4 janvier 2016

Chiffres pour tous

La diffusion des chiffres du recensement à l'heure de l'open data

Bernard Aubry APR

L'accès aux données mises en ligne par les institutions statistiques prend des formes variables à la fois parce que les publics sont différents, et parce que les attentes de chacun peuvent n'être pas les mêmes selon les usages qui sont faits des chiffres. En effet on peut souhaiter, dans le cadre de la composition d'un dossier ou lors de la rédaction d'un article, consulter le site afin de disposer de quelques chiffres, mais on peut aussi, dans un travail davantage orienté vers l'analyse statistique, souhaiter disposer d'une masse de données plus grande, par exemple, s'il faut pouvoir comparer des territoires entre eux.

En schématisant, l'information diffusée par les instituts nationaux de statistique prend deux formes. L'utilisateur accède soit :

** à des bases de tableaux déjà construits et mis en mémoire dans le système central de diffusion. C'est le cas le plus fréquent, mais les possibilités de croiser les variables entre elles sont toujours limitées.*

** à une ou plusieurs bases de données individuelles. Cette possibilité est généralement réservée aux seuls chercheurs. En France, l'Insee propose cependant en ligne quelques bases individuelles, ce qui – dans une certaine mesure – met alors l'utilisateur dans une situation comparable à celle des statisticiens de l'Insee dans leur capacité d'analyser les fichiers de façon approfondie. Avec les ordinateurs personnels, aujourd'hui très performants, nous sommes entrés dans l'ère de l'open data, ou si l'on préfère, des « Chiffres pour tous ».*

C'est autour de l'offre des fichiers détail, en open data, que l'on voudrait développer quelques réflexions. On en examinera plus particulièrement les avantages et les inconvénients, ainsi que les améliorations possibles.

Plan

1 - Les bases de données individuelles : pour accéder à toute l'information

La meilleure base de tableaux ne pourra jamais fournir qu'une infime partie de l'information dont un esprit curieux pourrait avoir besoin. Seules les bases individuelles, dénommées fichiers détail, permettent d'accéder, du moins en principe, à la richesse présente dans les fichiers informatiques détenus par les institutions qui exploitent la matière première, à savoir les bulletins de recensement et, plus généralement, les questionnaires d'enquêtes statistiques. Avec les progrès rapides dans l'exécution des programmes, l'accès aux données individuelles et leur traitement, encore relativement longs avec un ordinateur personnel, sont appelés à s'accélérer encore : quelques secondes pour produire des résultats portant sur l'ensemble de la France. Certes l'utilisateur doit disposer d'un logiciel de traitement de données volumineuses, mais ces logiciels se vulgarisent et leur coût deviendra accessible à de petites structures, voire à des particuliers. Dès à présent, l'Insee propose en ligne des fichiers détail suffisamment riches pour satisfaire de nombreuses exigences et, aux chercheurs, des fichiers plus détaillés encore.

Les fichiers détail des EAR (Enquêtes annuelles de recensement)

Sur son site, l'Insee propose actuellement plusieurs bases de données individuelles issues des enquêtes annuelles des recensements (EAR). Les fichiers les plus intéressants sont les suivants :

***un fichier régional** qui comprend une centaine de variables[1] et 20 millions de lignes. Il concerne toute la population enquêtée. Chaque individu est affecté d'une pondération (le plus souvent égale ou très proche de 1 pour les communes de plus de 10000 habitants et de 4 pour les autres, plus petites). L'unité territoriale est*

[1] *Une dizaine de variables sont propres aux départements d'outre-mer.*

l'ancienne région mais on y trouve aussi les départements et les EPCI (communautés de communes et d'agglomération) s'ils comprennent plus de 700000 habitants).[2]

Un fichier cantonal *(« Cantonville ») qui contient exactement le même nombre de lignes. Mais les variables, moins nombreuses, sont le plus souvent nettement plus pauvres en modalités. En revanche l'unité territoriale est le canton (et l'iris*[3]*, depuis 2007, pour les grandes communes), ce qui autorise des analyses locales.*

Un fichier des logements *qui contient seulement les variables relatives à la personne de référence des ménages. Il fournit également quelques informations sur le logement et ses occupants (nombre de personnes du ménage ayant 60 ans ou plus par exemple). En tout, une soixantaine de variables environ pour 24 millions d'enregistrements.*

Un fichier des migrations résidentielles[4] (une trentaine de variables)

Deux fichiers des navettes *(30 variables) pour les déplacements domicile-travail et un autre (20 variables) pour les déplacements des élèves et des étudiants.*

Chaque fichier EAR est l'agrégation de 5 fichiers de collecte qui représentent chacun 4 millions d'enregistrements (pour les fichiers région et cantonville). Ainsi le fichier EAR 2010 comprend les informations collectées en 2008, 09, 10, 11, 12. On retrouve donc dans l'EAR 2011 une partie des informations déjà présentes dans l'EAR 2010, à savoir 2009, 10, 11 (les pondérations sont toutefois modifiées à la marge pour assurer une cohérence temporelle).

Bien entendu, la tentation pour un statisticien est grande – et rien heureusement n'interdit de traiter librement les fichiers dûment mis en ligne - d'effectuer des opérations sur ces fichiers, en l'occurrence de créer un fichier unique permettant d'étendre le champ des applications. Il est en effet possible de fusionner ces fichiers et de généraliser à un échelon géographique plus fin les nombreuses variables disponibles au seul niveau régional. Certes, cette fusion s'opère en intégrant une incertitude aléatoire supplémentaire, mais cela n'a pas d'importance dès lors que les exploitations se font dans le cadre d'une investigation à caractère statistique, dans un esprit plus scientifique que polémique. En fait on ajoute une incertitude à une autre[5]. Rappelons que l'Insee ne met pas ***toutes*** les variables à disposition de ***tous***. Des fichiers plus complets sont réservés aux seuls chercheurs après signature d'une convention. La question est donc seulement de fixer une délimitation, ce qui n'est pas chose facile. Il y a toujours une part d'arbitraire dans la fixation d'une frontière et les règles évoluent dans le temps. A voir le nombre d'indicateurs à l'iris proposés aujourd'hui par l'Insee, on constate que les pratiques ont évolué depuis quelques décennies. Elles évolueront probablement encore avec le temps.

---

[2] *Voir en annexe les grandes lignes de la méthode de collecte et de traitement des données des recensements.*

[3] *Iris : découpage infra-urbain. C'est l'unité territoriale de diffusion la plus fine (environ 2000 habitants) pour les communes de plus de 10000 habitants et une partie des communes de 5000 à 10000.*

[4] *Jusqu'à présent, un seul fichier a été diffusé, celui de l'EAR 2008. Il concernait les mouvements migratoires entre le début 2004 et le début 2010 (période de cinq ans). Du fait d'un changement dans le questionnaire, plus aucun autre fichier n'a été diffusé depuis.*

[5] *On trouve dans les notes techniques mises en ligne sur le site de l'Insee l'ordre de grandeur de la précision des résultats. Ajouter une imprécision supplémentaire, faible et maîtrisée, c'est réduire encore les risques d'un usage abusif des résultats*

Dans le monde numérique qui nous attend, les enjeux de confidentialité seront plus déterminants en matière de *big data,* qui renvoie à de gigantesques fichiers d'origines disparates incluant souvent des données très personnelles, qu'en matière d'*open data,* qui concerne les fichiers issus du travail de l'administration contrôlés et qui sont contrôlés, notamment ceux des recensements de la population.

