

CONSEIL D'ORIENTATION DES RETRAITES

Séance plénière du 22 octobre 2013 à 14h30

« Les outils statistiques et les systèmes d'information sur les retraites »

<b>Document N°1</b>
---------------------

<i>Document de travail, n'engage pas le Conseil</i>
---

## **Note de présentation générale**

*Secrétariat général du Conseil d'orientation des retraites*



## Les outils statistiques et les systèmes d'information sur les retraites

Ce dossier s'intéresse aux sources statistiques d'information en matière de retraite, c'est-à-dire l'ensemble des sources de données – enquêtes par sondage, bases de données extraites de fichiers administratifs, appariements entre fichiers, etc. – qui renseignent et peuvent être mobilisées pour mieux connaître les caractéristiques des retraites, des retraités et des cotisants.

La problématique de la disponibilité de ces sources d'information dépasse le cadre purement technique. De cette disponibilité dépend en effet la capacité à connaître les caractéristiques des affiliés, cotisants ou retraités, à mesurer l'adéquation du système de retraite à ses objectifs ou encore à apprécier, de manière quantitative, les éventuelles inégalités ou inéquités. Ces sources d'information constituent également la brique de base nécessaire pour construire des modèles de projection : d'elles dépend donc la capacité à développer les outils d'une analyse prospective, et donc la capacité, en pratique, à piloter le système de retraite.

La disponibilité de sources d'informations est une problématique spécifique, qui ne doit pas être confondue avec la disponibilité d'études et analyses, ni avec l'existence et le contenu des fichiers de gestion des régimes de retraite.

Le fait qu'aucune analyse n'ait été publiée ou ne soit accessible au grand public sur telle ou telle question spécifique ne signifie pas forcément qu'aucune source d'information qui permettrait d'éclairer cette question ne soit (potentiellement) disponible. L'enjeu en termes de délais est à cet égard important : alors que la réalisation d'études nouvelles à partir de sources statistiques déjà existantes peut être relativement rapide, la constitution de nouvelles données statistiques est une opération lourde et généralement longue à finaliser. L'anticipation des besoins de connaissance doit de ce fait être beaucoup plus grande en ce qui concerne la construction des sources statistiques qu'en ce qui concerne les programmes d'études.

Par ailleurs, les fichiers de gestion et les bases de données statistiques sont de nature différente : l'existence d'un fichier de gestion n'implique pas nécessairement qu'on puisse en extraire un fichier contenant les mêmes informations dans un but d'analyses statistiques. Cela peut résulter d'obstacles juridiques, parce que les textes créant les fichiers de gestion ne prévoient pas – et donc n'autorisent pas – leur usage à des fins statistiques, ou du fait que cet usage n'est autorisé que pour les gestionnaires des fichiers, ce qui exclut les services d'étude et de recherche. Il peut également y avoir des raisons techniques, car les impératifs en termes de format et de contenu des fichiers ne sont pas les mêmes pour une finalité de gestion ou une finalité statistique<sup>1</sup>. Dans tous les cas, la constitution de systèmes d'information statistiques

---

<sup>1</sup> Cela concerne notamment la volumétrie des fichiers : les fichiers de gestion complets sont généralement trop gros pour permettre de réaliser des traitements statistiques dans des délais raisonnables. Cela peut également être lié à la complétude des données : une incomplétude partielle peut ne pas remettre en cause l'utilisation d'un fichier de gestion – les cas d'incomplétude pouvant être traités par un traitement manuel des opérateurs au cas par cas –, alors qu'elle peut représenter un obstacle majeure pour un fichier statistique, si elle remet en cause son caractère représentatif. Enfin, l'usage statistique d'une base de données nécessite une homogénéité des concepts utilisés pour toutes les observations (en particulier en ce qui concerne les dates d'observation, de mise à jour, etc.), ce qui n'est pas forcément le cas dans un fichier à finalité de gestion.

doit donc être vue comme indépendante, même si très complémentaire, de celle des fichiers de gestion.

