Réseau Quetelet

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

Réseau Quetelet

Banques de données pour les sciences sociales

Roxane Silberman

CESSDA-Réseau Quetelet

PROGEDO

FREDOC, Aussois le 8 oct2013

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

Sommaire

- Le périmètre des données
- Les Data Archives : des origines à l'infrastructure européenne CESSDA
- La France et le Réseau Quetelet
- La participation à l'infrastructure européenne
- Les grandes fonctions de Quetelet
- Retour sur quelques questions anciennes et nouvelles

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

Introduction: le périmètre de Quetelet

- Réseau Quetelet: la banque française de données pour les sciences sociales
  - Une composante de la TGIR PROGEDO (Production et gestion des données en SHS) qui comporte deux dimensions
    - Banque de données (archivage, documentation, accréditation, accès) = Quetelet
    - Grandes enquêtes pluridisciplinaires, notamment participation aux enquêtes européennes ESS et SHARE
- Le périmètre
  - Données individuelles (s'oppose à données agrégées) : Personnes, ménages, entreprises…
    - Grandes bases de données permettant des approches quantitatives : Grandes enquêtes provenant de la recherche, de la statistique publique (INSEE, services statistiques), données administratives et de gestion (CNAV, Pôle emploi …), de la sphère privée (instituts de sondage, bases de gestions diverses)…
    - 0012 1 2 13 4 5 0 …
    - Données individuelles qualitatives (provenant généralement de la recherche)
  - Intersection importante avec les données santé publique/ épidémiologie, environnement, géo-référencées
  - Marginalement pour l'instant des bases de données macro (séries conjoncturelles, données financières …)
- Le partage de ces bases de données, une question déjà ancienne pour les sciences sociales
  - Mais un contexte international, européen et national qui a beaucoup évolué sous plusieurs dimensions, juridique, technologique, économique et culturel
- Des enjeux communs avec d'autres types de données et des enjeux spécifiques liés au caractère individuel des données
  - Protection de la vie privée pour les personnes
  - Secret des affaires et droit de la concurrence pour les entreprises
  - Des conséquences importantes sur l'accréditation et le mode d'accès

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

Aux origines des banques de données pour les sciences sociales

- L'accès aux grandes bases de données individuelles pour les sciences sociales, une question déjà ancienne
  - Le mouvement du « sharing data »
  - Le contexte de l'après-guerre dès les années 50/60
  - Sciences politiques et recherche comparative
- Les bases de l'argumentaire déjà présentes
  - Valorisation de données déjà existantes
    - Des données sous-utilisées (« analyse secondaire »)
    - Accumulation nécessaire pour la comparaison dans l'espace et dans le temps
  - Le coût économique
  - Réplication indispensable à la validation scientifique
- Ainsi que la plupart des grandes difficultés
  - Convaincre les producteurs (dont les chercheurs)
    - Délai de priorité
    - Valorisation des producteurs (citation)
    - Retour vers les producteurs et amélioration des données
  - Archivage sur le long terme …
  - Les métadonnées (des metadonnées de base (variable, question, échantillonnage … au contexte de l'enquête)
    - Le minimum indispensable de la part du producteur (travail est peu valorisé)
    - L'enrichissement des métadonnées
- La question de l'anonymisation et de la protection des données individuelles est peu présente initialement
  - Il s'agit essentiellement d'enquêtes anonymisées venant du monde de la recherche
  - L'accès aux données de la statistique publique et aux données administratives peu présent hormis les recensements
- Le périmètre des utilisateurs est celui des chercheurs

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

Des premières banques au …CESSDA

- Le mouvement du *« sharing data »* est à l'origine de la fondation des premières grandes banques de données pour les sciences sociales
  - Le Roper Centre aux US (institut de sondage)
  - L'ICPSR de l'Université de Michigan (Ann Arbor) créé comme consortium d'universités avec un réseau de Data Librarians
- Un mouvement qui s'étend en Europe: UKDA, ZA, NSD…dès les années 70
  - Au départ surtout les données du monde académique (enquêtes menées par les chercheurs) et qqs instituts de sondage
- Un réseau européen, le CESSDA dès 1976 (Amsterdam) formalisé en 1992 avec la préoccupation centrale de l'accès transnational (recherches comparatives)
  - Un portail avec un catalogue central
  - Un standard pour les métadonnées (DDI en 1999)
  - Un thésaurus
  - Un *« transborder agreement »*
- Le CESSDA (27 pays membres) est identifié comme infrastructure européenne potentielle dans la feuille de route ESFRI (2006/2008)
- Juin 2013 Création du CESSDA AS hébergé en Norvège (devant prendre le statut d'ERIC en 2015), signature de 13 pays dont la France

