informatique ont rapidemment suscité des débats sur l'opportunité de légiférer à ce sujet en France

  Exemple : ***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) :

  Exemple :

  - 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 par **Le Monde** le 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**

« L'affaire » Safari

Tandis que le ministère de l'intérieur développe la centralisation de ses renseignements

Une division de l'informatique est créée à la chancellerie

En ordre dispersé, les départements ministériels tentent de développer à leur profit, à leur seul usage, l'informatique et son outil, l'ordinateur. Ce n'est pas tout à fait un hasard si, à l'époque où le Journal officiel va publier un arrêté créant une « division de l'informatique » au ministère de la justice, celui de l'intérieur met la dernière main à la mise en route d'un ordinateur puissant destiné à rassembler la masse énorme des renseignements grapillés sur tout le territoire; pas un hasard non plus si le projet SAFARI (Système automatisé pour les fichiers administratifs et le répertoire des individus) destiné à définir chaque Français par un « identifiant », qui ne définissait que lui, maintenant terminé, est l'objet de convoitises ardentes; le ministère de l'intérieur y souhaite jouer le premier rôle. En effet, une telle banque de données, soubassement opérationnel de toute autre collecte de renseignements, donnera à qui la possédera, une puissance sans égale.

Ainsi se trouve d'évidence posé un problème fondamental, même s'il est rebattu : celui des rapports des libertés publiques et de l'informatique. Son importance exigerait qu'il en fût, au Parlement, publiquement débattu. Tel ne paraît pas être, pourtant, la solution envisagée par le premier ministre dans les directives qu'il vient d'adresser au ministère de la justice, intéressé au premier chef si l'on s'en rapporte à la Constitution qui dans son article 66 fait de l'autorité judiciaire le gardien des libertés individuelles.

« Safari » ou la chasse aux Français

Rue Jules-Breton, à Paris-13e, dans des locaux du ministère de l'intérieur, un ordinateur Iris-80 avec bi-processeur est en cours de mise en marche. A travers la France, les différents services de police détiennent, selon la confidence faite par un très haut magistrat, 100 millions de fiches, réparties dans 400 fichiers. Ainsi se trouve posées — et, à terme, théoriquement résolues — les données d'un problème comprenant, d'une part, l'énormité des renseignements collectés ; de l'autre, la méthode à définir pour faire de cet ensemble une source unique, à tous égards, de renseignements.

L'histoire du très puissant appareil qu'est l'Iris-80 est exemplaire du secret qui entoure l'épanouissement de l'informatique dans les administrations, quelles que puissent être les informations qui tiennent ici et là.

Puissant, cet Iris-80, une comparaison la démontrire sans contestation. L'appareil employé pour engranger les données de l'opération Safari, qui concerne l'identification indiquerait de l'ensemble des 52 millions de Français, a une contenance de 2 milliards d'octets (1) ; celle de l'ordinateur du ministère de l'intérieur est de 3,2 milliards d'octets.

C'est dire que la mise en route d'Iris-80 — dont la location coûte 1 million de francs chaque mois — a été précédée d'études, de tests pour en éprouver les possibilités. D'autant qu'à lui seul, il doit remplacer les trois GE 400 et le 10070 de la C.I.I. qu'employait jusqu'alors l'origine, budgétairement, n'était pas du tout prévu pour la tâche qu'il a finalement assurée, mais pour « traiter » les données administratives du Fichier national des constructeurs (F.N.C.). Il s'agit donc apparemment d'un détournement manifeste de crédits d'études, ce qui n'était sans doute pas le vœu du Parlement qui les vota.

De vastes ambitions

Il n'y a pas que cela. Le ministère de l'intérieur a d'encore plus vastes ambitions. Détenteurs, déjà, du fichier national du remembrement, les services de M. Jacques Chirac font de grands efforts pour, affirmation, s'en adjoindre d'autres : le cadastre, le fichier de la direction nationale des impôts et, plus grave peut-être, celui du ministère du travail.

De telles visées comportent un danger qui saute aux yeux, et que M. Adolphe Touffait, procureur général de la Cour de cassation, avait parfaitement défini le 9 avril 1973 devant l'Académie des sciences morales et politiques, en disant : « La dynamique du système qui tend à la centralisation des fichiers risque de porter gravement atteinte aux libertés, et même à l'équilibre des pouvoirs politiques. »

C'est si vrai que la regie nationale des usines Renault, par exemple, dispose déjà d'une base de données, établies à partir d'un fichier du personnel.

Ce n'est pas, pourtant, que les avertissements aient manqué. Le Conseil d'Etat en 1970, puis le ministère de la justice en 1970 (qui avait rappelé le rôle dévolu à l'autorité judiciaire de « gardien des libertés individuelles » et donc reclamé voix au chapitre) ont insisté sur la nécessité d'une intervention législative qui préciserait les quelques éléments essentiels de l'emploi de l'informatique appliquée aux particuliers : réglementation de l'accès des tiers aux fichiers, de l'interconnexion de ceux-ci, droit de rectification des personnes fichées si les renseignements retenus sont inexacts, etc.

De plus, tous les exemples étrangers incitent à ce débat sur une utilisation de l'informatique à laquelle, par définition, il ne s'agit pas de renoncer, mais à qui doivent être tracées des limites, si grand est le danger qu'elle implique. La déclaration par le gouvernement d'une commission de « sachants » dans les semaines à venir ne saurait suffire à remplacer le débat parlementaire dont on se méfie si volontiers.

Ce fait de débats parlementaires, il y a d'ailleurs des précédents qui sont le fait, précisément, du ministère de la justice et n'ont pas compromis le développement des fichiers. Avec le casier judiciaire, depuis longtemps, la chancellerie a l'expérience de semblables fichiers. Quel que soit le jugement qui peut être porté sur le principe d'un tel outil, il n'apparaît pas — sauf erreurs négligeables, relativement — que l'accès des tiers ou le droit à contrôle des personnes visées — par demande d'un extrait — ait jamais provoqué des heurts préjudiciables à la légalité.

De même, le fichier national des conducteurs, dans sa partie judiciaire, est prévu par une loi, et il faut regretter que les textes d'application ait permis des illégalités injustifiables — mais connues (*le Monde* du 8 mars).

« A la hussarde »

Fort, pourtant, de ces exemples, le ministère de la justice paraît curieusement se laisser dépasser par des querelles internes peu compréhensibles. L'arrêté signé le 18 mars par M. Jean Taittinger le montre. La création d'une « division de l'informatique », place Vendôme, serait en soi une bonne chose, du point de vue de l'efficacité, si les conditions de sa création, engagée vraiment voici trois mois, ne prenait l'allure d'une peu élégante tentative d'élimination dirigée contre certains esprits novateurs ayant eu le mauvais goût de s'intéresser trop tôt à l'informatique.

Il serait, en effet, bien étonnant que les membres de la commission de l'informatique au ministère de la justice, que préside M. Adolphe Touffait, ne s'effusquent pas d'une décision qui, en soi, ne peut avoir pour but que de « vider de sa substance » la dite commission. D'autant qu'il est d'ores et déjà connu que M. Touffait a été rayé de la liste des « sachants ». Il semble d'ailleurs que les réactions vives qui sont enregistrées portent moins sur le renouvellement des structures, jugées inévitables, que sur la méthode « à la hussarde » employée par tel membre de l'entourage de M. Taittinger pour mener à bien ses projets de rénovation de la gestion dans le domaine judiciaire.

Est-ce à dire de plus que les choix que l'on entend promouvoir seront nécessairement les plus opportuns ? Tout indique, pour l'instant, que, si le ministère de l'intérieur a définitivement choisi le « matériel lourd » pour s'équiper, la chancellerie, au contraire, s'oriente vers un réseau de mini-ordinateurs placés auprès de chaque tribunal de grande instance important.

Dans cet ordre d'idée, le choix déjà décidé de M. Jean Matteoli, vice-président à Bobigny (Seine-Saint-Denis), comme futur chef de la division de l'informatique (au point qu'il a, dès à présent, effectué des missions d'information à Lille, Nice, Lyon et Marseille dans les semaines passées), est significatif. Il est, en effet, à Bobigny l'apôtre d'un système « mini » qu'il souhaite étendre à l'ensemble de l'institution judiciaire. Ce n'est sans doute pas non plus par hasard si la télévision, lundi, après avoir donné des extraits du discours de M. Taittinger à Gap sur la justice civile et les nécessités d'un aggiornamento technique, a illustré son discours par un large reportage sur les équipements du tribunal de Bobigny — plus réduits, donc plus rapides à réaliser, ainsi plus vite source d'orgueil pour leurs créateurs.

C'est donc un doute grave qui pèse sur les intentions du gouvernement, en général, et du ministère de la justice, en particulier ; ce dernier département, qui rappelle à tous sa mission de protection des libertés individuelles, a apparemment accepté sans broncher la suppression d'un éventuel débat public, ce qui jette sur les déclarations « libérales » de M. Taittinger et d'autres domaines une suspicion qui n'est pas de bon aloi.