2 - Avantage majeur des fichiers détail : la liberté de croiser les variables

La palette des investigations potentielles à partir d'une base de données individuelles est infinie. Or les institutions officielles ne peuvent présenter tous les résultats ni effectuer des analyses sur tous les sujets. Elles sélectionnent ce qui leur paraît le plus important compte tenu de l'appréciation qu'elles se font des besoins de la société. Et l'on ne peut nier qu'il existe des freins dans la façon d'aborder certains thèmes. Parce que sortir des sentiers battus peut nuire à la carrière, parce que les fonctionnaires sont formatés pour ne pas faire de vagues, on préfère le conformisme à l'audace. Il n'est qu'à se pencher sur le choix des thèmes et des indicateurs retenus le plus souvent dans les publications de l'Insee eu égard à la diversité des variables collectées dans les bulletins du recensement.
En ce sens la question de la diffusion des chiffres sur l'immigration est emblématique d'une timidité ambiante. Déjà, au recensement de 1962, on collectait des données sur le pays de naissance et la nationalité des personnes dans une nomenclature détaillée des pays et l'on pouvait accéder à de nombreuses informations qui ne sont pas accessibles aujourd'hui sur le site. On a continué à collecter cette information aux recensements suivants et, bien que les séries historiques se révélaient cohérentes[6], on s'est gardé de promouvoir l'information, on la conservait sous le boisseau. L'autocensure, une certaine inhibition entretenue sans doute par la pression politique et sociale a eu pour conséquence une retenue dans la diffusion des données.

L'une des fonctions premières des recensements et des statistiques en général devrait être de servir d'instrument d'alerte. Ça l'a été parfois. Ainsi en 1963, l'Assemblée nationale débattait des projections de population à l'horizon 2000 qui prévoyaient pour l'Île-de-France une fourchette de 16 à 18 millions d'habitants. Les mesures prises à l'époque pour enrayer la croissance jugée excessive ont permis de stabiliser le chiffre de la population de la capitale autour de 12 millions d'habitants. Semblablement, n'aurait-on pas pu enrayer l'extrême concentration de la population d'origine étrangère dans certaines portions du territoire ? Si des universitaires, ou des institutions exerçant hors du champ de l'administration avaient eu accès aux fichiers individuels, peut-être auraient-ils pu développer des approches différentes de celles de l'Insee et alerter l'opinion et les médias sur la rapidité du processus de concentration qui a caractérisé certains quartiers

[6] *Communication au colloque de l'AIDELF (Association internationale des démographes de langue francise - 2004) : Les immigrés, une perspective historique ;*
http://www.apr-strasbourg.org/membre/2054/aubry-bernard/document/684-les-immigrs-une-perspective-historique-1962-1999.html

au cours des dernières décennies. Trop souvent, l'Insee ne diffuse que des effectifs et des pourcentages[7]. Une plus large imprégnation statistique de la société aurait peut-être aidé à appliquer aux résultats des recensements des indicateurs adéquats (coefficients de GINI, indicateurs de dissimilarité, de voisinage, etc).

3 – Un outil pour la connaissance de son environnement

Un autre exemple, moins polémique, est celui de l'utilisation de la statistique publique pour la connaissance de son propre territoire. On imagine très aisément comment, encadrés par leurs enseignants, les élèves, plus tard les étudiants, pourraient apprendre à mieux connaître leur environnement immédiat en manipulant des fichiers détail, évidemment adaptés à cet usage : découverte du territoire, détermination des caractéristiques par rapport à des territoires voisins, etc. Au-delà, on aiderait à la formation civique des jeunes en les initiant aux difficultés que rencontrent les maires et plus généralement les responsables dans leurs efforts pour résoudre les problèmes d'organisation des territoires (favoriser l'emploi, assurer les services aux personnes âgées, etc.).

4 - Comprendre le système statistique, *son potentiel autant que ses faiblesses*

Pour la plupart d'entre nous, ce qui se passe entre la collecte et la diffusion des résultats est mystérieux. Savoir ce qu'il y a dans un fichier, apprendre à le disséquer, mettre la main à la pâte[8] : tout cela est indispensable si l'on veut comprendre ce qu'il y a derrière les résultats. On diffuse des chiffres sur les cadres supérieurs, les professions intermédiaires…, mais que recouvrent les statistiques et surtout leurs variations dans le temps ? L'utilisateur n'en sait rien et on ne l'aide pas. Il y a en effet dans chaque catégorie d'une nomenclature des emplois qui, relevant tantôt de la gestion publique et tantôt du privé, ne sont pas soumis aux mêmes contraintes. Grâce aux fichiers détail, on peut décortiquer à son gré les sous-groupes, voir ce qui les distingue les uns des autres etc. S'il le faut, on peut composer sa propre nomenclature. En réunissant des sous-groupes de professions ou d'activités économiques, on pourra sélectionner par exemple les emplois « créatifs », les activités « vertes ». On pourra s'intéresser à l'évolution de territoires singuliers, pour lesquels il n'existe pas de nomenclature reconnue ou, s'il y en a une, effectuer des variantes : communes de montagne, du littoral par exemple. Il peut être aussi utile de vouloir chercher à maîtriser les chiffres, en étudiant la sensibilité des résultats aux choix qui ont été faits. Ainsi on peut, de façon très légitime, remettre en cause le découpage en aire urbaine dans les zones où les communes et les bourgs s'enchevêtrent, formant un système urbain complexe. D'une façon générale, il faut pouvoir réunir ou au contraire dissocier les ensembles existants.

Cette liste des avantages procurés par la mise à disposition d'un fichier détail n'est pas limitative.

---

[7] *Quel intérêt y-a-t-il de calculer une densité moyenne pour l'Egypte dont la population est concentrée sur une bande de terre ?*

[8] *Cela vaut pour tous, notamment pour les agents dont c'est le métier.*

5 – Quelques critiques

Une fois saluée la décision de mettre en ligne des bases de données - décision dont on peut imaginer qu'elle n'a pu se faire qu'au terme de longues réflexions compte tenu des enjeux - un certain nombre d'insuffisances et d'incohérences dans son application apparaissent.

Des fichiers disjoints. Est-il bien nécessaire de dissocier l'information en créant plusieurs fichiers? Si la diversité des variables contenues dans le fichier régional est satisfaisante pour des analyses approfondies pour l'ensemble de la France et par régions, pourquoi l'analyse locale est-elle contrainte ? En effet, ce ne sont pas tant les régions qui bougent que les territoires infrarégionaux. Certaines aires urbaines sont à cheval sur plusieurs régions : ainsi celle de Paris, qui dépasse les limites de l'Ile-de-France, mais aussi quelques autres comme celle d'Avignon (450 000 habitants) qui, majoritairement située en région PACA, s'étend également dans la nouvelle région Languedoc-Roussillon-Midi-Pyrénées (10% de la population). Sur un autre plan, la maille canton ne présente pas beaucoup d'intérêt puisqu'elle est trop grande pour permettre, par agrégation, la reconstitution d'une aire urbaine[9].

L'absence de recul historique. Il est impossible de raccrocher les fichiers détail de l'EAR aux mêmes fichiers constitués sur la base des recensements classiques (jusqu'en 1999). Or l'ancrage historique est important à connaître pour établir un diagnostic territorial.
Certes, l'Insee a mis en ligne un fichier historique pour la période 1968-2012, mais le fichier détail proposé ne contient qu'une maigre sélection de variables. Le niveau territorial le plus fin est le département, ce qui est manifestement insuffisant (cf. ce qui a été dit plus haut sur les territoires en mouvement).