Certaines limites des sources statistiques actuelles, l'émergence de nouveaux référentiels ou traitements inter-régimes, ainsi que les besoins liés au développement des modèles de projection – et donc à la capacité à piloter le système de retraite – mettent en lumière deux voies de progrès possibles du système d'information actuel sur les retraites :

- malgré la richesse du système actuel, certaines thématiques nécessiteraient des sources d'information qui, aujourd'hui, n'existent pas. Ces besoins pourraient amener à la réalisation de nouvelles enquêtes, mais ils conduisent aussi à souligner l'intérêt que représenteraient de nouveaux appariements des données administratives sur les retraites, avec d'autres données administratives (par exemple, données sur les retraites et données fiscales) ou bien avec des données d'enquête ;
- l'expérience récente des nouveaux référentiels inter-régimes montre que, parce que les finalités statistiques de ces référentiels n'avaient pas été prévues initialement, leur mobilisation pour la connaissance du système de retraite a impliqué plusieurs années de délais – d'où la nécessité de prendre en compte, dès l'origine des projets de mise en place des systèmes d'information, les finalités statistiques, ainsi que les modalités d'extraction et d'utilisation afférentes.

## 1. Le système statistique actuel d'information sur les retraites

Les revenus des retraités proviennent en grande partie des pensions versées par les régimes de retraite légalement obligatoires. Dans le système statistique actuel d'information sur les retraites (**document n°2**), les données extraites des fichiers administratifs de gestion de ces régimes constituent donc naturellement la source d'information principale.

Chaque régime gère ainsi son ou ses propres fichiers statistiques (à titre d'exemple, le Système National Statistique Prestataires – SNSP – du régime général ou la base des pensions du Service des retraites de l'Etat).

Plusieurs régimes ont récemment amélioré très sensiblement leurs fichiers statistiques ou sont en train de le faire. Les possibilités en termes d'analyses et d'études devraient donc s'accroître dans un avenir proche. Ces améliorations des bases de données statistiques font généralement écho aux refontes des systèmes d'information, pour répondre en particulier à la mise en place du droit à l'information en matière de retraite. C'est notamment le cas dans les régimes de la fonction publique : service des retraites de l'Etat (**document n°4**) et CNRACL (**document n°5**). D'autres fichiers statistiques nouveaux ont été mis en place suite à la création de régimes, comme c'est le cas pour le régime additionnel de la Fonction publique, ou RAFP (**document n°6**). Dans le même ordre d'idées, la fusion en 2002 des moyens des fédérations AGIRC et ARRCO a conduit à revoir entièrement le système d'information statistique de ces régimes (**document n°7**).

La connaissance d'un seul régime à la fois ne fournit toutefois qu'une information partielle sur les retraites de chaque individu, puisque la plupart des retraités perçoivent des pensions de plusieurs régimes – régime(s) de base et régime(s) complémentaire(s). Des sources de données inter-régimes sont de ce fait nécessaires pour avoir une image plus globale des droits

à retraite. Les échantillons inter-régimes construits par la DREES – en rapprochant, pour un échantillon anonyme d'individus, les données des différents régimes français légalement obligatoires – constituent, en ce domaine, les sources de référence : l'échantillon inter-régimes de retraités (EIR), qui permet de reconstituer le montant de la retraite globale (tous régimes confondus, droits directs et droits dérivés) ainsi que ses divers déterminants, et l'échantillon inter-régimes de cotisants (EIC), qui permet de reconstituer l'intégralité de la carrière année après année (**document n°2**).

Du fait de la lourdeur de leur processus de collecte, ces échantillons inter-régimes ne sont toutefois constitués que tous les quatre ans. D'autres sources sont donc nécessaires, afin de disposer d'une actualisation annuelle et plus récente des principaux résultats concernant les retraites tous régimes. Ces sources peuvent se situer à la limite entre production statistique et modélisation, comme l'illustre notamment l'exemple du modèle de microsimulation statique ANCETRE, qui permet de calculer annuellement le nombre de retraités et le montant de pension tous régimes, en synthétisant au mieux les données du dernier EIR disponible et de l'enquête annuelle auprès des caisses de retraite de la DREES (**document n°9**).

Une limite importante des données issues de fichiers administratifs est que ces données contiennent généralement très peu d'informations autres que celles directement utiles pour calculer les retraites versées par les régimes, et en particulier très peu d'informations sur les caractéristiques sociodémographiques des personnes. Pour étudier certaines problématiques concernant les retraités, d'autres sources statistiques doivent donc être mobilisées. Il s'agit soit d'enquêtes générales, couvrant toute la population française et incluant, à ce titre, des retraités (notamment les enquêtes en population générale de l'INSEE : enquête Patrimoine, enquête Revenus Fiscaux et Sociaux, etc.) ; soit d'enquêtes spécifiques, étudiant une thématique en particulier (l'enquête annuelle de la DREES sur la retraite supplémentaire facultative, l'enquête sur les bénéficiaires de minima sociaux – dont ceux du minimum vieillesse, etc.) ou ciblant une sous-population de retraités ou de personnes âgées (par exemple, l'enquête européenne SHARE sur les personnes de 50 ans et plus).