Mais, dans cette entreprise, le ministère de la justice, même s'il fait preuve d'une grande mollesse pour la défense de ses idéaux, car il ne s'agit pas seulement à présent de « protéger des délinquants », n'est pas essentiellement en cause. Ce qui l'est, c'est une entreprise dont on a tout lieu de suspecter la pureté tant on prend soin de cacher sa réalisation.

PHILIPPE BOUCHER.

(1) L'octet, ensemble de huit « bits », est l'unité de mesure de la puissance des ordinateurs. Quand on transcrit un texte dans la mémoire, chaque caractère du texte occupe un octet. Un milliard d'octets représente, en gros, la capacité de mémoire de cinquante bandes magnétiques.

L'article de P. Boucher publié dans *Le Monde* du 21 avril 1974

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 statistique 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**

  *dans ses **modalités** d'application que dans ses principes*

- 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 avant ça 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 (cf. p. 221),...

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

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**

    Exemples : *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** »*
- **Note :** la gratuité ne doit pas ici être confondue avec le le logiciel libre

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 G. 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

    dont certaines ont coûté cher en réparations à Google pour leur utilisation de DCP *(`Google Buzz`)*
- 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 à plus des 2/3 des PIB des pays de la planète (160 m^{ds} \$ en 2019)

    contre 136 m^{ds} \$ en 2018, 110 m^{ds} \$ en 2017 et 66 m^{ds} \$ en 2014
  - 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 c

  Bref, les techniques de surveillance sont toujours plus efficaces pour un coût toujours moindre

- la multiplication de ces traces offre donc des possibilités **de croisements inédites**

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

- et 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

RGPD : *Αποκάλυψις*

- même si l'entrée en application du RGPD a attiré **beaucoup d'attention**
- et a parfois été décrite en termes **apocalyptiques**
- le règlement s'inscrit pourtant dans **la continuité** de près de quarante de politiques publiques
  - différentes **obligations** sont parfois attribuées au RGPD de façon **erronées**

    un texte qui commence par « conformément au RGPD *» est souvent un mauvais signe*
  - les principes et de nombreuses dispositions restent **les mêmes** ou se voient renforcés
- la principale différence est plutôt que la réglementation est beaucoup plus difficile **à ignorer** aujourd'hui
- eut égard **au niveau de sanction** (p. 123)

- à l'instar de la LIL, les circonstances de l'élaboration du RGPD peuvent sembler **très éloignées** 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 DCP encore **largement négligée**, voire (sciemment) ignorée

  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

  Note : cette recension a été réalisée avant l'entrée en application du RGPD

Le traitements de données à caractère personnel 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

  Note :
  - l'anonymisation est souvent confondue avec la pseudonymisation
  - la pseudonymisation est en effet définie de façon précise dans le RGPD (cf. *infra* p. 197)

- 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

La charge de la normativité

- autre point (délicat)
- la loi doit être affectée **d'une charge normative**

> *« Non seulement la loi doit énoncer un impératif, mais encore cet impératif doit avoir une prise claire sur la réalité. Non seulement, la loi doit relever de la contrainte et non de l'invitation, mais ce degré de contrainte ne doit pas être laissé dans l'indétermination. »* Mathieu[2007]

- la réglementation sur les DCP **ne fait pas exception**
- l'application de la réglementation peut avoir un effet :
  - sur ce que vous pouvez **collecter** (et comment)
  - mais aussi sur ce que vous pouvez **en faire**
  - son application peut même prendre une dimension **épistémologique**
  - en deçà de toute **éthique**

> *voir notamment infra p 150 sur l'application du principe de minimisation*

- en conséquence, la présentation revêt un caractère nécessairement **normatif**...

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
  - **corollaire :** 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**)

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**

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

  Note : 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. 236
- 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 **délégués à la protection des données** (DPD)

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

    Exemple : pas de distinction entre **collecte**, **analyse** ou encore **publication**, pas de distinction de « personnes publiques »
  - et réciproquement (notamment en fonction de la finalité)

    Exemple : la minimisation des données
  - les définitions de données identifiantes, traitement, responsable de traitement, anonymisation, pseudonymisation,. . . **ne correspondent pas forcément** à l'idée que vous vous en faites
  - et ces différences peuvent avoir des implications **très concrètes**
- importance, aussi, **d'associer votre** DPD à vos projets de recherche
  - le DPD ne veille pas seulement à **la conformité des traitements** de DCP réalisés par le responsable de traitement
  - il a aussi une mission **de conseil et d'information**
  - cf. *infra* p. 125

La réglementation

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

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

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

La réglementation 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. 227)
  - 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. 204)
- plutôt que d'anonymat, il est donc préférable de parler **de possibilité de réidentification** des personnes

Traitement

RGPD **art. 4 § 2 :** 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

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.** *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. 108)

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 (RGPD **art. 9 § 1**) :
  - **l'origine raciale ou ethnique** (cf. p.172)

    « 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).
    La LIL fait, elle, référence à « la prétendue origine raciale ou l'origine ethnique »
  - **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
- auxquels s'ajoutent 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,. . .)
- sont **interdits**

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

Dérogations à l'interdiction de collecte des données sensibles (suite)

- 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)

  *avec, entre autres conditions,la **proportionnalité** à la finalité*

- cette possibilité a été reprise dans la LIL pour les traitements nécessaires à la recherche publique, sous réserve (**LIL art. 44 § 6**) :
  - que des motifs d'intérêt public important les rendent nécessaires

    dans les conditions prévues par l'**art. 9 § 2 (g)** : minimisation des données (cf. p. 95)
  - après avis de la CNIL

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

RGPD **art. 4 § 11** : 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

- **« 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 p. 100)

- **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**)

Toutefois, le consentement n'est **pas le seul fondement juridique** du traitement :

- le recueil du consentement n'est donc **pas toujours nécessaire** quand un autre fondement est mobilisable (cf. licéité *infra* p. 99)
- l'utilisation du consentement comme fondement peut aussi se révéler **contraignant** car il doit être conservé sur **le long terme**

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

Le recueil du consentement

- dans les faits, on peut observer **une trop grande focalisation** sur **le recueil du consentement**

  dans la représentation de la relation aux personnes concernées

- **en omettant** les autres aspects
- à commencer par **l'enregistrement** –obligatoire– du traitement **au registre du** DPD
- mais aussi **l'information** et **les droits** des personne

  - se passer du recueil du consentement est plus facilement envisageable que se soustraire à l'exercice du droit des personne
  - et, peut-être plus encore, à l'obligation d'information des personnes

Le recueil du consentement

- le recueil du consentement est **un traitement de** DCP
- et c'est parfois **le seul** dans certaines collectes...
- en effet, soit
  - par méconnaissance de la réglementation
  - respect d'un code professionnel (comme la **déclaration d'Helsinki**)
  - ou des critères imposés par une revue
- le recueil du consentement est parfois ajouté à **des traitements anonymes** ou seulement indirectement identifiants
- dans ce cas, **la seule finalité du traitement**, c'est le recueil du consentement

  le traitement doit alors faire l'objet d'un enregistrement uniquement du fait du recueil du consentement

Traitement ne nécessitant pas l'identification

À ce sujet, l'art. 11 du RGPD dispose que :

> RGPD **art. 11 § 1** : Si les finalités pour lesquelles des données à caractère personnel sont traitées n'imposent pas ou n'imposent plus au responsable du traitement d'identifier une personne concernée, celui-ci n'est pas tenu de conserver, d'obtenir ou de traiter des informations supplémentaires pour identifier la personne concernée à la seule fin de respecter le présent règlement.

> RGPD **art. 11 § 2** : Lorsque, dans les cas visés au paragraphe 1 du présent article, le responsable du traitement est à même de démontrer qu'il n'est pas en mesure d'identifier la personne concernée, il en informe la personne concernée, si possible. En pareils cas, les articles 15 à 20 ne sont pas applicables, sauf lorsque la personne concernée fournit, aux fins d'exercer les droits que lui confèrent ces articles, des informations complémentaires qui permettent de l'identifier.