La discontinuité géographique qui perturbe les analyses effectuées sur les fichiers communaux. Certes, entre 2006 et 2012, il n'y a eu que peu de modifications territoriales (quelques fusions de communes, souvent de petite taille, à l'exception notable de celle de Dunkerque[10]), mais les réformes territoriales en cours vont conduire à n'en point douter à accélérer les regroupements de communes. Si l'on ne fait pas un effort pour maintenir le maillage statistique communal pour suivre et rétropoler l'information dans n'importe quel territoire, le suivi dans le temps du contenu des cartes géographiques deviendra problématique. Il n'est pas anormal, en effet, que des élus ou des responsables

[9] *En revanche les fichiers migrations, composés au niveau des communes, pallient en partie à l'insuffisance évoquée.*
[10] *La commune de Dunkerque est l'une de celles qui a connu les plus fortes discontinuités territoriales avec, en 1970, l'absorption de la commune de Malo-les-Bains puis, en 1972, celles de la Petite-Synthe et de Rosendaël, enfin, en 2011, celles de Saint-Pol-sur-Mer et de Fort-Marly.*

territoriaux établissent, de leur propre initiative, des variantes pour anticiper les conséquences des projets de fusion qui concernent leurs propres communes.

La documentation est inadaptée. Sans grande importance sur le fond, cette critique doit être soulignée car l'utilisateur trouve matière à irritation. On a toujours l'impression que ceux qui préparent les fichiers ignorent les difficultés rencontrées par ceux qui exploitent en solo les données[11]. Par ailleurs, si le fichier régional contient une centaine de variables, la plupart d'entre elles sont des recompositions… Proposer des regroupements est souhaitable, mais dans bien des cas, il semble que l'on n'ait pas beaucoup réfléchi à la variété des applications.

6- Ce qu'il reste à faire[12]

Ce que l'on attend en premier lieu de l'Insee, parce que lui seul peut le faire, c'est de produire des fichiers de qualité et les diffuser pour le meilleur usage des institutions et des citoyens. Il y aurait encore beaucoup d'améliorations à apporter aux produits mis en ligne sans qu'il soit pour autant nécessaire d'engager de gros investissements. On propose plusieurs pistes.

La documentation des fichiers : proposer des modes d'emploi, adapter les notes méthodologiques afin qu'elles disent de façon concise ce qui est nécessaire et qu'elles reportent en annexe ce qui est secondaire. Une même activité routinière amène parfois l'Insee à maintenir des codifications hors d'usage[13].

Les erreurs.

Voici deux exemples

1) La sous-estimation de la population au recensement de 1999 : il aura fallu attendre plusieurs années avant que l'Insee ne consente à reconnaître, sous la forme d'un communiqué, la différence importante observée entre les résultats attendus et les chiffres diffusés. A l'inverse, dès la publication des premiers résultats du recensement allemand de 2011, Destatis a souligné les écarts constatés et a proposé des explications nécessaires à la compréhension des chiffres.

---

[11] Une critique : malheur à celui qui imprime la documentation et qui est soucieux d'économiser et le papier et l'encre de son imprimante (dont on connaît le coût exorbitant) : des pages entières composées d'une répétition de listes de toutes les années, de 1884 à 2014 ou les âges (de 000 à 123). On aurait pu ne faire figurer que les années ou les âges extrêmes (variables AEMM, AGED, AGEREV, AN, ANEM)! Ce simple détail confirme bien, s'il le fallait, le peu d'intérêt porté aux problèmes concrets rencontrés par l''utilisateur.

[12] Les aspects évoqués dans ce chapitre ont fait l'objet d'un courrier adressé à l'Autorité de la statistique publique avec copie au Président du CNIS et au Directeur général de l'Insee. Aucun accusé de réception. Ce courrier a-t-il seulement été lu ?
http://www.apr-strasbourg.org/membre/2054/aubry-bernard/document/683-rflexion-sur-l-accs-aux-statistiques-publiques.html

*[13] Il en va ainsi de la date d'entrée en France collectée dans les recensements. On nous propose 9 modalités. Modalité 1, période 1884-1899: **aucune** installation, modalité 2, entre 1990 et 1917 : **532** …. Modalité 9, depuis 1998 : **1970000** installations (chiffres pour l'EAR 2012). On se demande à qui peut bien servir pareil découpage si on ne détaille pas davantage les modalités utiles à l'analyse !*

2) L'emploi concernant la commune de Mulhouse (encadré).

Extrait des Chiffres-clés pour la commune de Mulhouse

Plusieurs tableaux et graphiques sont proposés sous la référence « Activité et emploi ». Des tableaux EMP.T5 et EMP.T8 on retient quelques informations contradictoires pour les recensements 1999, 2007 et 2012 :*

EMP.T5 Emploi et activité	*1999*	*2007*	*2012*
*Ensemble des **emplois** (actifs comptés au lieu de travail*	*61820*	*62780*	*55676*
***Actifs** occupés (comptés au lieu de résidence)*	*42034*	*39354*	*35573*
Mention : exploitation principale au lieu de résidence et au lieu de travail			

EMP.T8 Emploi selon le secteur d'activité	*1999*	*2007*	*2012*
*Ensemble des **emplois** (actifs comptés au lieu de travail*	*53567*	*61035*	*55676*
*Dont **emplois dans l'industrie***	***6148***	***10339***	***6114***
Mention : exploitation complémentaire au lieu de travail			

En comparant ces données entre elles, le lecteur remarque deux incohérences. D'une part la différence sur les chiffres sur l'emploi (entre exploitation principale EMP.T5 et complémentaire EMP.T8, considérable en 1999, puisque l'écart était de plus de 8000 postes de travail ; plus faible en 2007, il est nul en 2012). D'autre part (EMP.T8), ***le nombre d'emplois dans l'industrie qui aurait fait un bond de 4200 de 1999 à 2007 (+68%) ; il serait retombé d'autant de 2007 à 2012 (-4200, soit – 41%)****.*

*[*http://www.insee.fr/fr/themes/donnees-locales.asp?typgeo=COM&typesearch=territoire&search=Mulhouse+%2868224%29&ref_id=emp2012*](http://www.insee.fr/fr/themes/donnees-locales.asp?typgeo=COM&typesearch=territoire&search=Mulhouse+%2868224%29&ref_id=emp2012)

Les anomalies dans les statistiques diffusées sont courantes, mais celles qui affectent la commune de Mulhouse sont particulièrement choquantes. Il peut s'agir d'erreurs dans la collecte ou dans le traitement des données. En l'occurrence la collecte est correcte, c'est le traitement qui est en cause. Les chiffres subissent en fait deux traitements parallèles et l'on ne s'est pas préoccupé de leur cohérence. Cela signifie qu'avant de diffuser les résultats, l'Insee n'a pas enclenché un système d'alerte qui aurait permis d'attirer l'attention avant la diffusion des résultats. Certes figurent à côté des tableaux des mentions. L'une a trait aux militaires du contingent (diffusion EAR 2007)[14] et n'a pas beaucoup d'intérêt. L'autre concerne les traitements subis par les chiffres. De fait pour le tableau T5, on lit « ex*ploitation principale au lieu de résidence et lieu de travail*» et pour le tableau T8 : « *exploitation complémentaire au lieu de travail* ». Mais il s'agit là d'un langage inaccessible à tous ceux – et ils sont nombreux - qui ne sont pas familiers des

[14] *Rappelons que l'appel sous les drapeaux a été suspendu par la loi du 28 octobre 1997 pour tous les Français nés après le 31 décembre 1978.*

arcanes des recensements[15]. L'utilisateur n'a que faire des subtilités des statisticiens. Les chiffres diffusés doivent être crédibles, immédiatement compris des personnes auxquels ils sont destinés. Puisque que le problème a été identifié et corrigé (dès lors que les chiffres sont maintenant cohérents pour l'EAR 2012[16]), pourquoi l'Insee ne recalcule-t-il pas des séries historiques pour donner du sens aux variations?

Bien sûr, et c'est un argument supplémentaire en faveur de *l'open data*, les utilisateurs qui ont téléchargé les fichiers détail peuvent, s'ils en ont la patience, « bricoler » les chiffres de façon à donner une image plus proche de la réalité que ne fait l'Insee avec ses chiffres-clés. Mais n'est-ce pas plutôt aux spécialistes rompus aux méthodes de redressements enseignées à l'école de la statistique de faire le travail de mise en cohérence des chiffres proposés au public?

