

Diffuser en open data des données individuelles pour les recensements depuis 2006, et restreindre pour tous, et notamment pour les chercheurs, l'accès aux recensements antérieurs n'est pas cohérent.

Récemment un jeune chercheur rencontré à sa demande a évoqué les difficultés d'accès aux données des recensements classiques (en 1999 et avant). Des données qu'il faut des semaines, voire des mois pour acquérir et rendre cohérentes, quand elles pourraient être accessibles en quelques minutes si elles étaient disponibles en open data.

Avec son accord, nous reproduisons ses remarques (suite à la fourniture des fichiers par le centre Quételet) :

1- les dictionnaires de codes ne sont pas toujours mis à jour. Ainsi certaines variables figurant dans les dictionnaires n'apparaissent pas dans les fichiers Quételet.

2. On nous dit que les données géographiques n'existent pas dans le RP1999 au ¼, mais seulement dans le 1/20.

3. Les recensements rénovés après 2006 sont diffusés sous la forme de sous-fichiers thématiques (mobilité professionnelle, immigration...), mais nous ne pouvons pas avoir l'ensemble des informations désirées sur un même fichier. Donc impossible de les utiliser.

4. Le fichier SAPHIR nous paraît le plus pertinent pour notre sujet de recherche (je crois également pour beaucoup de sujets de recherche sociologiques et démographiques), mais ce fichier n'est disponible que sous forme de tableaux sur mesure, rendant impossible une exploration libre. De plus, les extractions de tableaux sont surtout payantes.

5. À cause de différences méthodologiques, nous avons mis beaucoup de temps pour trouver la correspondance entre les recensements successifs.

Quelques réflexions :

L'Insee met en ligne un « fichier harmonisé des recensements » pour la période 1968-2013 (<https://www.insee.fr/fr/statistiques/2414232>). Cependant ce fichier, malgré sa taille (38,7 millions d'enregistrements) est extrêmement pauvre, et donc totalement sans intérêt pour les chercheurs -car le niveau géographique le plus fin est le département (4 variables). Il contient 9 variables aux nomenclatures très peu détaillées (les 4 variables régionales proposées n'apportent aucune information supplémentaires et ne font qu'alourdir inutilement, la taille des fichiers). Par ailleurs, pourquoi continuer à proposer les 1968-1999, 1968-2006, -1968-2012 qui se ressemblent tant au lieu de mettre à jour un seul fichier comprenant une année de plus tous les ans ? Ces mauvaises pratiques dénotent un manque de cohérence et donc de réflexion dans la politique de diffusion de l'institution.

Ce fichier harmonisé est en fait une extraction du fichier Saphir, créé à la l'Insee-Alsace dans les années 90, et validé par la Direction générale en 2008 à fin de diffusion. Pour en savoir plus, on pourra consulter plusieurs documents figurant sur mon blog intégré au site de l'Association de prospective rhénane (APR). Parmi ces documents :

1) <http://www.apr-strasbourg.org/membre/2054/aubry-bernard/document/681-saphir-prsentation-partie-1.html> - document de travail décrivant les objectifs et la démarche de Saphir (août 1992)

2) <http://www.apr-strasbourg.org/membre/2054/aubry-bernard/document/605.html> - réflexions sur l'accès aux données de la statistique publique (annexe à un courrier adressé au Présidente de l'Autorité de la statistique publique).

3) <http://www.apr-strasbourg.org/membre/2054/aubry-bernard/document/679-les-immigrs-une-perspective-historique-1962-1999.html> - communication au colloque de l'Aidelf de Budapest en 2004 : application du fichier Saphir à l'étude de l'immigration (suivi sur longue période par génération).

4) <http://www.apr-strasbourg.org/membre/2054/aubry-bernard/document/699-le-service-public-en-recherche-quelle-modernisation.html>- pour ceux qui s'intéressent à la modernisation du service public . Comment l'insee s'est engagé à la fin des années 80 dans un « *Projet de service public* » qui incitait les agents à innover. Cette fenêtre ouverte pendant quelques années a permis à quelques agents de prendre des initiatives (d'où le projet Saphir), mais très vite, la hiérarchie a décidé d'étouffer les initiatives qu'elle n'avait pas elle-même suscitées. D'où probablement, la suppression de l'acronyme Saphir pour dénommer le fichier assez pauvre qui en est sorti et se trouve aujourd'hui en ligne sur le site de l'Insee.