- autre application **du principe de minimisation** développé après
- on ne collecte pas de DCP à la seule fin **de satisfaire** aux obligations de la réglementation

Traitement ne nécessitant pas l'identification

- quand on se retrouve dans l'entre-deux des données **indirectement identifiantes**

  Exemple : questionnaire diffusé via des listes -ie : sans invitations personnelles

- le croisement avec des données auxiliaires pourrait conduire à **des ensembles d'anonymats** trop réduits pour empêcher toute réidentification

  certaines combinaisons de caractéristiques correspondent à un nombre limité de personnes physiques

- donc, **la réglementation s'applique**
- mais **l'exercice des droits** des personnes devient –quasiment– impossible

  Droit de consulation : on ne va pas communiquer des renseignements sur une personne si on ne peut pas être certain que le demandeur est bien la personne concernée

- de par l'art. 11, il n'est **pas nécessaire** de collecter des informations supplémentaires pour que les personnes puissent exercer leurs droits

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
- les finalités de recherche **n'échappent pas** à ce principe
- en pratique, c'est un des aspects les plus **délicats** de l'application de la réglementation aux sciences sociales (cf. *infra* p. 150)

Licéité, loyauté et transparence de la collecte des traitements

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

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

  Note :
  - la licéité est une condition nécessaire mais **non suffisante**
  - une fin de recherche **ne suffit pas** *en soi* à rendre un traitement conforme
- **loyauté et la transparence** : la personne concernée doit être informée de l'existence du traitement et de ses finalités (c60) ainsi que de ses 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** – LIL **art. 1 § 2**, voir aussi plus haut p. 41)
- 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)

Droits 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**)

Droits des personnes

Toutefois,

- en cas de traitements à des fins de recherche scientifique ou historique ou à des fins statistiques (RGPD **art. 89 § 2**)
  - **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)
- concernant le droit d'effacement (RGPD **art. 17**) :
  - le RdT peut refuser de faire droit aux demandes d'effacement s'il est en mesure d'établir le fait que l'exercice de ce droit est susceptible de **rendre impossible ou de compromettre gravement** la réalisation des objectifs dudit traitement
  - le droit d'opposition n'est pas applicable dans l'hypothèse où le traitement (poursuivant des finalités de recherche) est fondé sur **l'exécution d'une mission d'intérêt public**

Droit des personnes

- le droit d'accès **ne s'applique pas** lorsque les DCP
  - sont conservées sous une forme excluant manifestement tout risque d'atteinte à la vie privée
  - et à la protection des données des personnes concernées

  Exemple : *questionnaires non-nominatifs (à défaut de non-idnetifiants)*
- conformément à la loi protection des données, les dérogations pour **la recherche** ont été déterminées par le décret pris en Conseil d'État du 1er août 2018
- le décret prévoit que :
  - les dérogations se limitent au cas où l'exercice des droits **aux articles 15, 16, 18 et 21** du RGPD
  - risqueraient de **rendre impossible ou d'entraver sérieusement** la réalisation des finalités spécifiques
  - dans ce cas, le RdT **prend des mesures appropriées** pour protéger **les droits et libertés** ainsi que les **intérêts légitimes** de la personne concernée

En résumé,

- pour les traitements **à des fins de recherche scientifique ou historique ou à des fins statistiques**
  - il est **possible de ne pas à exercer** des droits des personnes
  - si ces droits risquent de **rendre impossible ou d'entraver sérieusement** la réalisation des finalités spécifiques
- **question (ouverte) :** dans quels cas ces dérogations s'appliquent-elles ?

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 obligation **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**) :

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

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

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

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

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* RdT *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**)
- et ne sert pas à des **prises de décisions** à l'égard des personnes concernées (**Lil art. 4 § 2**)
- dans ce cas, **plus d'obligation d'information** des personnes (**Lil art. 79**, RGPD **art. 14 § 5 (b)**) si :
  - la fourniture de telles informations se révèle impossible ou exigerait des efforts disproportionnés
  - l'obligation [information des personnes] est susceptible de rendre impossible ou de compromettre gravement la réalisation des objectifs dudit traitement

La réutilisation des DCP à des fins scientifiques

- de plus (décret pris en Conseil d'État du 1er août 2018) :
  - les données issues de ces traitements [*à fins de recherche scientifique ou historique ou à des fins statistiques*] conservées par le responsable du traitement ou son sous-traitant **ne sont accessibles ou modifiables que par des personnes autorisées**
  - ces données ne peuvent pas être diffusées **sans avoir été préalablement anonymisées**
  - sauf si **l'intérêt des tiers** à cette diffusion **prévaut** sur les intérêts ou les libertés et droits fondamentaux de la personne concernée
- le droit de l'UE et le droits français prévoient donc un ensemble **de dérogations**

  pour l'utilisation des données **à d'autres fins** que celles pour lesquelles elles ont été collectées
- pour autant, ces dérogations **n'abolissent** pas tous les droits des personnes
- en pratique. . . ?

Modalités et agents de la protection des données

Le responsable de traitement

RGPD **art. 4 § 7 :** 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

- le responsable de traitement n'est **pas nécessairement** une personne physique
- le RdT est soumis à différentes obligations :
  - le responsable de 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 RdT est aussi responsable de **la sécurité des données**
- 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 (faute séparable des fonctions)

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 imposait jusqu'à présent 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 DPD 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** DPD

Protection des données dès la conception et par défaut

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

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

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

Analyse d'impact relative à la protection des données

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

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

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

Remarques

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

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

- les études d'impact illustrent de plus la spécificité de la réglementation sur les DCP :
  - la réglementation édicte **des grands principes**
  - **les modalités de son application** telles que la minimisation des données et, plus généralement, les mesures de protection à adopter doivent être déterminées **au regard du traitement**
  - en s'appuyant notamment sur **la doctrine** de la CNIL et **les recommendations** du G29
- manque **d'un référentiel** propre aux sciences sociales (cf. *infra* p. 162)

Destinataires de données

RGPD **art. 4 § 9** : 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

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

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

RGPD **art. 4 § 8 :** 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

- **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)

Notes :

- 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.***
- d'autre part, le RGPD (et la LIL) s'appliquent de manière **différenciée** dans l'outre-mer

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 avait 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

  > *dès l'entrée en application du* RGPD, *plusieurs dizaines d'actions de groupe ont été engagées auprès de la* CNIL, *notamment contre différents* GAFAM

- le RGPD ouvre de plus le droit à des réparations en cas de dommage (RGPD **art. 82**, **LIL art. 37**)
- la LIL avait **déjà été modifiée** en ce sens par la loi de modernisation de la justice du XXI^{e} siècle du 16 novembre 2016

Le délégué à la protection des données

- 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 a été remplacé par le **délégué à la protection des données** (DPD) à l'entrée en vigueur du RGPD
- les fonctions du DPD (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 DPD n'en est pas pour autant une émanation de la CNIL
- le DPD, 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. 114 et suivantes

La désignation du DPD

La désignation du DPD 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 DPD 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 DPD de l'université
- si le DU est personnel CNRS, il faut désigner le DPD du CNRS

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

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

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

Le délégué à la protection des données

Le DPD dans vos projets de recherche

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

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

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

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

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

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

Le DPD dans vos projets de recherche

- la première action consiste donc

À PRENDRE CONTACT AVEC SON DPD

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

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

Le DPD dans vos projets de recherche

- reproche fréquent du **manque de médiation** entre les chercheurs et les agents de la protection des données à caractère personnel Rossi et Bigot[2018]
- du point de vue adverse, plainte des DPD du **manque d'association** aux projets de recherche
  - les DPD sont souvent associés trop tard

    quand il le sont
  - toutes les informations nécessaires à l'instruction des dossiers ne sont pas toujours transmises

    ne soyez pas pudiques

Le DPD dans vos projets de recherche

- la première mission des DPD est **la protection juridique** de leur établissement
- dans l'ESR, les DPD peuvent avoir à prendre en charge de nombreux traitements

  RH, médecine du travail, scolarité, enseignements (normalement),. . .

- au niveau local (université) ou national (EPST)
- pour la recherche, problème n'est pas seulement **le volume** mais aussi **la nature** des traitements
  - de nombreux traitements hors-recherche sont **balisés**

    notamment par la doctrine de la CNIL
  - pour la recherche, les choses plus **difficiles**
  - **infinité** de traitements envisageables
  - il faut donc travailler **au cas par cas**
- l'instruction des traitements à fin de recherche est généralement **plus ardu**

Le DPD dans vos projets de recherche

- situations **contrastées** entre disciplines
  - les traitements de la recherche médicale sont très réglementés
  - mais peuvent s'appuyer sur la doctrine de la CNIL

    notamment les méthodologies de référence
- pour les SHS, **peu d'équivalents** (cf. p 163)
- faute de mobilisation

  les délibérations de la CNIL ne tombent pas du ciel
- ce qui **complique** l'analyse juridique
- et peut conduire les DPD à adopter des positions **très défensives**
- cadre juridique général, etc. . .

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

Les grands principes de la réglementation

Récap : les grands principes de la réglementation

Les grands principes de la réglementation sont :

- **information :**

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

- **limitation de la finalité :**

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

- **minimisation des données :**

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

- **limitation de la conservation :**

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

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

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

La mise en conformité du traitement

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

  la réglementation est **parcimonieuse** *dans ses principes (cf. p. 154)*
- **toute la difficulté** de la mise en conformité réside donc dans **la traduction pratique de ces principes**
  - la généralité des principes permet une certaine **souplesse** dans l'application de la réglementation

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

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

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

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

      des personnes, des produits et des systèmes de données ou de procédures
    - et encore **ses délibérations**
  - qui forment **la doctrine** de la Cnil
  - et n'ont **aucun caractère obligatoire**

La doctrine de la CNIL

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

    avec des nuances, particulièrement disciplinaire (-cf. démographie du fait de l'INED)
  - comme le montre **l'analyse thématique** des délibérations de la CNIL
- ce qui **complique** d'autant plus la mise en conformité
  - car les délibérations, préconisations,. . . portent le plus souvent sur des traitements **très éloignés** des sciences sociales
  - cf. *infra* p. 162

Au de-là de l'application de la réglementation

- de plus, comme la réglementation repose sur des **grands principes**
- tout **n'y fait pas** l'objet de dispositions

  tant s'en faut. . .