Quoi qu'il en soit, l'institution doit impérativement corriger ses fichiers des verrues qui sèment le doute quand elles sont découvertes et qui conduisent à des risques d'interprétations aberrantes quand elles ne le sont pas. Ces erreurs ne peuvent que contribuer à disqualifier la statistique publique aux yeux de ses contempteurs. Elles sont d'autant plus impardonnables que, dans le cas de Mulhouse, aucune mise en garde spécifique n'est susceptible d'attirer l'attention de l'utilisateur potentiel.

Le contenu des données. On doit faire un effort pour aider l'utilisateur à réfléchir sur les données, leur valeur, la signification des résultats. Il faut davantage insister sur la précision des résultats, mettre en garde sur les pièges de l'interprétation. Pourquoi ne pas mettre en ligne un manuel « grand public » - coproduit par les utilisateurs, mais coordonné par un groupe de pilotage - qui replacerait le recensement dans le système statistique du pays (finalité, méthode, outil d'analyse, diffusion, comparaisons internationales…) ?

Hiérachiser les données pour pouvoir facilement étudier les liens entre les personnes d'une même famille (ou d'un même ménage).

D'autres orientations:

Redresser les données plus anciennes (1962-1999) de façon à donner du sens aux séries longues. Cela suppose de faire des hypothèses appropriées et d'affecter à chaque observation devant être corrigée une pondération adéquate. Ce travail conduirait à harmoniser les chiffres issus des exploitations par sondage (au quart) avec les chiffres de population issus du dénombrement. Il faudrait aussi corriger sans doute les chiffres du dénombrement eux-mêmes, quand ils sont manifestement erronés. On est là devant des investissements théoriques qui peuvent être difficiles, mais qui ne pourraient que valoriser la formation initiale des cadres issus des écoles d'application de la statistique,

---

[15] *La direction régionale de l'Insee-Alsace a pourtant, à de nombreuses reprises, tenté d'attirer l'attention de sa hiérarchie parisienne à ce sujet. Ce problème des divergences entre les données issues de deux traitements (l'exploitation principale et l'exploitation complémentaire) conduisait déjà dans le passé à des divergences importantes, mais qu'un usage raisonné des tableaux permettait d'éviter, ce que ne permet guère la diffusion des EAR. Pour en savoir plus sur l'importance des écarts entre les traitements (ils ne concernent pas seulement Mulhouse) : http://www.apr-strasbourg.org/membre/2054/aubry-bernard/document/685-recensement-des-rsultats-divergents.html*

[16] *En revanche les variations de 2007 à 2012 sont totalement aberrantes.*

trop souvent affectés à des taches de gestion ou d'administration pour lesquelles ils n'ont pas été préparés.

Etendre l'open data à d'autres enquêtes - On peut se réjouir de trouver en ligne des fichiers détail relatifs à quelques autres enquêtes auprès des ménages, enquêtes Emploi, enquêtes permanentes sur les conditions de vie (EPCV). D'autres enquêtes pourraient également faire l'objet d'une mise en ligne à l'instar des recensements. Leur exploitation par un public non spécialisé serait facilitée par le fait que les ménages sont décrits de façon identique et que les variables structurelles de base relèvent de nomenclatures identiques (sexe, âge, localisation, pondération…). Les pondérations ne seraient pas, comme au recensement, voisines de 1 ou de 4. Elles seraient beaucoup plus grandes, de l'ordre de 800 pour les fichiers des enquêtes sur le logement. On aurait ainsi des informations complémentaires comme par exemple sur la question du mal-logement (non repérable au recensement) ou sur celle de la réhabilitation des logements. Bien entendu les résultats n'auraient de sens que pour des ensembles géographiques suffisamment grands. L'avantage de tels fichiers harmonisés serait d'éviter des travaux fastidieux préalables de prise en main par ceux qui cherchent à dépasser le champ des variables des recensements, forcément limité ou qui cherchent à comparer les sources entre elles.[17]

7 - A propos du contrôle de la diffusion

Quel que soit le mode de diffusion retenu - fichiers de tableaux ou fichiers détail -, les statisticiens se heurtent au problème du risque de diffusion de statistiques confidentielles[18]. Des pistes sont proposées par les chercheurs pour pallier cet inconvénient, mais il ne semble pas qu'il y ait encore d'application systématique de ces méthodes au recensement français. On peut par exemple brouiller aléatoirement, mais faiblement, les résultats quand ils portent sur de petits nombres. C'est ce que fait le site Nomis, le site de diffusion britannique. Il n'y a pas de solution miracle, il faut trouver des compromis entre la satisfaction des besoins des utilisateurs et les risques, qu'ils soient de confidentialité ou tout simplement d'interprétation erronée.

---

[17] *L'auteur a réalisé un travail de ce type qui consistait à comparer les flux migratoires inter-régionaux : http://www.apr-strasbourg.org/membre/2054/aubry-bernard/document/682-recensement-enqutes-emploi.html*

[18] *On renvoie aux articles publiés d'Alfred Dittgen (professeur émérite – Paris 1). Voir l'ouvrage Démographie locale. Relations : population, logement migration, Néothèque 2012 et plusieurs articles parus dans la Lettre Blanche (association Pénombre http://www.penombre.org/Encore-une-statistique-singuliere). A. Dittgen aime à ironiser sur certaines faiblesses et incohérences dans la diffusion des données. L'auteur évoque le cas de la commune de Cumières-le-Mort-Homme : en 1990, un habitant (un seul !), sexe masculin, tranche d'âge 40-59 ans, propriétaire d'un logement de 5 pièces, etc. Cette diffusion se faisant, évidemment, en totale contradiction avec les règles générales en vigueur à l'époque en matière de préservation d'anonymat.*

La responsabilité des chiffres produits.

L'Insee décline toute responsabilité dans l'usage fait des résultats établis à partir de fichiers détail, ce que l'on comprend aisément. Si des chiffres aberrants ou faux sont diffusés sur le net par des groupes ou des individus, ils ne feront que s'ajouter à beaucoup d'autres sottises mises en ligne et avec lesquelles la société doit composer.

8 - Conclusion

Le site de l'Insee propose gratuitement pour les enquêtes EAR (depuis 2006) des fichiers détail. Mais, à l'inverse, il bloque la diffusion des mêmes informations concernant les recensements antérieurs aux enquêtes annuelles (1962-1999), ce qui ne permet pas d'engager, hors du regard de l'administration, des réflexions sur la dynamique des territoires puisque toute profondeur historique est impossible à mettre en évidence.

Avec l'offre des fichiers détail, la France s'est engagée partiellement dans *l'open data*, mais comme cette initiative ne profite qu'à quelques utilisateurs privilégiés, l'application du concept n'est que théorique. Une solution innovante serait de proposer un système de menus déroulants permettant de croiser beaucoup plus de variables que ne le font les autres instituts de statistique, tout en respectant les règles de confidentialité qui s'imposent. En ce cas de croisements multiples, la mise à disposition des résultats se ferait parfois en différé, ce qui n'est pas un gros inconvénient.

S'agissant des recensements, ce système de diffusion devrait couvrir une période débutant en 1962, puisque notre pays, contrairement à bien d'autres, a le privilège d'avoir pu effectuer des collectes de façon régulière sur la base d'un questionnaire qui n'a changé qu'à la marge[19]. Quant aux autres enquêtes (sur l'emploi, le logement, la formation et la qualification professionnelle, la famille[20]…) dont la structure des fichiers s'apparente à celle des recensements, elles pourraient, *mutatis mutandis*, être proposées également à la diffusion sous une forme analogue (du moins pour certaines variables non retenues dans les questionnaires du recensement).

L'accès aux données du système statistique français en *open data* n'est donc qu'entr'ouvert.