Même si le système actuel d'information statistique sur les retraites est déjà très riche, plusieurs limites subsistent.

Pour ce qui concerne les données issues des fichiers administratifs des régimes, ces limites tiennent notamment à la difficulté de reconstituer l'information sur les périodes passées – antérieures à la création des fichiers. Cela implique des délais longs avant que les nouveaux fichiers statistiques deviennent effectivement opérationnels, et puissent être mobilisés pour des études et analyses. A titre d'exemple, la complétude de tous les comptes individuels de retraite des fonctionnaires d'Etat n'est pas attendue avant la fin de 2014, et la complétude des carrières des fonctionnaires territoriaux et hospitaliers, pour ce qui concerne les périodes antérieures à 2011, est encore en cours d'expertise à la CNRACL. En outre, certains obstacles juridiques peuvent subsister, qui font que l'amélioration des systèmes d'information des régimes ne suffit pas pour accroître la connaissance. Par exemple, même si les régimes gérés par la Caisse des dépôts ont une proximité d'organisation, qui fait que leurs données pourraient en théorie être rapprochées pour des études, ce rapprochement n'est en pratique pas toujours possible<sup>2</sup>.

---

<sup>2</sup> Des rapprochements sont possibles entre les informations de la CNRACL et de l'IRCANTEC, permettant d'étudier les personnes passant d'un statut de contractuel à un statut de titulaire des fonctions publiques territoriale ou hospitalière. En revanche, des rapprochements – qui viseraient à étudier les diverses composantes,

D'autres limites peuvent être liées à la taille des bases de données statistiques actuelles. Par exemple, le nombre de retraités ou de cotisants observés dans les échantillons interrégimes de la DREES peut, malgré la taille relativement élevée de ces échantillons, être insuffisant pour étudier de petits régimes (par leur nombre d'affiliés ou de retraités) ou certaines sous-populations spécifiques. De même, les effectifs enquêtés dans certaines enquêtes, comme l'enquête SHARE (pour ses premières vagues) ou certains panels généraux de l'INSEE comme SRCV (statistique sur les ressources et conditions de vie), peuvent être trop faibles pour étudier des problématiques particulières.

Certaines thématiques nécessiteraient enfin des sources d'information statistique qui, à l'heure actuelle, n'existent pas. La capacité à les étudier empiriquement reste donc très limitée. Il n'y a pas, ainsi, de sources ayant à la fois une information riche sur les droits acquis à la retraite – issue de données administratives – et sur les liens entre personnes d'un même ménage : il est alors difficile d'étudier les décisions conjointes de départ à la retraite entre individus d'un même ménage. De même, il est difficile d'apprécier en quoi la prise en compte des retraites supplémentaires d'entreprise modifient le diagnostic sur les niveaux de pensions et les disparités entre assurés, faute de bases individuelles rassemblant des informations sur les retraites légalement obligatoire et sur les retraites supplémentaires facultatives.

Ces besoins conduisent à souligner l'intérêt que pourraient représenter des appariements des données administratives sur les retraites avec d'autres données administratives – comme la DREES envisage de le réaliser entre les données de l'EIR et des données fiscales – ou bien avec des données d'enquête.

## **2. Une montée en charge des référentiels inter-régimes**

Des évolutions réglementaires du système de retraite ont modifié récemment le champ et le contenu des référentiels de gestion, ce qui devrait conduire à s'interroger sur leurs conséquences pour la constitution de fichiers statistiques. De nouveaux fichiers inter-régimes de gestion sont ainsi créés afin, notamment, de mettre en œuvre les mécanismes de calcul faisant intervenir le niveau de pension « tous régimes » (pour l'écêtement des minima de pension ou pour la majoration de pension de réversion) ou dans le cadre des campagnes d'information en matière de retraite. A ce jour, ils ne font toutefois pas l'objet d'une utilisation statistique.