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

La France et le Réseau Quetelet

- Une entrée tardive de la France malgré la présence de chercheurs français dans les débats sur le « *sharing data » (Stoetzel, Boudon)*
- Comme ailleurs à l'initiative de chercheurs mais pas relayé au niveau institutionnel (CNRS) ni gouvernemental
  - Les ancêtres: deux laboratoires du CNRS, le LASMAS (1986, Paris), héritier du DAS/CES), et le CIDSP, Grenoble
- Vers une politique publique de données pour les sciences sociales
  - Le rapport Silberman *Les sciences sociales et leurs données,* 1999
  - Décret de création du Comité de concertation pour les données en SHS (CCDSHS) 12 février 2001, présidé par le ministre en charge de la recherche , associe plusieurs grands ministères détenteurs de données
  - L'article 10 confie au CNRS la mise en place d'une banque de données
  - Création du Centre Quetelet en 2002 qui devient le Réseau Quetelet en 2005 avec 3 partenaires en charge de différents domaines de données
  - Et un partenariat avec des plates-formes universitaires de données (PUD) en appui aux utilisateurs
- La consolidation
  - Inscrit dans la feuille de route française sur les TGIR (2008) comme composante de PROGEDO
  - Un nouveau partenaire, le CASD pour l'accès aux données confidentielles, avec un EQUIPEX (2011)
  - La participation à deux autres EQUIPEX (DIME-SHS dont données quali et données web) et D-FIH (données financières historiques)
  - Création de la TGIR PROGEDO au CNRS (2012)

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

La participation à l'infrastructure européenne

- La participation au CESSDA, le réseau européen des Archives de données
  - En pointillé jusqu'en 2001
  - Une participation formalisée depuis 2002
    - L'adoption de la norme documentaire internationale DDI et la publication sous NESSTAR qui permet le référencement dans le catalogue CESSDA
  - La participation à la phase préparatoire CESSDA PPP (6ème PCRD)
  - La signature de la France en juin 2013 et l'engagement du CNRS dans CESSDA
  - Des opportunités nouvelles pour la France mais aussi des requisits élevés et des questions
    - Un processus de certification des services providers
    - Quelle subsidiarité dans un système distribué ?

- La coordination du projet Data without Boundaries (DwB) 7ème PCRD qui construit la coopération entre le CESSDA et le Système Statistique Européen (INS coordonnés par Eurostat, BC coordonnés par la BCE)
  - 28 partenaires
  - Un accès transnational pour les données confidentielles de la statistique publique nationale et européenne
  - Un point d'accès unique avec un standard unifié pour les metadonnées en lien avec le portail CESSDA
  - Un réseau de centres d'accès sécurisés
  - Une accréditation européenne

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

Le fonds de données Quetelet

- Données nationales
  - Base de données représentatives
    - Statistique publique: Insee et SSM, bases de données administratives
    - Données issues de la recherche: INED, CEREQ, enquêtes socio-politiques de Science po…
    - Autres détenteurs: CERTU, IRDES, OVE, écoles d'ingénieur, instituts de sondage
  - Données qualitatives: BeQuali (EQUIPEX DIME-SHS) et à terme données web
- Données internationales
  - Membre du Luxembourg Income Study (enquêtes budget et patrimoines harmonisées) accès gratuit pour les chercheurs français via Quetelet
  - Membre de l'ICPSR, accès gratuit pour les chercheurs français via Quetelet
  - Partenaire de l'IPUMS (recensements harmonisés), MTUS (enquêtes emploi du temps harmonisées)

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

Le périmètre des utilisateurs Quetelet

- Metadonnées accessibles à tous
- Première exploration des données possible en ligne sous NESSTAR (tris simples et selon type de fichier, tris croisés, régressions simples)
- Microdonnées: accès après accréditation
  - Public Use Files: procédure simplifiée, téléchargement après enregistrement
  - Scientific Use Files: uniquement finalité de recherche, téléchargement après accréditation
  - Données confidentielles ou très détaillées (dé-identifiées), appariements de données, via le CASD en accès distant sans téléchargement (SdBox) et avec contrôle des sorties : accréditation par le Comité du secret statistique
- Le périmètre des utilisateurs recherche
  - Chercheurs et enseignants chercheurs, étudiants des masters, doctorants et post-docs des universités et centres de recherche
  - France, UE et pays associés, autres pays au cas par cas
  - Le problème de la définition du périmètre recherche

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

Les grandes fonctions de Quetelet

- Dépôt des données
  - Veille sur les sources de données, sélection (critère de qualité), convaincre les producteurs…
  - Licence de dépôts pour les chercheurs individuels, conventions pour les institutions détentrices de données (INSEE, SSM etc…)
  - Quetelet diffuseur, droit d'usage, encadré par la législation sur la protection des données individuelles, obligation de citation, metadonnées de bases …
- Archivage en lien avec les Archives nationales et plus récemment avec CINES
  - Vérification, conversion en différents formats
- Documentation
  - Mise au standard DDI
  - Enrichissement des metadonnées,
  - Routines, nomenclatures internationales harmonisées
  - Publication des métadonnées sous NESSTAR qui permet le référencement sur le catalogue du CESSDA et la recherche par questions et variables sur le portail Quetelet
- Diffusion de l'information: portail avec catalogue commun, base de questions et variables, première exploration en ligne, journées d'information …
- Accréditation selon les procédures en vigueur (CCDSHS) en fonction des types de fichiers demandés
- Mise à disposition des données en fonction des types de fichiers: téléchargement ou uniquement travail à distance sans téléchargement des données (CASD)
- Formation des utilisateurs pour l'accès sécurisé (CASD)
- Support aux utilisateurs et suivi des utilisations
- Retour aux producteurs