[19] *De nombreux pays proposent des séries temporelles longues. Le recensement de 1962 n'ayant été exploité qu'au 20ème (un logement sur 20), le choix des variables disponibles serait plus limité que pour les recensements ultérieurs.*

[20] *associée à la collecte 2011 du recensement.*

Annexe I - Méthode : la collecte et les traitements

A - L'organisation de la collecte des EAR

Dans le passé, l'ensemble de la population était observée dans sa situation un jour donné (début du mois de mars). Pour différentes raisons, l'Insee s'est engagé depuis 2004 dans un nouveau système de collecte dont les grandes lignes sont les suivantes :

- Les communes de moins de 10 000 habitants (la moitié de la population du pays) continuent d'être recensées exhaustivement, comme par le passé, mais une fois tous les 5 ans au lieu de tous les 8 ou 9 ans.
- Les autres communes font désormais l'objet d'une enquête annuelle auprès d'un échantillon de 8 % de la population, dispersé aléatoirement sur l'ensemble de leur territoire.

Au bout de 5 ans, **70% de la population** reçoit la visite d'un agent recenseur (100% pour les communes de moins de 10000 est 40% de la population des plus grandes).

B – Les traitements

Dans le passé, le traitement des bulletins des recensements a beaucoup varié en raison des mutations technologiques. Le recensement de 1962 fut le premier à bénéficier d'une exploitation informatique. Plus tard, avec l'introduction de la lecture optique, les temps de traitement se sont accélérés. D'une façon générale, on peut distinguer trois types de traitement des bulletins collectés. En premier lieu un simple décompte des logements et de leurs habitants (le dénombrement) qui conduisait à fixer pour chaque commune les chiffres de *population légale.* Puis deux autres chaînes d'opérations : un traitement exhaustif et un traitement par sondage (tirage au sort d'un logement sur 4, voire de un sur 5 en 1975). Le traitement *exhaustif* (appelé aujourd'hui *exploitation principale*) porte sur les questions à codification simple, le traitement par *sondage* (exploitation *complémentaire*) porte sur les variables à codification complexe, notamment la catégorie socioprofessionnelle et l'activité économique de l'entreprise (voir plus loin). Pour des raisons d'ordre purement statistique, les chiffres diffusés divergent selon le mode de traitement, le sondage créant une incertitude dont l'amplitude relève des règles du calcul des probabilités. Ces divergences peuvent être importantes et troubler les utilisateurs. Mais à ces écarts statistiques se surajoutent parfois des erreurs dues à une insuffisante prise en compte de la localisation des établissements dans lesquels travaillent les personnes recensées et qui détermine **l'emploi** (compté au lieu de travail) et la **population active** (comptée au lieu de résidence). Ainsi l'établissement Peugeot, qui regroupe plusieurs milliers de salariés, est localisé tantôt à Mulhouse (exploitation principale), tantôt à Sausheim (exploitation complémentaire). Ce qui aboutit à des résultats tout à fait incohérents. Les statisticiens nationaux, soucieux de fournir des résultats sur des aires géographiques de grande taille, n'ont jamais porté beaucoup d'intérêt à des divergences qui gênent surtout ceux qui travaillent sur des aires géographiques restreintes. Plus la taille du territoire est grande et plus l'incertitude aléatoire est faible et plus grande est la probabilité pour que les deux communes à risque soient incluses dans la même entité (les chiffres de l'agglomération ou ceux de l'aire urbaine de Mulhouse ne supportent les conséquences de l'erreur de localisation de

l'établissement).[21] Pour en savoir plus on pourra se référer aux nombreux documents de méthode mis en ligne sur le site de l'Insee, notamment http://www.insee.fr/fr/ppp/bases-de-donnees/recensement/resultats/doc/pdf/fiche-ponderation.pdf pour la pondération et http://www.insee.fr/fr/ppp/bases-de-donnees/recensement/resultats/doc/pdf/fiche-exploitation.pdf *pour les méthodes de traitement. Voici ci-après un extrait :*

> *Les questionnaires renseignés par les personnes recensées et collectés par les agents recenseurs sont saisis par un dispositif de lecture optique et font l'objet d'un vidéo-codage. L'Insee procède ensuite à une série de traitements statistiques : contrôle de la cohérence des informations collectées, imputation des valeurs manquantes, codification des variables de diffusion, calcul des pondérations en vue d'obtenir un fichier de données individuelles, anonymes et apurées, prêt à être exploité. Comme lors des recensements généraux traditionnels, l'exploitation statistique des questionnaires collectés est réalisée en deux temps : une exploitation dite « principale », puis une exploitation qualifiée de « complémentaire». Depuis la rénovation du recensement, le délai qui sépare les deux exploitations est très court et les résultats qui en sont issus sont diffusés simultanément.*
>
> *1. L'exploitation principale*
>
> *L'exploitation « principale » porte sur l'ensemble des questionnaires collectés. Elle est donc exhaustive pour les communes de moins de 10 000 habitants et porte sur environ 40 % des logements dans les communes de 10 000 habitants ou plus. Elle permet de produire un « fichier détail » contenant l'ensemble des logements et des individus recensés. Ces observations sont assorties d'un poids spécifique de l'exploitation principale.*
>
> *L'exploitation principale traite toutes les informations pouvant être codifiées aisément après la saisie des questionnaires. Les résultats statistiques issus de cette exploitation couvrent la plupart des critères d'étude permis par les questionnaires du recensement (bulletin individuel et feuille de logement). Ils portent sur de nombreuses variables et peuvent être classés en 11 thèmes :*

- *Population (âge, sexe, état matrimonial, actifs ou inactifs...) ;*
- *Activité des résidents (taux d'activité par sexe et âge, chômage, statut professionnel...) ;*
- *Emploi au lieu de travail (salariés ou non, temps partiel ou complet, conditions d'emploi...) ;*
- *Déplacements domicile - lieu de travail ;*
- *Formation (scolarisation et lieu d'études, diplômes...) ;*
- *Migrations (lieu de résidence cinq ans ou un an plus tôt, lieu de naissance) ;*
- *Nationalité (Français de naissance ou par acquisition, nationalité selon le sexe et l'âge...) ;*
- *Immigration (pays de naissance, année d'arrivée en France...) ;*
- *Ménages (taille du ménage, sexe et âge de la personne de référence...) ;*
- *Logements (type, époque d'achèvement...) ;*
- *Résidences principales (statut d'occupation, date d'emménagement, nombre de pièces, surface, sanitaires, mode de chauffage, nombre de voitures...).*

> *2. L'exploitation complémentaire*
>
> *La seconde phase de l'exploitation statistique, dite « complémentaire », est destinée à produire les variables dont l'élaboration est complexe. Il s'agit de celles qui décrivent :*

[21] *Du reste, il s'agit de l'une des motivations qui ont conduit à la création de Saphir (fichier historique des recensements : http://www.apr-strasbourg.org/membre/2054/aubry-bernard/document/681-saphir-prsentation-partie-1.html Voir, en particulier, pages 27 et 28 « Abondance de données nuit parfois ».*

• la structure familiale des ménages : détermination précise de la personne de référence du ménage, identification, le cas échéant, de familles au sein du ménage et composition de ces familles ;
• les secteurs d'activité dans lesquels les emplois sont exercés ;
• la profession et la catégorie socioprofessionnelle des personnes, notamment de celles qui exercent un emploi

Annexe II - La diffusion des chiffres du recensement : une comparaison des sites britannique et français

La mise en regard du site de l'Insee et de celui de l'ONS *(Office For National Statistics) anglais (Nomis)*, donne une bonne idée du manque d'audace qui caractérise la diffusion des résultats des recensements français, alors même que la technicité qui les a engendrés est très sophistiquée