- car tous les aspects du traitement n'ont pas vertue à faire l'objet de disposition
- son application est donc pour partie **conventionnelle**

Conventionnement

- dans les cas de recherches associant des membres **de plusieurs équipes** différentes, il peut être nécessaire de rédiger une convention de recherche (cf.*supra*)

  ce qui constitue aussi une protection en cas de dissension au sein du projet

- le conventionnement peut aussi se révéler nécessaire dans la relation **au terrain d'enquête** :

  - il peut arriver que des fichiers soient **transmis** par des institutions mais aussi que des fichiers leurs soient transmis (éventuellement en retour)
  - là encore, ces transferts doivent faire l'objet d'une convention et, le cas échéant d'une information des personnes concernées par les DCP transmises

- mais aussi dans les situations de **sous-traitance** (cf.*supra*)

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

L'enquête en questions

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

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

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

L'enquête en questions

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

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

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

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

    notamment pour déterminer le DPD *compétent*

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

  publication, rediffusion de base de données,. . .

La fin justifie les moyens

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

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

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

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

Remarques

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

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

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

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**

  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 données à caractère personnel dans les enseignements de méthodes :

- l'enseignement est une mission de **service publique**, la collecte de données à caractère personnel 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 données à caractère personnel sont nécessaires à la démonstration et pourquoi elles sont nécessaires
- il faut donc **formuler** toutes vos hypothèses *a priori*

  Note : finalité 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

Le traitement

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

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

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

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

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

Information, consentement et droits des personnes (rappel)

La mise en conformité du traitement **ne se limite pas** au recueil du consentement :

- il faut que le traitement soit enregistré **au registre du** DPD (p. 125)
- les personnes doivent être **informées du traitement**, de sa finalité, de l'identité du RdT, des destinataires des données, . . . (p. 100, p. 41)
- les personnes ont aussi **des droits** (p. 101)
- de plus, le recueil du consentement n'est **pas toujours nécessaire** (p. 92)

  et peut même se révéler problématique

Note : ***des exceptions*** *à certaines obligations sont toutefois envisageables*

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 (-cf. minimisation des données p.98)

- 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. . .

Le principe de parcimonie

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

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

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

    simple et non simpliste

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

Le principe de parcimonie

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

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

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

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

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

    la simplicité est *relative*. . .

Critères d'information

- en statistique la parcimonie s'incarne dans différentes critères de sélection des modèles
- tels que
  - **le critère d'information d'Akaike** (`AIC`)

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

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

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

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

L'information des critères

- ces critères permettent de relier **complexité des hypothèses** et **quantité d'information** disponible pour les tester

  en mobilisant diéfférentes théories mathématiques de l'information : l'information de Shannon (cf. p. 215) et l'information de Fisher

- et illustrent le principe (simple...) que, **plus une hypothèse est complexe, plus elle demande de l'information pour la tester**

  - on ne peut tester que des hypothèses aussi complexes que les données le permettent
  - le rejet d'une hypothèse plus complexe se fait donc au regard des données disponibles – cf. réfutabilité
  - la complexification du modèle rajoute de l'incertitude

    arbitrage biais–variance

L'information des critères

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

  en lien avec la complexité des hypothèses et la variance des variables mobilisées pour la tester (puissance des tests)

- mais on est toujours **contraint à la parcimonie** au regard de l'information disponible

  - car le recueil de l'information a **un coût**

    pas seulement monétaire

  - qui **limite** la quantité d'information disponible globalement
  - cf. Soubiran[2017b] dans le cas des données auxiliaires

Parcimonie, proportionnalité et pertinence

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

  du moins sur le principe, la pratique peut se révéler plus ardue – cf. après

- la référence au principe de parcimonie permet aussi de **modérer** une application **trop drastique** du principe de minimisation
  - on observe en effet **rarement directement** ce que l'on souhaite mesurer

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

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

- Soubiran[2019a]

Les concupiscences

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

Fazit

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

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

Mais, en règle générale,

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

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

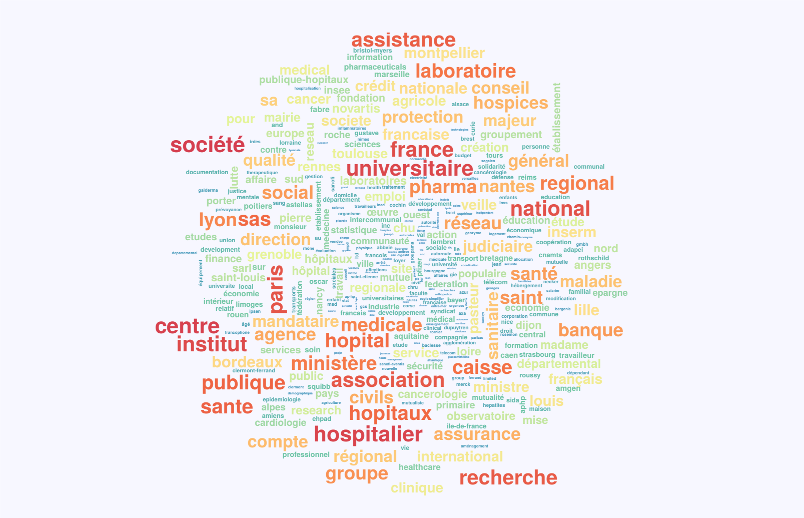
Note : il est important de souligner que ce n'est pas toujours le cas et que la proportionnalité et la pertinence des traitements peuvent être établis dans de très nombreuses situations

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

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

- les délibérations de la Cnil ont été classées de modèles thématiques (*Topic models*)
  - le terme de modèles thématiques désigne **un ensemble de méthodes** pas nécessairement probabilistes (factorisation de matrice)
  - pour **dégager** des thèmes (*topics*) à partir d'un corpus de textes
    - méthodes plutôt **issues de l'informatique**
    - dans une perspective **d'organisation de l'information** et non lexicale
  - l'objet est au final de regrouper des documents **similaires**
  - plutôt que de faire ressortir, p. ex. des « mondes lexicaux »
- la classification a été réalisée au moyen de **l'allocation de Dirichlet latente** (*latent Dirichlet allocation*) Blei, Ng et Jordan[2003]
- voir Soubiran[2019b] pour une version mise à jour

Allocation de Dirichlet latente

- chaque document est vu comme **un mélange de thèmes**

  même si, au final, on n'obtient que des probabilités plus ou moins marquées d'appartenir à une classe

- généralisation **de l'analyse sémantique latente probabiliste** (pLSA) :

$$p(w, d) = \sum_z p(z)p(d|z)p(w|z) = p(d) \sum_z p(z|d)p(w|z)$$

-

  Problèmes :

  - le nombre de paramètres de ce modèle **croît linéairement** avec le nombre de mots et de classes
  - risque **de surajustage** (*overfitting*)
  - le résultat ne peut **pas être généralisé** à d'autres documents

    car $p(c)$ est inconnu en dehors du corpus d'apprentissage

Allocation de Dirichlet latente

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
- tous les mots d'un document peuvent être issus (tirés) de plusieurs thèmes
- tous les mots documents ont leur propre distribution de thèmes θ_d

Allocation de Dirichlet latente

- étant donnés α et θ, la probabilité totale du mélange de thèmes est donnée par :

$$p(\theta_{1:D}, \phi_{1:K}, \mathbf{z}, \mathbf{w}|\alpha, \beta) = \prod_{k=1}^{K} p(\phi_k|\beta) \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{i=1}^{N} p(z_{di}|\theta_d) p(w_{di}|z_{di}, \phi_{1:K})$$

- la loi marginale

$$\begin{aligned} p(\mathbf{z}, \mathbf{w}|\alpha, \beta) &= \int_\theta \int_\phi p(\theta_{1:D}, \phi_{1:K}, \mathbf{z}, \mathbf{w}|\alpha, \beta)\, \mathrm{d}\theta\, \mathrm{d}\phi \\ &= \int_\phi \prod_{k=1}^{K} p(\phi_k|\beta) \prod_{d=1}^{D} \prod_{i=1}^{N} p(w_{di}|z_{di}, \phi_{1:K})\, \mathrm{d}\phi && p(\mathbf{w}|\mathbf{z}) \\ &\int_\theta \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{i=1}^{N} p(z_i|\theta_d)\, \mathrm{d}\theta && p(\mathbf{z}) \end{aligned}$$

permet de calculer les paramètres en s'appyant sur la conjugaison des lois multinomiale et Dirichlet