- Des compétences très diverses
  - Data manager, un métier qui requiert des compétences en statistique, en informatique, en sciences sociales

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

Retour sur quelques questions anciennes et nouvelles

- Les grandes bases de données quantitatives (individuelles et agrégées), un enjeu essentiel pour l'état et les politiques publiques, la démocratie, les acteurs économiques, les partenaires sociaux, les sciences
- Des acteurs multiples qu'il s'agisse des producteurs, des intermédiaires ou des utilisateurs
    - Producteurs :
        - la sphère gouvernementale (INS, SSM, Banques centrales, administrations, agences gouvernementales, collectivités territoriales, agences internationales, européennes…)
        - acteurs de la recherche publique et privée
        - acteurs économiques …
    - Utilisateurs:
        - acteurs gouvernementaux (national et international)
        - acteurs économique et partenaires sociaux
        - recherche
    - Intermédiaires:
        - producteurs eux-mêmes, archives nationales, banques de données à caractère public, banques de données à caractère commercial …

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

Un nouveau paysage

- L'évolution technologique
  - La croissance exponentielle des données: bases de données administratives, appariements, entrepôts de données, données web, *big data*
  - La multiplication des sources pour les mêmes données
  - Nouvelles possibilités de calcul et de traitements statistiques qui nécessitent l'accès à des données très détaillées
  - Nouvelles possibilités d'accès (tabulations en lignes, accès sécurisé distant (job submission et remote access)
- Une valeur économique accrue de la donnée : le nouveau contexte de *l'Open data*
  - Une multiplication des « banques » et des nouveaux opérateurs en retraitement de données, en développement d'applications, avec souvent une montée des coûts
  - Une pression sur l'accès qui s'accroît de manière générale
  - Des évolutions en sens contraires sur le caractère payant ou pas des données
- Une montée des questions sur la protection des données individuelles
  - Les lois sur la protection des données individuelles (loi sur les Archives, loi sur le secret statistique, loi sur la protection des données personnelles …)
  - Le relèvement des seuils d'anonymisation à partir des années 90
  - La prise en compte progressive de la finalité de recherche et son introduction dans les différentes lois archives, secret statistique, protection de la vie privée, code de procédure fiscale, données medico-administratives …
  - Le wab et les données individuelles

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

…qui bouscule les frontières

- Et repose les questions sur la conservation sur le long terme, les métadonnées, l'identification des données, la protection des données individuelles, les droits et les conditions d'accès, les infrastructures en matière de données

  Dans un contexte qui ignore de plus en plus les frontières nationales

➢ Des frontières plus floues et discutées sur le périmètre des utilisateurs: des définitions variables et discutées sur le périmètre recherche, des frontières qui s'estompent dans le contexte de *l'Open data*

➢ La question de la protection des données individuelles

Des évolutions en sens divers au niveau européen et international

- Le règlement sur l'accès des chercheurs aux données européennes Eurostat de juillet 2013 devrait faciliter l'accès aux données très détaillées via des centres accrédités nationaux (instituts de stat et à terme banques de données recherche) mais un processus lourd d'accréditation des universités utilisatrices
- Le projet de règlement européen sur la protection des données personnelles qui doit se substituer à la directive européenne de 95 est en l'état moins favorable à la recherche
- Les recommandations de OCDE sur l'accès transnational pour les données confidentielles: la voie du « *circle of trust* » (OCDE) = les équivalences en matière de conditions de sécurité et de pénalités (en cas de rupture de la confidentialité) et le transfert des responsabilités sur le pays de l'utiilisateur

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

Et la France dans ce nouveau contexte

- Des rapports nombreux et pour certains anciens (cf le rapport Braibant)
- Une évolution du cadre juridique sur la finalité de recherche: modification des lois Informatique et Libertés, loi de 51 sur le secret statistique dans le cadre de la loi sur les Archives, livre des procédures fiscales …
- Le nouveau contexte de *l'Open data et* ETALAB, service du premier ministre en charge de l'ouverture des données publiques et du développement de la plate-forme française de Open Data
- Des avancées qui n'ont pas été sans mal dans le domaine de la statistique publique, des données géo-référencées
- Des discussions toujours en cours dans le domaine des données medico-administratives
- Une très lente mobilisation des institutions de financement (ANR) et des universités et institutions de recherche en ce qui concerne les bases de données produites par les chercheurs en sciences sociales et une non reconnaissance du travail nécessaire à la valorisation de ces bases
- Des moyens encore largement insuffisants

Réseau Quetelet

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

Pour en savoir plus

http://www.reseau-quetelet.cnrs.fr/

roxane.silberman@ens.fr