Une première remarque, en forme de parenthèse. Le domaine choisi pour le test, l'immigration, illustre les différences de pratique de part et d'autre de la Manche. L'ONS intègre depuis 1991 une question sur l'appartenance ethnique selon une nomenclature qui n'a pas cours en France (le « référentiel ethno-racial »). Depuis 2001, une question porte également sur la religion (question facultative, 7,2% de non réponses, voir tableau T.4). Côté français, l'Insee ne s'est guère aventuré, dans ses recensements, au-delà du recueil du pays de naissance et de la nationalité. Pourtant, depuis 2007, la CNIL est favorable à ce que le recueil du pays et de la nationalité de naissance des parents, déjà courant dans les grandes enquêtes de l'Insee (enquête Emploi par exemple), figure aussi dans les enquêtes annuelles de recensement. Ces questions étant bien acceptées par la grande majorité des personnes interrogées, l'Insee, ou plutôt le Conseil national de l'information statistique (CNIS) pourrait œuvrer à l'intégration de ce complément dans les questionnaires des EARs. En effet, il ne fait pas de doute que l'information collectée serait d'une grande utilité pour mesurer le niveau d'intégration de la population d'origine étrangère sur deux générations, qui varie selon l'origine, le type d'habitat (rural-urbain)[1] , etc.

Timide dans l'interprétation des textes en matière de collecte, la statistique française l'est aussi en matière de diffusion. Pourtant, compte tenu de la place prise dans le débat social par l'immigration, une connaissance moins parcellaire du phénomène éviterait sans doute la propagation d'idées fausses[*1].*
Notons qu'en raison du statut spécial en vigueur en Alsace et en Moselle, quelques questions ont été posées dans le passé sur la religion et la langue (notamment aux recensements de 1936 et 1962)
Mais on ne saurait focaliser l'attention sur la seule question de l'immigration. En effet le déficit de l'offre statistique est peut-être encore plus flagrant quand il s'agit des données sociodémographiques en un temps où le pays s'est engagé dans de vastes recompositions des territoires (modifications cantonales et régionales). Il semble que l'on ait fait bien peu d'efforts pour s'adapter et intéresser les citoyens à la découverte de leur nouvelle région par un accès aisé aux indicateurs basiques. Un exemple : il n'est pas possible d'obtenir, facilement, un panorama des 31 zones d'emploi de la région Alsace-Lorraine-Champagne-Ardenne. C'est, bien entendu, la même chose si l'on veut comparer des villes, des communautés de communes, des territoires quelconques entre eux.
Un fichier détail fournit quasi-immédiatement l'information, pour autant que l'on ait formulé clairement les questions.

*[*1] La palette des questions posées est sensiblement différente d'un pays à l'autre. On trouve sur le site de l'IPUMS (université du Minnesota) des échantillons de données individuelles concernant les recensements de 82 pays. Le site propose des tableaux synoptiques des variables disponibles par thème (à titre d'exemple, pour la religion et l'ethnie : https://international.ipums.org/international-action/variables/group?id=ethnic*

On observe ainsi que :

- Le site de l'Insee donne accès à des données basiques d'un côté et à des fichiers détail de l'autre. Entre les deux, l'offre est peu satisfaisante.
- Nomis ne propose pas de fichiers détail, mais il offre un choix considérable de tableaux auxquels sont toujours associés un accès et une documentation adéquate bénéficiant d'une convivialité appréciable puisque, d'un simple clic, on obtient des informations précises sur le contenu du tableau. Les données portant sur des effectifs réduits (moins de 100 personnes ou 50 ménages) sont brouillées de façon à ce que la confidentialité soit respectée. Mais les possibilités de croisement des variables sont limitées et l'on ne s'approche donc pas de ce que peut produire un fichier-détail.

Du côté de l'Insee, c'est l'inverse. On télécharge aisément les fichiers détail mais, en dehors de cette opportunité inaccessible au plus grand nombre, on doit se contenter de tableaux dont le contenu est figé, qu'il s'agisse des chiffres-clés – ce que l'on comprend bien - et des fichiers de tableaux communaux qui sont à la fois pauvres en variables et peu avenants. Autant le système de collecte élaboré par l'Insee pour recenser en continu la population est l'alliance de la science statistique et de la créativité (le nouveau recensement est une construction complexe adaptée au contexte juridico-administratif du pays ; en effet contrairement à d'autres pays, il n'existe pas de répertoire de population), autant le système de diffusion semble se situer hors du temps, hors des préoccupations de la société. Ce n'est pourtant pas la matière grise qui fait défaut. Faut-il y voir un déficit de curiosité intellectuelle, l'enlisement dans des pratiques routinières peu favorables à l'émergence des idées ? Assurément, la notion de *« statistique participative »*[22] est une notion qui, de toute évidence, est aujourd'hui étrangère à l'institution[23]. Pourquoi l'Insee n'a-t-il pas cherché à imiter, pour les dépasser, les meilleures expériences de ses homologues ? Pourquoi ne pas inciter les agents à se brancher, chacun dans son domaine, sur les sites statistiques de la Grande-Bretagne, de la Suisse, des Pays-Bas, de la Norvège et de bien d'autres encore, de les inviter à discuter sur les points forts et les points faibles de chacun d'entre eux. Au manque de curiosité, ne faut-il pas aussi ajouter une propension récurrente au jacobinisme qui voudrait tenir pour peu ce qui ne vient pas de la capitale et de ses têtes pensantes ? Ainsi comment expliquer autrement le désintérêt des services centraux de l'Insee à l'égard de la DR d'Alsace qui avait, dès les années 90, créé un fichier historique des recensements réalisés depuis 1962 ?

A – Le site de l'Insee

[22] *L'idée d'une « statistique participative » est un peu analogue à celle d'une « science participative », un concept que l'Université de Strasbourg s'efforce de promouvoir.*

[23] *Cela n'a pas toujours été le cas puisque, dans le cadre de la « Nouvelle société » promue dans les années 70 par Jacques Chaban-Delmas, l'Insee avait créé des Observatoires économiques régionaux qui avaient la mission de diffuser l'information statistique dans le sens d'un rapprochement entre les producteurs et les utilisateurs de données chiffrées travaillant sur le terrain.*

On distingue, en simplifiant, trois types de produits (T.1):

T.1 – les liens permettant d'accéder aux principaux résultats

1.**Résultats pour un territoire (chiffres-clés)** (fichiers-tableaux)	http://www.insee.fr/fr/bases-de-donnees/default.asp?page=recensement/resultats/2012/rp2012.htm
2.B**ase de données** (fichiers tableaux)	http://www.insee.fr/fr/bases-de-donnees/default.asp?page=recensement/resultats/2012/donnees-detaillees-recensement-2012.htm
3.**Fichiers détail** (fichiers individuels)	http://www.insee.fr/fr/themes/detail.asp?reg_id=99&ref_id=fd-rp2012&page=fichiers_detail/RP2012/telechargement.htm

1- Résultats pour un territoire - L'utilisateur a le choix entre 11 types de territoires et 9 thèmes, sachant que pour un même thème le nombre de tableaux proposés est variable (voir plus loin T.3a)

2-Base de données - On choisit le thème d'après un tableau (voir plus loin T.3b). On télécharge (.txt ou .xls) un tableau de 36675 lignes (autant que de communes, yc DOM, sauf Mayotte).

L'image présentée (T.3c) correspond au tableau NAT1, l'un des 6 tableaux proposés sur le thème *Nationalité et immigration*. En bas du tableau, l'un des onglets précise les définitions.