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

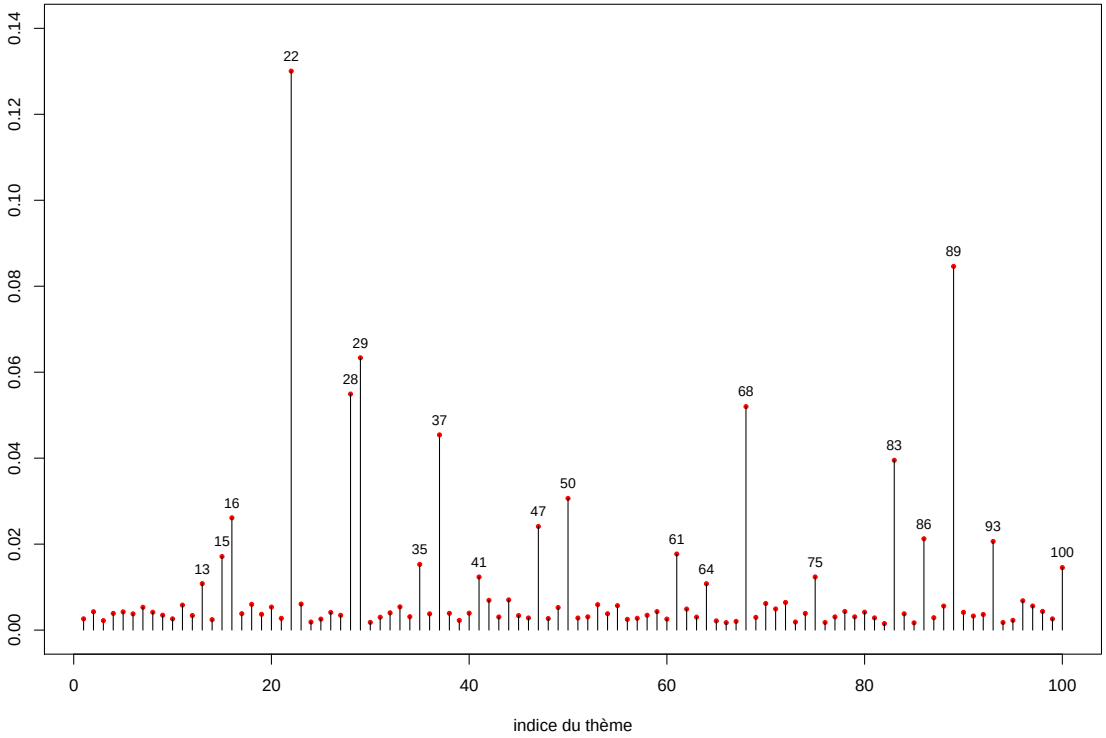
Résultats de la classification

- présentation des résultats provisoires d'un modèle **à 100 thèmes**
- sur **la (quasi-)intégralité** des +18 5000 délibérations de la CNIL de 1979 à janvier 2017

  quelques délibérations sont en effet manquantes

- à partir des fichiers `xml` des délibérations récupérés sur **data.gouv.fr**
- l'analyse porte donc sur **tous les types** de délibération :
  - autorisations (uniques, d'évaluation, de transferts, de recherche,...)
  - avis, recommandations
  - avertissements, sanctions
  - ...

Résultats de la classification



Résultats de la classification

- le nombre de thèmes témoigne de **la diversité** des sujets abordés par la CNIL
- poids des **autorisations de transferts de données** dans le total ($\simeq$ 33 %)
- et des recherches dans **le domaine de la santé** ($\simeq$ 29 %)

Note : le nombre de délibérations relatives au secteur santé en général est plus grand

CPAM, mutuelles, EPHAD, organisation du système de santé (téléservices, systèmes d'échanges de données des CHRU), les suivis médico-sociaux,. . .

- les traitement de **la statistique publique**, notamment en association avec certains EPST (INED, INSERM)
- ainsi que :

les traitements réalisés dans le l'exercice des fonctions régaliennes, les dispositifs de lutte contre la fraude, les dispositifs d'alerte professionnelle, la vérification des identités (dont biométrie),. . .

Les sciences sociales dans les délibérations de la Cnil

- les sciences sociales se distinguent surtout par **leur absence** du corpus

  quatre délibérations concernant les traitements de deux UMR (3 + 1)

- ces délibérations sont associées par le modèle aux enquêtes de **la statistiques publiques**
- dont les traitements comportent effectivement **des similarités** avec les traitements de sciences sociales
- toutefois, les traitements réalisés par les SSM sont régis par **un cadre spécifique**

  loi de 1951, Cnis

- et font aussi l'objet de dispositions spécifiques

  lors du traitement de données sensibles (art. 8), norme simplifiée n° 26

- de plus, ces traitements sont **loin de couvrir** tous les traitements réalisés en sciences sociales

Un point de doctrine

- l'effet des délibérations de la CNIL peut être illustré par **le recueil des « origines »**
- la liste des catégories particulières de données à caractère personnel (données sensibles) inclut notamment le « traitement des données à caractère personnel qui révèle l'origine **raciale ou ethnique** » (RGPD **art. 9 § 1**)
- difficultés spécifiques car, en France, c'est aussi une question **constitutionnelle**
- différentes décisions ont toutefois permis de **définir un cadre** (très contraint) autorisant la collecte de ce type de données

> *dont quatre délibérations de la CNIL relatives à l'enquête Trajectoires et origines (TeO) réalisée par l'INSEE et l'INED en 2008-2009*

Statistique des origines

- débat **récurent** depuis le début des années quatre-vingt dix et difficile à aborder sans fâcher qui que se soit
- importance de distinguer, tout d'abord, la question de **la nationalité** :
  - la statistique publique collecte la nationalité des personnes résidant sur le territoire nationale depuis la milieu du XIX^{e} siècle

    d'abord par intermittence puis de façon définitive à partir de 1886
  - l'INSEE collecte cette information de façon routinière dans ses enquêtes (p. ex. dans le TCM)
  - pratique qui a été validée par la CNIL à plusieurs reprises
- la collecte de la nationalité doit évidemment être traitée de façon **conforme**
- mais elle n'est **pas considérée** comme une donnée sensible
- et **ne pose donc pas** les problèmes spécifiques de la collecte d'informations sur « l'apparence extérieure » ou en référence à un « référentiel éthno-racial » ou encore les « statistiques éthniques »

L'interdiction constitutionnelle d'un référentiel ethno-racial

- lors de l'examen de la loi n° 2007-631 du 20 novembre 2007 **relative à la maîtrise de l'immigration, à l'intégration et à l'asile** (dite loi Hortefeux), **un amendement** fut voté proposant une modification de l'art. 8 la LIL
- il visait à soumettre les études de la diversité **à un régime d'autorisation** et non plus seulement de déclaration

> **Note :** cet amendement suivait une recommandation de la CNIL publiée quelques mois plus tôt

- cet amendement fut **censuré par le Conseil constitutionnel** après une saisine de députés et de sénateurs et ce, à la fois :
  - **sur la forme :** le Conseil constata **l'absence effective de lien** avec le projet déposé sur le bureau de l'Assemblée nationale qui portait sur le code de l'entrée et du séjour des étrangers et du droit d'asile (regroupement familial, asile et immigration pour motifs professionnels)
  - **mais aussi sur le fond :** le Conseil estima de plus que

> *« si les traitements nécessaires à la conduite d'études sur la mesure de la diversité des origines des personnes, de la discrimination et de l'intégration peuvent porter sur **des données objectives**, ils ne sauraient, sans méconnaître le principe énoncé par **l'article 1er** de la Constitution, reposer sur l'origine ethnique ou la race.*
> *Ces données objectives pourront, par exemple, se fonder sur **le nom, l'origine géographique ou la nationalité antérieure à la nationalité française**.»*

L'interdiction constitutionnelle d'un référentiel ethno-racial

- la décision du Conseil constitutionnel s'appuyait sur **l'art. 1er** de la Constitution :

> *« La France est une République indivisible, laïque, démocratique et sociale. Elle assure l'égalité devant la loi de tous les citoyens sans distinction d'origine, de race ou de religion. »*

- au regard de cette décision, il apparaissait que cette disposition ne concernait pas seulement l'égalité devant la loi mais **pouvaient être étendue** aux nomenclatures utilisées dans le cadre d'enquêtes
- et **limitait les données mobilisables** pour la mesure de la diversité des origines des personnes, de la discrimination et de l'intégration aux données administratives
- cette décision fut abondamment commentée et **un alinéa supplémentaire fut inséré** dans la version du 1er mars 2008

> *« Le Conseil n'a pas jugé pour autant que seules les données objectives pouvaient faire l'objet de traitements : il en va de même pour des données subjectives, par exemple celles fondées sur **le "ressenti d'appartenance"**.*
> *En revanche, serait contraire à la Constitution **la définition, a priori, d'un référentiel ethno-racial**. Telle est la limite constitutionnelle qui a été posée par la décision du 15 novembre 2007. »*