Trois nouveaux répertoires avec une vocation inter-régimes ont ainsi été créés récemment ou doivent l'être prochainement (**document n°3**) :

- l'Echange Inter-Régimes de Retraite (EIRR), mis en place en décembre 2009, qui permet de stocker dans un lieu unique l'ensemble des informations fournies par les régimes de retraite afin de calculer de façon automatique la majoration de pension de réversion, la majoration de pension non salariée agricole et le minimum contributif « tous régimes » ;
- le Répertoire national commun de la protection sociale (RNCPS), créé par la loi de financement de la sécurité sociale pour 2007 et étendu lors de la réforme des retraites de 2010, qui recense les bénéficiaires de prestations sociales obligatoires (retraites de

---

primes et hors primes, des rémunérations des fonctionnaires – ne sont à l'heure actuelle pas possibles entre les informations de la CNRACL et du RAFF.

- base et complémentaire, maladie, famille, chômage, etc.) ainsi que toutes les prestations qui leur sont versées, sans en indiquer les montants ;
- enfin, le Répertoire de Gestion des Carrières Unique (RGCU), créé à l’occasion de la réforme des retraites de 2010 et dont le projet de loi de réforme des retraites de 2013 prévoit l’extension, qui doit permettre d’avoir – lorsque ce répertoire sera opérationnel dans quelques années – une connaissance de l’ensemble des éléments concernant la carrière des assurés, pour l’ensemble des régimes de retraite légalement obligatoire, et ce quel que soit l’âge de l’assuré.

Ces référentiels de gestion ne prévoyaient pas, à leur création, de finalité statistique, et n’ont pour l’instant pu être mobilisés pour des travaux d’étude. La finalité statistique a toutefois été ajoutée depuis. En ce qui concerne l’EIRR, la mise en œuvre d’un flux spécifique dédié aux statistiques permettra, à partir du 1er janvier 2014, la mise à disposition des données de l’EIRR à tous les régimes demandeurs. En ce qui concerne le RNCPS, un projet pilote a été lancé<sup>3</sup> en partenariat avec l’INSEE, qui porte sur la fourniture à ce dernier d’une extraction anonymisée du répertoire, afin de réaliser des comptages de population. Pour le RGCU, enfin, le chantier n’a pas prévu à ce jour de projet précis d’utilisation statistique mais il incorpore déjà en amont les requêtes des services statistiques qui s’appuieraient sur lui pour étendre ou perfectionner les travaux déjà menés sur les référentiels de carrières de chacun des régimes.

Ces projets d’exploitation à des fins statistiques, qui devraient déboucher dans un avenir proche, ne couvrent pas, néanmoins, tout le potentiel de ces référentiels inter-régimes pour les études et analyses. Par exemple, les possibilités d’accès pour les chercheurs à des échantillons anonymisés extraits de ces référentiels, de même que leur mobilisation dans le cadre de la production des bases de données d’étude du système statistique public – notamment pour compléter et enrichir les échantillons inter-régimes EIR et EIC de la DREES voire, à plus long-terme, pour revoir leurs modalités de constitution – mériteraient d’être étudiées.

Des données inter-régimes de retraite sont également rassemblées, depuis 2007, pour la réalisation des campagnes d’information retraite (**document n°8**). Il s’agit d’une part de données sur les nombres d’affiliés aux différents régimes de retraite, rassemblées dans le cadre de l’annuaire du GIP Info Retraite, et d’autre part de données sur les droits acquis pour la retraite (trimestres validés, points acquis, etc.), rassemblées – chaque année pour les générations concernées – pour la constitution des relevés individuels de situation (RIS) et des estimations indicatives globales (EIG). Là-encore, ces informations n’ont pas fait et ne font pas, pour l’instant, l’objet d’une exploitation statistique systématique<sup>4</sup>, le GIP Info Retraite n’ayant pas, d’après les missions qui lui ont été confiées, vocation à produire des statistiques sur les carrières.

Ces situations, même si elles sont compréhensibles au vu des finalités initiales des référentiels mis en place (dans une optique de gestion), ne sont guère satisfaisantes, compte tenu des besoins d’informations statistiques pour le suivi et le pilotage du système de retraite. On ne peut donc que souligner la nécessité de prendre en compte, dès l’origine des projets de mise

---

<sup>3</sup> La problématique du RNCPS est spécifique, dans la mesure où ce répertoire n’est pas une base de données classique, renvoyant à la demande les différentes prestations perçues par l’assuré.