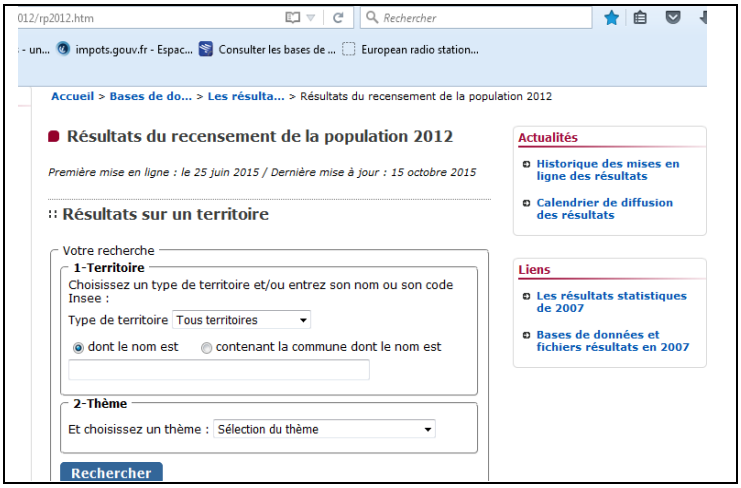
3 -Fichiers détail- Le tableau (T.3d) montre, à titre d'exemple, les choix fournis : téléchargement complet (France entière) ou en 5 éléments, formés d'un ensemble de régions.

T.2 – Site de l'Insee : avantages et inconvénient de chacune des catégories (voir T.1)

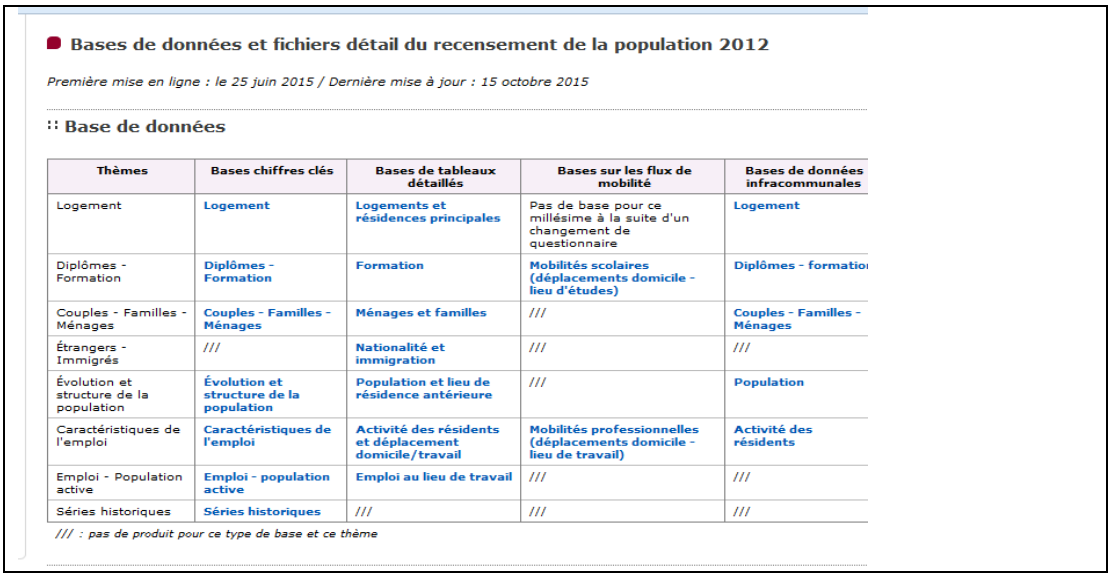
	Avantages	**Inconvénients**
1. (tableaux)	Simplicité Adapté à la consultation Accès à EAR 2007 et 2012	Pas possible de comparer immédiatement des territoires entre eux Peu de résultats historiques Pas d'accès immédiat à la documentation Les anomalies ne sont pas signalées
2. (tableaux)	Données communales (parfois à l'iris)	Lourdeur excessive Regroupements difficiles
3. (détail)	Croisements des variables à volonté, notamment au niveau régional	Fichiers disjoints Nécessité d'un logiciel ad hoc

T.3 – Images correspondant aux liens

T.3a



T.3b



Bases de données et fichiers détail du recensement de la population 2012

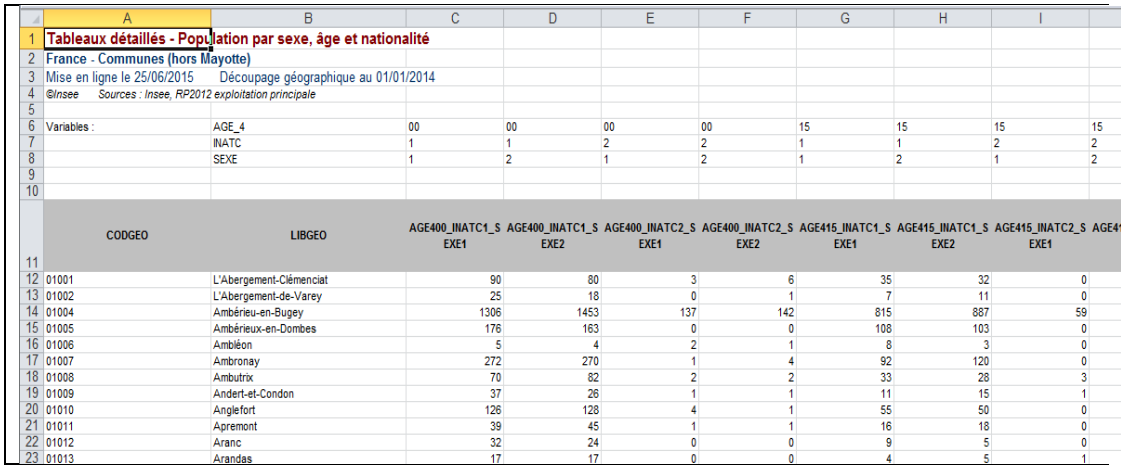
Première mise en ligne : le 25 juin 2015 / Dernière mise à jour : 15 octobre 2015

Base de données

Thèmes	Bases chiffres clés	Bases de tableaux détaillés	Bases sur les flux de mobilité	Bases de données infracommunales
Logement	Logement	Logements et résidences principales	Pas de base pour ce millésime à la suite d'un changement de questionnaire	Logement
Diplômes - Formation	Diplômes - Formation	Formation	Mobilités scolaires (déplacements domicile - lieu d'études)	Diplômes - formatio
Couples - Familles - Ménages	Couples - Familles - Ménages	Ménages et familles	///	Couples - Familles - Ménages
Étrangers - Immigrés	///	Nationalité et immigration	///	///
Évolution et structure de la population	Évolution et structure de la population	Population et lieu de résidence antérieure	///	Population
Caractéristiques de l'emploi	Caractéristiques de l'emploi	Activité des résidents et déplacement domicile/travail	Mobilités professionnelles (déplacements domicile - lieu de travail)	Activité des résidents
Emploi - Population active	Emploi - population active	Emploi au lieu de travail	///	///
Séries historiques	Séries historiques	///	///	///

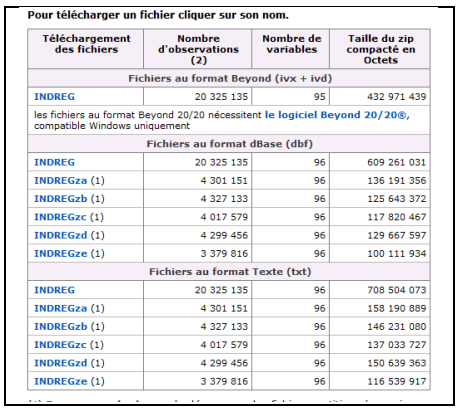
/// : pas de produit pour ce type de base et ce thème