- autorisant donc l'utilisation **de données « subjectives »** pour ce type de mesures

L'interdiction constitutionnelle d'un référentiel ethno-racial

- la décision du Conseil constitutionnel a notamment été interprétée comme **le refus de la mise en place de « statistiques ethniques »** dans l'administration française DEBET et al.[2015]
- et la légalisation **d'une nomenclature préétablie** des « ethnies » et « race » ayant une valeur normative dans l'administration
- il est à noter que la décision du Conseil constitutionnel intervînt dans **un contexte tendu** :
  - le projet de loi sur l'immigration prévoyait la possibilité de pratiquer **un test ADN** sur les candidats au regroupement familial issus de pays dans lesquels « l'état civil présente des carences ou est inexistant »
  - suscitant **une vive polémique**

    Note : les décrets d'application n'ont finalement pas été signés
  - polémique, qui avait elle-même été précédée quelques mois auparavant par un autre débat tout aussi virulent autour du **projet d'enquête TeO** et de la question des « statistiques ethniques »

L'enquête TeO

- selon ses promoteurs, l'enquête TeO est « une enquête spécifiquement dédiée à **l'étude de la diversité des populations en France** et au thème des discriminations » et **leur impact** sur les trajectoires des personnes
- elle a fait l'objet **d'une autorisation** de la CNIL (délibération n° 2008-055 du 6 mars 2008)
- toutefois, la décision du Conseil constitutionnel a eu pour conséquence **la suppression de deux questions** sur la couleur de peau
  - « de quelle couleur de peau vous diriez-vous ? »
  - « d'après vous, de quelle couleur de peau les autres pensent-ils que vous êtes ? »
  - et cela, alors qu'il s'agissait de questions ouvertes sans aucunes suggestions
- par contre, figure dans la questionnaire une question invitant la personne interrogée à indiquer, **selon elle et au regard de son histoire familiale**, quelles seraient ses origines
- la CNIL a en effet considéré qu'il s'agisait **de données subjectives**
- et que cette question relevait donc de ce que le Conseil avait qualifié de **« ressenti d'appartenance »**

Mesurer la diversité

Pour la mesure de la diversité des origines des personnes, de la discrimination et de l'intégration, il faut donc :

- **obtenir le consentement de la personne** (sauf dans le cas où la loi prévoit que l'interdiction de traiter ces données ne peut être levée par le consentement), ou
  - justifier de l'intérêt public de l'étude
  - appartenir à la statistique publique (après avis favorable du Cnis et autorisation de la Cnil)
- **ne poser que des questions subjectives** sur le ressenti d'appartenance, de manière ouverte sans référence à une nomenclature spécifique
- ou en recueillant des informations considérées comme objectives

  comme le lieu de naissance et la nationalité à la naissance de l'intéressé et de ses parents, la ou les langues parlées,. . .
- la réponse doit avoir un caractère **explicitement facultatif**
- et prendre **les mesures de sécurité** appropriée

La portée de la décision du Conseil constitutionnel

- toutefois, toujours en 2008, la CNIL **a aussi autorisé** la collecte d'informations sur l'appartenance **à un groupe ethnique** dans la cas de l'enquête MAFE
  - la Commission a en effet estimé que l'enquête visait à analyser **les déterminant de la migration**
  - et que **la décision du Conseil constitutionnel ne s'appliquait donc pas**
- en 2009, la CNIL a de plus donné un avis favorable au rétablissement de la collecte de la communauté d'appartenance dans le **le recensement en Nouvelle-Calédonie**
  - avec les modalités « Européenne », « Kanak », « Tahitienne »,. . .
  - en **se fondant** notamment sur le décret n° 2003-485 du 5 juin 2003 relatif au recensement de la population et l'avis favorable du CNIS
  - **une question similaire** figurait dans les recensement précédents
  - et **avait été validée** par la CNIL pour des opérations de recensement précédentes
  - mais elle avait été **supprimée en 2004** à la suite de la réforme du RGP

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

La protection des données

La protection des données à caractère personnel

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

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

  protection physique ou logicielle contre les accès non autorisés aux données

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

  concerne les données elles-mêmes

- lors **des différentes étapes** du traitement

  collecte, conservation, analyse ou (re)diffusion

La protection des données à caractère personnel

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

La sécurisation des données

La sécurisation 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 données à caractère personnel ou de s'en débarrasser (moins de données à caractère personnel, 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 données à caractère personnel
  - stockages séparés des données à caractère personnel

    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**

  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 cryptographie asymétrique

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

    Exemple : le chiffre de César
    décalage du numéro d'ordre des lettres de l'alphabet, k correspondant au décalage
  - **asymétrique :** on génère deux clefs

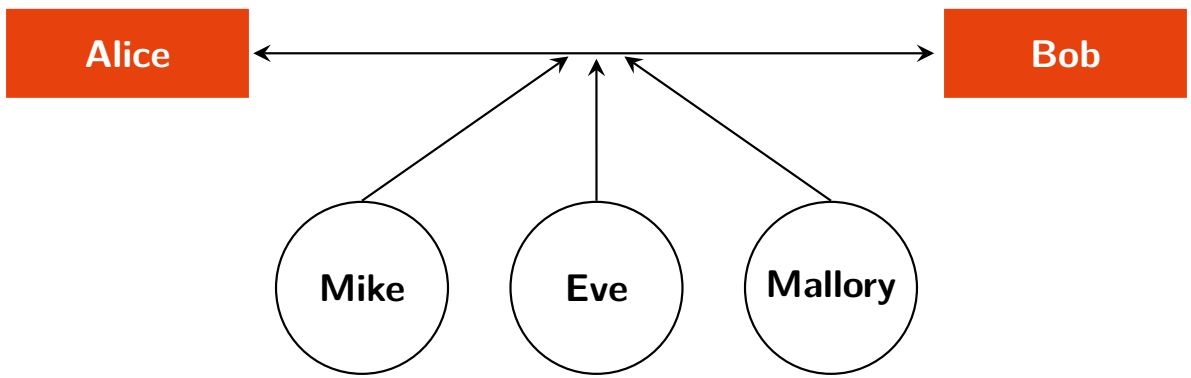
    Exemple :
    - une clef publique k_{pub} qui sert **à chiffrer** m
    - une clef privée k_{priv} qui sert **à déchiffrer** m
    - k_{pub} peut être diffusée sans restriction alors que k_{priv} doit restée cachée
- Exemple d'alogrithmes de chiffrement symétrique :

