Réseau Quetelet

Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

La protection des données individuelles et la recherche en France et en Europe.

Évolution des cadres juridiques, critères et procédures d'accréditation, systèmes d'accès.

Roxane Silberman, CNRS, Directrice du Réseau Quetelet
Principal Investigator of the FP7 Data without Boundaries (DwB) project
DREES, 10 décembre 2014

Réseau Quetelet
Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

Introduction

- Données individuelles (micro-données)
  - Données personnelles (protection de la vie privée)
  - Données sur les entreprises ou autres entités individuelles (secret des affaires)

- Une évolution importante des cadres juridiques tant au niveau national qu'au niveau européen
- Une terminologie internationale et européenne en cours de consolidation
- Mais des contenus qui continuent à varier d'un pays à l'autre et d'un producteur à un autre, notamment quant aux règles d'anonymisation et à la perception du risque de ré-identification
- L'accès contrôlé: des modes d'accréditation et d'accès qui varient également et impliquent différents acteurs selon les pays
- Des problématiques identiques mais des évolutions séparées dans le domaine des SHS et de la santé

- Cette intervention:
  - Un bref aperçu général des évolutions et des points de discussions
  - Un focus sur un mode d'accès intermédiaire entre open data et accès sécurisé: Quetelet

Réseau Quetelet
Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

Sommaire

- Du « *sharing data* » à « *l'open data* »
- Les évolutions des cadres juridiques: grandes tendances
- Une terminologie plus précise et mieux partagée
- Mais des règles qui restent peu homogènes
- Critères d'accréditation et modes d'accès: quelques points de débat
- Entre *open data* et *accès sécurisé*: l'expérience de Quetelet

Réseau Quetelet
Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

DU « *SHARING DATA* » À « L'*OPEN DATA* », UN CONTEXTE NOUVEAU POUR LA QUESTION DE LA RÉ-IDENTIFICATION

- Le *Sharing data* de l'immédiat après guerre
  - Une problématique qui vient des chercheurs en science politique
  - Coût des données, réplication pour validation scientifique
  - Il s'agit essentiellement des enquêtes de chercheurs en sciences politiques
  - Peu de questions sur la protection de la vie privée, la ré-identification
  - La question de l'accès aux données statistique publique et données administratives très marginale ou absente (sauf recensements)
- Les progrès de l'informatique et *les premières lois sur la protection de la vie privée*
  - La question de la finalité de recherche généralement absente et peu de réactions
  - Une prise en compte progressive de la recherche
  - D'abord du côté de la santé (épidémiologie) avec des dispositions spéciales dont la trace persiste
  - La notion d'intérêt public apparaît inégalement
- L'*Open data* introduit la question d'autres acteurs
  - Les acteurs économiques
  - Déjà présent aux US (cf données payées sur argent public)
  - Les *big data,* élément également nouveau de ce contexte: données intéressant les acteurs économiques, données individuelles, détenteurs privés

Réseau Quetelet
Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

LES ÉVOLUTIONS DES CADRES JURIDIQUES EN EUROPE

- Une inscription juridique à des niveaux parfois très différents (lois, constitution, réglementation) et une articulation très diverse des différents cadres (archives, protection de la vie privée pour les données à caractère personnel, loi statistique publique, données administratives- , données fiscales etc…) selon les pays
- Une prise en compte progressive et générale de la finalité de recherche et de statistique (interprétation plus souple, modification des textes) pour l'accès aux données personnelles (et individuelles/entreprises)
  - Souvent au départ pour la santé (avec des modes d'accès relativement souples initialement)
  - Une généralisation ensuite de la possibilité d'accès pour la recherche (sans unification systématique des dispositions)
  - Qui s'est accentuée avec l'évolution des modes d'accès sécurisés
  - De nouvelles possibilités pour les appariements
  - L'accès transnational en progression (*silence de la loi*, notion de *cercle de confiance* reposant sur équivalences pour pénalités et sécurité, cf OCDE 2013-2014)
- Des différences notamment sur:
  - la notion *d'indirectement identifiable/ré-identification* par des *moyens que l'on peut raisonnablement mettre en œuvre* (ex. France vs Directive européenne 1995)
  - les termes : recherche, études, statistique, histoire, intérêt public …

Réseau Quetelet
Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

UNE TERMINOLOGIE PLUS PRÉCISE ET MIEUX PARTAGÉE

- Types de fichiers de données individuelles (ex. Réglement européen pour données Eurostat)
  - Public Use Files (PUF) / Fichiers grand public
    - *Super anonymisés,* parfois disponibles librement sur les sites web
    - *Si pas de PUF, des Campus Use Files ou Teaching Files pour l'enseignement (extractions)*
  - *Scientific Use Files (SUF)/ Fichiers scientifiques (équivalent des Fichiers dits de production et recherche dans la terminologie INSEE)*
    - *Plus détaillés, risque faible de ré-identification essentiellement lié à une détention d'informations extérieures*
    - *Accès contrôlé mais léger*
  - *Bespoke tabulations/Tabulations sur mesure, à façon*
  - *Secure Use Files (ScUF)/ Fichiers très détaillés ou fichiers dé-identifiés*
    - *Sur accréditation et accès sécurisé*