<sup>4</sup> Des travaux techniques ont néanmoins été réalisés en 2007 et 2008, afin d’étudier la faisabilité d’une éventuelle utilisation statistique des données du GIP Info Retraite. Quelques opérations ponctuelles, impliquant une extraction de données du GIP pour un usage statistique, ont également été menées (appariement en 2009 entre données des RIS et données de l’EIC, appariement en 2011 et 2012 entre EIG et données sur les pensions effectivement liquidées à la CNAV et à l’AGIRC-ARRCO).

en place des systèmes d'information, les finalités statistiques, ainsi que les modalités d'extraction et d'utilisation afférentes.

### 3. A la limite des sources d'information : les outils de projection

La constitution d'indicateurs pour le suivi et le pilotage nécessite parfois des informations qui ne peuvent être fournies par aucune source statistique existante. Il est alors nécessaire d'avoir recours à des techniques de modélisation économique et de simulation pour les estimer. Dans ce cadre, la problématique de la disponibilité des sources d'information statistique reste cruciale, puisque les outils de modélisation et de simulation sont généralement des outils empiriques, qui s'appuient – en les complétant – sur les données statistiques existantes. En particulier, l'apport des modèles est bien souvent de synthétiser des informations diverses, déjà disponibles mais séparées dans des sources statistiques distinctes, en vue de répondre à des problématiques spécifiques.

On a déjà cité l'exemple du modèle ANCETRE de la DREES (**document n°9**), dont l'objectif est de permettre une estimation actualisée chaque année du nombre de retraités et des montants de pension « tous régimes ». Cet outil est à la limite entre production statistique et modélisation, au sens où il n'intègre aucune dimension de simulation ou de modélisation de comportements économiques. Son objet est de synthétiser au mieux les données du dernier EIR disponible et de l'enquête annuelle auprès des caisses de retraite de la DREES, c'est-à-dire finalement de produire un EIR « actualisé » – à défaut de pas pouvoir itérer chaque année l'EIR du fait de la lourdeur de son processus de collecte.

La dimension de synthèse de sources est plus grande encore en ce qui concerne les modèles de microsimulation dynamique, nécessaires pour construire des indicateurs prospectifs. Ces modèles s'appuient chacun sur une source particulière, qui constitue son échantillon de base, mais ils utilisent souvent, de surcroît, d'autres sources d'information, mobilisées pour fournir des données de paramétrages, de calage ou de calibrage<sup>5</sup>.

Ainsi, le modèle TRAJECTOIRE (**document n°10**), récemment créé par la DREES<sup>6</sup>, utilise un échantillon tiré de l'EIC, mais repose aussi sur les données d'Etat-civil produit par l'INSEE (données de calage pour les effectifs par génération), sur des données du service statistique de l'immigration (données de calage sur les flux migratoires) et sur l'EIR (utilisé pour l'estimation des probabilités de sortie d'activité et de départ à la retraite).

Le modèle DESTINIE de l'INSEE (**document n°11**), dont l'échantillon a récemment fait l'objet d'un rebasage, repose principalement sur l'enquête Patrimoine, mais utilise également de très nombreuses autres sources, comme données de calage ou pour estimer les probabilités associées aux diverses transitions : enquête Emploi, EIC, enquête Génération, enquête Histoire familiale, enquête Formation et Qualification Professionnelle, etc.

---

<sup>5</sup> Les données de *paramétrage* servent à estimer les valeurs numériques des paramètres du modèle (c'est-à-dire les probabilités associées aux différents changements d'état démographique et professionnel, ainsi que les déterminants de niveau de salaire ou de revenu d'activité). Les données de *calage* définissent les valeurs de certains agrégats, qu'on souhaite contraindre dans le modèle (par exemple, le nombre de personnes dans chaque génération). Enfin, les données de *calibrage* fournissent des cibles qui, sans que les modèles soient nécessairement contraints pour reproduire exactement ces cibles, aident à affiner les modélisations.

<sup>6</sup> Ce modèle succède au modèle de projection PROMESS.

S'ils ne constituent pas des sources statistiques en tant que telles, les outils de projection sont évoqués dans ce dossier afin de souligner les besoins statistiques que leur développement implique, ainsi que le lien étroit qui, par leur intermédiaire, relie la thématique de la disponibilité des sources statistiques à la capacité d'assurer, en permettant de développer une vision prospective, le suivi et le pilotage du système de retraite.