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

  `RSA, Diffie-Hellman,...`

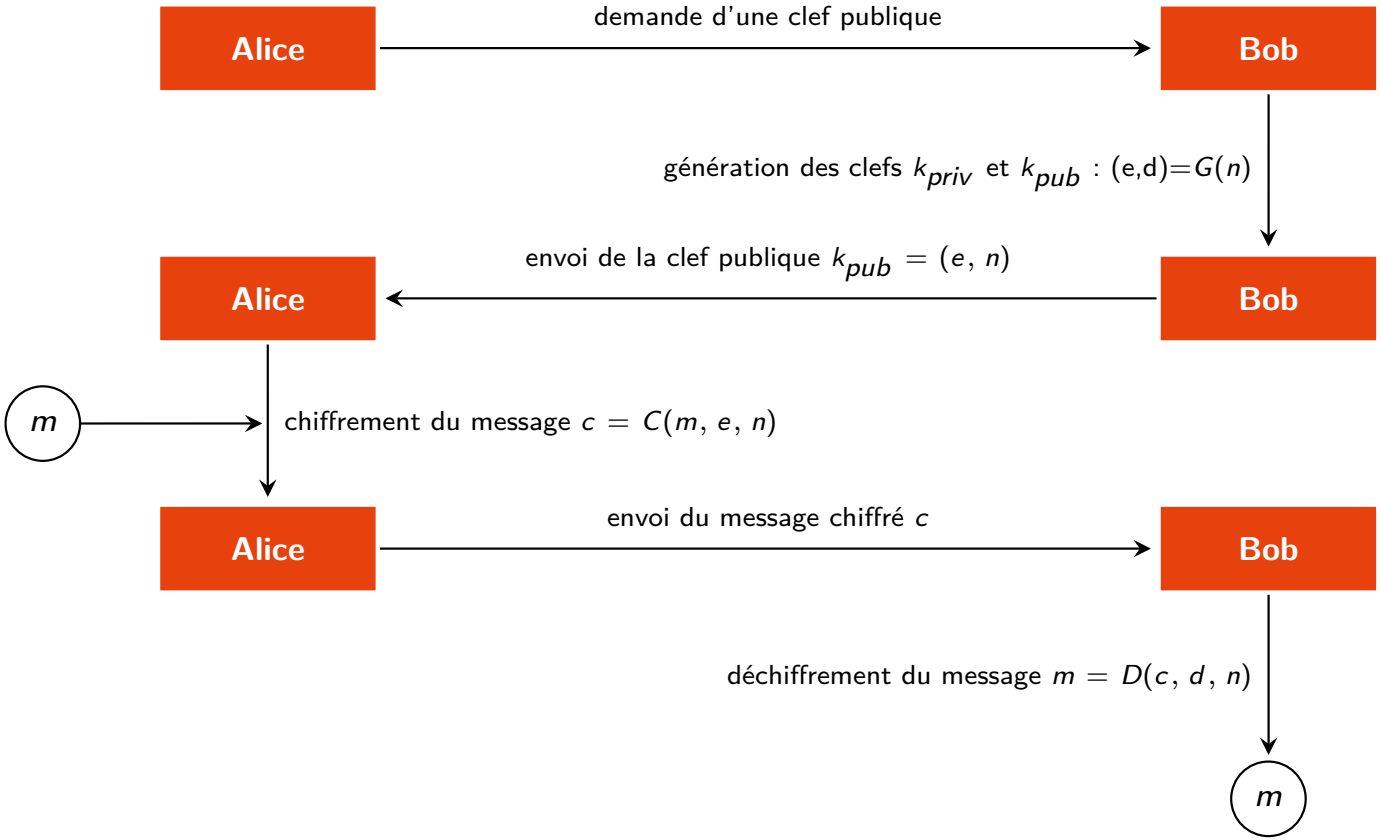
Alice, Bob, Eve et les autres

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



Chiffrement asymétrique : RSA

Alice veut écrire à Bob :



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

Cryptographie sommaire

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

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

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

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

- la racine carrée peut aussi s'écrire

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

- en effet,

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

Application

- `RSA` : chiffrement

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

- `RSA` : déchiffrement

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

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

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

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

Pour aller plus loin...

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

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

où

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

  avec p et q deux nombres premiers générés aléatoirement

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

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

La sécurité RSA

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

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

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

Notes :

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

Cryptographie hybride

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

La pseudonymisation

La pseudonymisation

RGPD **art. 4 § 5** : traitement de données à caractère personnel de telle façon que celles-ci

- **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

Pseudonymisation de BdD

En pratique,

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

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

- la table permettant la mappe entre les deux doit être **stockée à part**
- **Note :** pour attribuer un numéro pour identifier les individus, il est préférable :
  - d'utiliser un générateur de nombre pseudo aléatoire **de qualité cryptographique**
  - de réaliser **une permutation** $\sigma(\#\texttt{oid})$ avant l'attribution (sinon le nombre correspondra à la ligne et l'ordre permettra la réidentification)

La génération des identifiants publiques

- la génération des identifiants publiques doit être réalisée **avec le plus grand soin**
- pour écarter toute possibilité **de réidentification**
- **Exemple :** *MBS/PBS data*
  - en août 2016, le département de la santé australien a diffusé une fichier « anonymisé » renseignant **les remboursements médicaux** de 10% de la population

    soit $\sim$ 2, 9 millions de personnes
  - **les identifiants publiques** des patients et des fournisseurs de soins ont été générés de la façon suivante :

    1. générer un nombre pseudo aléatoire en utilisant l'identifiant comme graine (*seed*)
    2. extension de la longueur du nombre et mélange avec une partie de l'identifiant de départ
  - **Problème :** en devinant le `prng`, l'opération peut être **inversée** pour retrouver l'identifiant original Culnane, Rubinstein et Teague[2017]
- cette étude fournit de plus un exemple de **réidentification** à partir de données publiées (cf. infra)

La pseudonymisation

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

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

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

  nécessite au préalable une analyse morpho-syntaxique

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

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

Exemple : l'enquête MILITENS

MILITENS : enquête par questionnaires en ligne sur les enseignants des premier et second degrés à partir d'un échantillon national aléatoire stratifié tiré de la base de sondage de la DEPP

- qui a d'abord fait l'objet d'**une convention** avec la DEPP
- le transfert et la conservation **des informations** de contact sur des supports chiffrés
- gestion des invitations **distinctes** de la gestion des réponses (pas d'informations de contact stockées sur le même serveur que le gestionnaire d'enquête)
- conservation **des réponses et des traitement** sur un support chiffré
- **diffusion** des données auprès des membres du projet :
  - la cryptographie asymétrique
  - agrégation des données potentiellement indirectement identifiantes issues de sources externes à l'enquête (taille de l'établissement, informations sur le quartier issues du recensement)

Pseudonymisation des questionnaires en ligne

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

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

La « pseudonymisation » des entretiens

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

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

La « pseudonymisation » des entretiens

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

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

La 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`,...

La protection contre la réidentification

Données à caractère personnel

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

- **toute information**
  - y compris **« subjectives »**
  - prouvables **ou non**
  - quelles que soit **les caractéristiques** de la personne considérées
  - ou **le format**

    *écrit, numérique, photographique, sonore, avec une attention toute particulière aux **données biométriques***
- **personne physique identifiée ou identifiable**
  - la réglementation s'applique dès qu'il est **possible** d'identifier quelqu'un

    que cela fasse partie du traitement ou non
  - ne se **limite pas** aux noms

    données (in)directement identifiantes

Données à caractère personnel

- portée **volontairement large** de la notion de DCP
  - tout ce qui permet(rait) **de distinguer** une personne à l'exclusion de toute autre
  - distinction qui peut avoir **un effet** (majeur ou mineur) sur elle

    Exemple : profilage publicitaire
    - pas (nécessairement) de désanonymisation mais vise à produire un effet sur la personne visée (achat)
    - cf. infra empreinte digitale d'appareil p.221
- la défintion ne se limite pas aux données permettant l'identification mais inclut **l'ensemble des données** associées

  plus large que que des données personnellement identifiantes (*Personally Identifiable Information*)
- la réglementation peut s'appliquer **aux personnes décédées**
  - notamment du fait des relations familiales
  - associer une origine à une personne conduit à faire de même pour d'autres membres de sa famille (parents, frères, sœurs,. . .)

La protection contre la réidentification

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

  et pas seulement en mettant des mesures pour en contrôler l'accès
- le problème ne se limite pas aux entretiens mais concerne aussi les `BdD`
- différents études publiées montrent que, dans les faits, **une quantité limitée** d'information est nécessaire pour réidentifier les personnes
- et sans avoir recours à l'état de l'art en matière d'`Ai`

Réidentification

Quelques exemples **de réidentifications** publiés :

- *The Massachusetts Governor* (Latanya Sweeney)

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

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

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

- *The Netflix Dataset* Narayanan et Shmatikov[2008]

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

- *Riding with the Stars* (Antony Tockar)

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

- *Dark Data* (Svea Eckert – Andreas Dewes)

  réidentification de personnes à partir de leurs historiques de navigation

- *MBS/PBS data* Culnane, Rubinstein et Teague[2017]

  réidentification à partir de dossiers médicaux

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 ?

Ensembles d'anonymat

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
  - une classe d'équivalence est un sous-ensemble de S de la la forme :

    $$\{x \in S \mid x \sim a\}$$

    dont les éléments sont équivalents à a
  - partition de S en sous-ensembles selon une relation R

    toutes les classes d'équivalence sont soit égales, soit identiques
- 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

l'anonymat est d'autant plus fort que les caractères des personnes sont plus uniformément distribuées

- la taille de l'ensemble est **souvent insuffisante** pour garantir l'anonymat
- il faut aussi prendre en compte :
  - **la distribution** et **la variabilité** des données
  - mesurée par leur **entropie**

    les personnes n'ont pas les mêmes probabilités de présenter un ou plusieurs des caractères observés
  - et, de façon corrélative, **les 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
- l'anonymat ne se conçoit pas dans l'absolu mais **relativement à une situation et des données**

  *l'anonymat n'est pas un état mais une **relation***

Entropie

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

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

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

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

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

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

Entropie de l'information de Shannon

L'entropie de l'information de Shannon :

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

Interprétation de l'entropie de Shannon :

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

Exemples :

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

Note :

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

Entropie de l'information de Shannon

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

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

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

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

$$H(X) = \sum_{i=1}^{n} p(x_i)\mathbb{I}(X) = \sum_{i=1}^{n} p(x_i) log_2 \left(\frac{1}{p(x_i)}\right) = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i) \qquad (3)$$

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

Note : le terme d'entropie a été suggéré par J. von Neuman à C. Shannon de par la relation de (3) avec l'entropie de Boltzmann et parce que « personne n'y comprenait rien »

Entropie de l'information de Shannon

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

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

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

Identification des personnes

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

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

Identification des personnes

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

Exemples :

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

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

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

Empreinte digitale d'appareil

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

*par analogie à **la biométrie**, désigne un ensemble de techniques permettant de **pister** p. ex. la navigation d'une personne sur internet (connexions multiples à un même site mais aussi entre sites) mais sans laisser de traces sur la machine de l'utilisateur (`cookies`)*