T.3c



	A	B	C	D	E	F	G	H	I
1	Tableaux détaillés - Population par sexe, âge et nationalité								
2	France - Communes (hors Mayotte)								
3	Mise en ligne le 25/06/2015	Découpage géographique au 01/01/2014							
4	©Insee	Sources : Insee, RP2012 exploitation principale							
5									
6	Variables :	AGE_4	00	00	00	00	15	15	15
7		INATC	1	1	2	2	1	1	2
8		SEXE	1	2	1	2	1	2	1
9									
10									
11	CODGEO	LIBGEO	AGE400_INATC1_S EXE1	AGE400_INATC1_S EXE2	AGE400_INATC2_S EXE1	AGE400_INATC2_S EXE2	AGE415_INATC1_S EXE1	AGE415_INATC1_S EXE2	AGE415_INATC2_S EXE1
12	01001	L'Abergement-Clémenciat	90	80	3	6	35	32	0
13	01002	L'Abergement-de-Varey	25	18	0	1	7	11	0
14	01004	Ambérieu-en-Bugey	1306	1453	137	142	815	887	59
15	01005	Ambérieux-en-Dombes	176	163	0	0	108	103	0
16	01006	Ambléon	5	4	2	1	8	3	0
17	01007	Ambronay	272	270	1	4	92	120	0
18	01008	Ambutrix	70	82	2	2	33	28	3
19	01009	Andert-et-Condon	37	26	1	1	11	15	1
20	01010	Anglefort	126	128	4	1	55	50	0
21	01011	Apremont	39	45	1	1	16	18	0
22	01012	Aranc	32	24	0	0	9	5	0
23	01013	Arandas	17	17	0	0	4	5	1

T.3d



Pour télécharger un fichier cliquer sur son nom.

Téléchargement des fichiers	Nombre d'observations (2)	Nombre de variables	Taille du zip compacté en Octets
Fichiers au format Beyond (ivx + ivd)			
INDREG	20 325 135	95	432 971 439
les fichiers au format Beyond 20/20 nécessitent le logiciel Beyond 20/20®, compatible Windows uniquement			
Fichiers au format dBase (dbf)			
INDREG	20 325 135	96	609 261 031
INDREGza (1)	4 301 151	96	136 191 356
INDREGzb (1)	4 327 133	96	125 643 372
INDREGzc (1)	4 017 579	96	117 820 467
INDREGzd (1)	4 299 456	96	129 667 597
INDREGze (1)	3 379 816	96	100 111 934
Fichiers au format Texte (txt)			
INDREG	20 325 135	96	708 504 073
INDREGza (1)	4 301 151	96	158 190 889
INDREGzb (1)	4 327 133	96	146 231 080
INDREGzc (1)	4 017 579	96	137 033 727
INDREGzd (1)	4 299 456	96	150 639 363
INDREGze (1)	3 379 816	96	116 539 917

B – Le site britannique NOMIS (Office For National Statistics) :

https://www.nomisweb.co.uk/census/2011

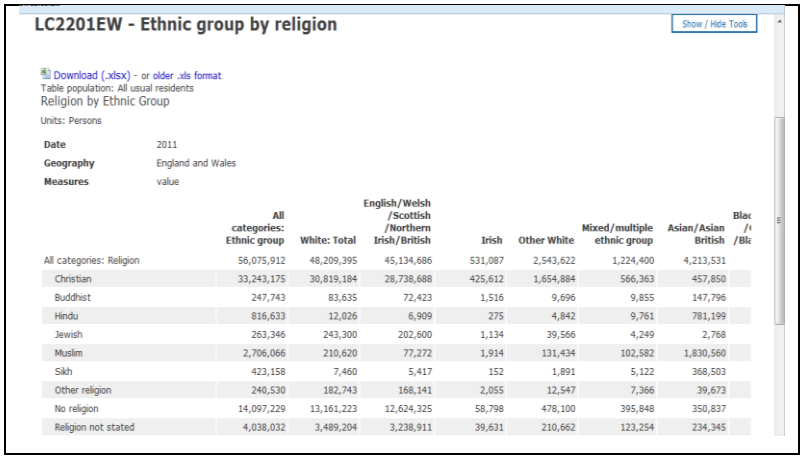
La page d'accueil annonce la mise à disposition de plusieurs centaines de tableaux couvrant une large variété de thèmes :

***Table Finder** liste les principaux thèmes (disponibles à tous les niveaux géographiques)
***Key Statistiss (KS)** propose des chiffres-clés pour chacun des variables (sans croisement de variable), y compris au lieu de travail
***Quick Statistics (QS)** pour un accès à 75 tableaux de base
***Local Characteristics (LC)** pour croiser des variables entre elles
***Detailled Characteristics (DC) ,** idem, mais avec davantage de détails
***Out Of Term-time Population (OT)** pour les navettes (étudiants…)
***Origin Destination Data** pour les flux migratoires

Exemple : *Local Characteristics* : groupes ethniques par religion (LC2201)
Le lien suivant nous fournit le tableau suivant :

https://www.nomisweb.co.uk/census/2011/LC2201EW/view/2092957703?rows=c_relpuk11&cols=c_ethpuk11 :

T.4a Groupe ethnique et religion



LC2201EW - Ethnic group by religion

Show / Hide Tools

Download (.xlsx) - or older .xls format
Table population: All usual residents
Religion by Ethnic Group
Units: Persons

Date	2011
Geography	England and Wales
Measures	value

	All categories: Ethnic group	White: Total	English/Welsh /Scottish /Northern Irish/British	Irish	Other White	Mixed/multiple ethnic group	Asian/Asian British	Blac /(/Bla
All categories: Religion	56,075,912	48,209,395	45,134,686	531,087	2,543,622	1,224,400	4,213,531	
Christian	33,243,175	30,819,184	28,738,688	425,612	1,654,884	566,363	457,850	
Buddhist	247,743	83,635	72,423	1,516	9,696	9,855	147,796	
Hindu	816,633	12,026	6,909	275	4,842	9,761	781,199	
Jewish	263,346	243,300	202,600	1,134	39,566	4,249	2,768	
Muslim	2,706,066	210,620	77,272	1,914	131,434	102,582	1,830,560	
Sikh	423,158	7,460	5,417	152	1,891	5,122	368,503	
Other religion	240,530	182,743	168,141	2,055	12,547	7,366	39,673	
No religion	14,097,229	13,161,223	12,624,325	58,798	478,100	395,848	350,837	
Religion not stated	4,038,032	3,489,204	3,238,911	39,631	210,662	123,254	234,345	

Après téléchargement, on obtient un fichier Excel comprenant trois onglets :

1- *Contents* (brève description) – 2 : *Metadata* (information sur les variables) – 3 sheet (tableau).

T.4b Les mêmes données, présentées en milliers.

Ethnic Group	All categories: Ethnic group	White: Total	English/ Welsh/Scottish/Northern Irish/British	Irish	Other White	Mixed/multiple ethnic group	Asian/ Asian British	Black/African/Caribbean/Black British	Other ethnic group
Religion	1	2	3	4	5	6	7	8	9
All categories: Religion	56076	48209	45135	531	2544	1224	4214	1865	564
Christian	33243	30819	28739	426	1655	566	458	1288	111
Buddhist	248	84	72	2	10	10	148	3	4
Hindu	817	12	7	0	5	10	781	5	8
Jewish	263	243	203	1	40	4	3	2	11
Muslim	2706	211	77	2	131	103	1831	272	290
Sikh	423	7	5	0	2	5	369	1	41
Other religion	241	183	168	2	13	7	40	7	4
No religion	14097	13161	12624	59	478	396	351	137	52
Religion not stated	4038	3489	3239	40	211	123	234	149	43

1: ensemble – 2 : Blancs (total)- 3 : dont Anglais, Gallois, Ecossais, Irlandais du Nord, Britanniques
4: Irlandais- 5: autres Blancs – 6: Mixtes/groupe ethnique multiple- 7: Asiatiques/Asiat. Britanniques –
8: Noirs/Africains/Caribéens/Noirs britanniques -9: Autres groupes ethniques