- Modes d'accès pour les ScUF
  - *On site*
  - *Remote execution ou job submission: pas d'accès direct aux données, contrôle de toutes les sorties*
  - *Remote access: données visibles mais pas de déchargement, contrôle des sorties finales*

Réseau Quetelet
Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

MAIS DES RÈGLES QUI RESTENT PEU HOMOGÈNES

- L'anonymisation
  - Pas de règles homogènes tant pour les données à caractère personnel que pour les données entreprises, pas de fondement très clair sur les règles
  - Tant pour les jeux de données que pour le contrôle des *outputs* (pour l'accès sécurisé)
  - Les *core variables:* forcément en nombre limité; quelle justification ?
  - Le problème de l'hétérogénéité (territoires, secteurs …)

- Et le risque de ré-identification
  - Une graduation du risque (aucun, risque faible, possible) qui détermine le statut des fichiers
  - Mais un débat sur une relativisation du risque
    - Risque de ré-identification et informations externes potentiellement disponibles
    - Risque cd ré-identification et professionnalité/intérêt du chercheur
    - Risque de ré-identification et sensibilité des données
    - Risque de ré-identification et intérêt public

- En dépit d'une terminologie partagée, des différences sur le contenu des PUF et des SUF

Réseau Quetelet
Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

CRITÈRES D'ACCRÉDITATION ET MODES D'ACCÈS: QUELQUES POINTS DE DÉBAT

- Critères d'accréditation et périmètre des utilisateurs
  - Recherche: des définitions et des appréciations variables selon pays, contexte et domaine
  - Institution vs chercheur vs projet de recherche
  - Publications dans journaux/éditions scientifiques vs publications
  - Recherche vs intérêt public
  - Recherche vs recherche et études
  - Exclusion de la finalité commerciale (quid si financement recherche, ex. financement CE ?)
- Procédures d'accréditation et instances de décision
  - Producteur vs comité spécifique vs conseil scientifique vs banque de données

- Modes d'accès
  - *Remote execution/job submission vs remote access*
    - *Le remote access en croissance*
  - *Remote access*
    - *Conditions ou pas sur les points d'accès (bureau fermé, data centers …)*
    - *Contrôle des sorties, systématique vs sondage vs responsabilité du chercheur*

Réseau Quetelet
Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

ENTRE OPEN DATA ET ACCÈS SÉCURISÉ: L'EXPÉRIENCE DE QUETELET

- Les « fichiers scientifiques » et les PSM à façon
  - Un accès contrôlé intermédiaire entre les Fichiers grand public et l'accès sécurisé (CASD)
  - La voie première pour les chercheurs
- Une accréditation plus rapide et légère que la procédure pour l'accès sécurisé
  - Des critères définis par un conseil scientifique et avalisés par les producteurs
  - Une délégation à Quetelet avec possibilité de recours au conseil scientifique et au producteur
  - Un critère central: la finalité de recherche appréciée sur la base d'une évaluation par les pairs
  - *Une procédure simplifié*e pour :
    - Statut de chercheurs, enseignants-chercheurs, post doc, doctorants, masters des universités et des établissements de recherche publics français et étrangers
    - Libellé court de la recherche
    - Signature d'un engagement (chercheur et responsable département ou établissement) sur finalité scientifique, protection des données, citation du producteur…)
  - *Une procédure sur examen* déléguée à Quetelet pour tt autre demandeur/établissement
    - Projet de recherche et objectif de publication
    - Part de la recherche dans l'établissement (évalué sur publications dans journaux/éditions scientifiques), financement recherche, étanchéité avec fonction de gestion
  - La possibilité de labelliser un établissement autre pour la procédure simplifiée (via avis du CS))

Réseau Quetelet
Réseau français des centres de données pour les sciences sociales
French Data Archives for social sciences

EN CONCLUSION

- *Des frontières de plus en plus poreuses*
- *Une progression forte des modes d'accès sécurisés*
- *Mais qui demande …*
  - *Un équilibre entre les différents types de fichiers et modes d'accès si l'on ne veut pas que l'accès sécurisé devienne la voie principale (et demanderait dès lors une autre approche en termes de moyens comme de procédures)*
  - *Des différences à prendre en compte selon les domaines en fonction de:*
    - *La sensibilité des données*
    - *L'Intrication de la recherche et des acteurs économiques*