- cette technique est lié au développement d'un web toujours plus **interactif et dynamique**
- lorsque `JavaScript` est activé sur un navigateur, la page chargée peut récupérer **un très grand nombre d'informations** de façon passive ou active :
  - **matérielles :** taille et résolution de l'écran, caractéristiques de la carte graphique via `WebGL`
  - **logicielles :** `OS`, navigateur, protocoles supportés, modules installés
  - **autres :** fuseau horaire, langue, polices, exécution cachée de code (canva -rendu 2d ou 3d-)

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

Empreinte digitale d'appareil

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

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

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

De plus,

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

Entropie des empreintes digitales d'appareil

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

Source : Eckersley[2010]

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

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

Entropie des empreintes digitales d'appareil

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

Source : Laperdrix, Rudametkin et Baudry[2016]

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

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

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

Entropie des empreintes digitales d'appareil

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

    comme c'est souvent le cas sur les données issue d'internet
- exemple d'identification **sans désanonymisation**
  - on attribue **une identité** à des personnes
  - via **leur caractéristiques**
  - et **non pas en leur attribuant un nom**
  - qui tombe dans **le champ d'application de la réglementation**

33 Bits of Entropy

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

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

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

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

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

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

The Netflix Prize

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

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

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

The Netflix Prize

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

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

- les données sont pourtant **éparses** (chaque individu n'a évalué qu'une infime portion de l'ensemble des films)
- l'entropie des données est donc **faible**

  la k-anonymisation (cf. *infra*) est ici impraticable

- mais c'est pourtant l'éparpillement des données qui va servir de fondement à la réidentification
- en utilisant la mesure de **similarité**

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

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

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

The Netflix Prize

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

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

The Netflix Prize

- la réidentification nécessite **des données auxiliaires** mais comme dans la plus part des scenarii d'attaques

  mais ce type de données peut être facile à obtenir (cf. `IMDb` ou *infra*)

- la réidentification de ce type de données peut paraître **vénielle** parce qu'elles ne renseignent pas des données sensibles

  - mais l'évaluations peut aussi être utilisée pour **inférer** l'orientation politique ou sexuelle des personnes
  - et donc déterminer les orientations de personnes si elles peuvent être réidentifiées

- de plus, l'algorithme n'est **pas spécifique** à `Netflix` ou `IMDb`

  A. Narayanan et V. Shmatikov ont appliqué une démarche similaire **aux réseaux sociaux** pour la réidentification d'utilisateurs de `Twiter` en les croisant avec des profils `Flickr`
  Narayanan et Shmatikov[2009]

Réidentification à partir d'attributs

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

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

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

- en utilisant **les listes électorales**, elle a aussi réussi à identifier **William Weld**, l'ancien gouverneur du Massachusetts dans une base médicale de séjours à l'hôpital des agents publics de l'État

  - six personnes à Cambridge avait la même date de naissance, trois étaient des hommes et une seule correspondant à son *ZIP code*

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

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

Unique in the Crowd

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

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

- et ne cherchant pas tant à **réidentifier** les personnes
- qu'à démontrer **l'unicité** des traces laissées par les utilisateurs du réseau
  - l'étude montre ainsi que **quatre points** choisis **au hasard** suffisent pour identifier 95% des personnes ($\epsilon > .95$)

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

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

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

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

Réidentification

- à partir de ces quelques exemples, on peut notamment remarquer que :
  - la suppression des noms **n'empêche pas** la réidentification
  - pas plus que la diffusion de données peu discriminantes **prises une à une**
  - ou la diffusion de **données éparses**
  - dans des bases renseignant les caractéristiques **d'un grand nombre de personnes**
  - **les données géographiques** sont de plus particulièrement discriminantes
- les études adoptent des point de vue **différents** (diversité et entropie des populations, unicité des individus)
  - mais elles ont ceci de commun de montrer l'hétérogénéité des populations est **le moteur de la réidentification**
  - en pratique, c'est cette hétérogénéité qui est **exploitée** pour le profilage des personnes à des fins commerciales, de surveillance, criminelles,. . .
- elles illustrent combien la réglementation applicable est **une protection contre des risques effectifs**

Modéliser la probabilité de 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** Machanavajjhala et al.[2007]
- mais plus l'entropie est forte, **plus la probabilité de réidentification** est importante (cf. *The Netflix Prize supra*)
- l'information auxiliaire dont peut disposer un attaquant, particulièrement **s'il appartient à la population**
- « *the culprit is auxiliary information* » Dwork et Roth[2014]

La protection contre la réidentification

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

  k-anonymity Sweeney[2002], *l-diversity*, *t-closeness*, *differential privacy* Dwork et Roth[2014], ...

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

    touche aux limites de la théorie de l'information ?

Libéralisation des données

- plusieurs des études reposent sur des données **rediffusées** par leurs producteurs
- ce qui illustre **les tensions**

  voire les injonctions contradictoires

- actuelles entre
  - **protection** des données
  - et **diffusion** des données
    - en *Open Data*
    - mais aussi pour la réplication de résultats
  - que ces données soient publiées ou seulement transférées à destinataires précis
- ces tensions se manifestent aussi dans le cas des publications de documents reposant sur des **données qualitatives**

Conclusion

Conclusion

- la réglementation **encadre** la collecte de données à caractère personnel 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*

Conclusion

- 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 constitue **une protection** contre cet aléas
- dans tous les cas, importance d'associer les personnes compétentes (DPD, RSSI) à vos projets de recherche
- et ce, dès **la conception du projet**
- et en intégrant la protection des données dans **le plan de gestion de données**

Merci pour votre attention

Bibliographie

Bibliographie I

Blei, David M., Andrew Y. Ng et Michael I. Jordan (mar. 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.

Culnane, Chris, Benjamin I. P. Rubinstein et Vanessa Teague (2017), « Health Data in an Open World », *CoRR*, abs/1712.05627, url : `http://arxiv.org/abs/1712.05627`.

Debet, Anne, Jean Massot, Nathalie Metallinos, Anne Danis-Fatôme et Olivier Lesobre (2015), *Informatique et libertés : La protection des données à caractère personnel en droit français et européen,* Lextenso édition, 1290 p.

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

Dwork, Cynthia et Aaron Roth (août 2014), « The Algorithmic Foundations of Differential Privacy », *Found. Trends Theor. Comput. Sci.* 3&#8211 ;4, 9, p. 211-407.

Eckersley, Peter (2010), « How Unique is Your Web Browser ? », PETS'10, Berlin, Springer, p. 1-18.

Bibliographie II

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.

Golle, Philippe (2006), « Revisiting the Uniqueness of Simple Demographics in the US Population », WPES '06, Alexandria, ACM, p. 77-80.

Laperdrix, Pierre, Walter Rudametkin et Benoit Baudry (mai 2016), « Beauty and the Beast : Diverting modern web browsers to build unique browser fingerprints », San Jose, IEEE Computer Society, https://hal.inria.fr/hal-01285470.

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

Mathieu, Bertrand (jan. 2007), « La normativité de la loi : une exigence démocratique », *Cahiers du conseil constitutionnel*, 21, URL : https://www.conseil-constitutionnel.fr/nouveaux-cahiers-du-conseil-constitutionnel/la-normativite-de-la-loi-une-exigence-democratique.

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

Bibliographie III

Narayanan, Arvind et Vitaly Shmatikov (2008), « Robust De-anonymization of Large Sparse Datasets », SP '08, Washington, IEEE Computer Society, p. 111-125.

– (2009), « De-anonymizing Social Networks », SP '09, Washington, IEEE Computer Society, p. 173-187.

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

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 (déc. 2013), "Selling Off Privacy at Auction", working paper or preprint, URL : `https://hal.inria.fr/hal-00915249`.

Quantin, Catherine et Benoît Riandey (2012), « Les techniques d'appariements sécurisés. De l'épidémiologie à la démographie », Chaire Quetelet, Louvain, Presses univ. de Louvain, p. 483-498.

Rossi, Julien et Jean-Édouard Bigot (2018), « Traces numériques et recherche scientifique au prisme du droit des données personnelles », *Les Enjeux de l'information et de la communication*, 19, p. 161-177.

Bibliographie IV

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.

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

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

Soubiran, Thomas (nov. 2017a), « La réglementation relative aux données à caractère personnel en sciences sociales », journées Data-SHS, `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 personnel*, Lille, 28 avr. 2017, `https://hal.archives-ouvertes.fr/hal-01589980`.

– (sept. 2019a), « La mise en conformité des traitements de données personnelles en SHS : les bases pour la négociation », colloque inaugural de la PUD-GA, 13 sept. 2019, `https://pro.univ-lille.fr/fileadmin/user_upload/pages_pros/thomas_soubiran/dcp/pudga2019-dcp.pdf`.

Bibliographie V

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

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

– (oct. 2002), « K-anonymity : A Model for Protecting Privacy », *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 5, 10, p. 557-570.