

Big data et protection sociale : au-delà de la lutte contre la fraude, des opportunités à saisir pour améliorer l'accès aux droits¹

Par Ange CHEVALLIER et Géraldine TAUBER, élèves de la 56^{ème} promotion de l'EN3S

Introduction

En 2016, les Caisses d'allocations familiales (CAF) ont détecté près de 43 000 fraudes, soit 8 % de plus que l'année précédente, pour un montant de 275 millions d'euros. La Caisse nationale (CNAF) a expliqué cette hausse par l'accroissement de l'efficacité des dispositifs de contrôle et non par l'augmentation de la fraude elle-même. Ces résultats, ajoutés à ceux des années précédentes sont notamment liés au développement du *data mining*, un procédé qui repose sur le *big data*. À l'instar des administrations fiscales, les organismes de protection sociale ont en effet un usage de plus en plus poussé de ces nouvelles méthodes d'analyse.

La notion de « big data » renvoie aux données de masse produites par les administrations, les entreprises, les utilisateurs et toute autre entité en capacité d'en créer. C'est le volume en présence, la quantité d'information qui en fait sa particularité. Une fois traitées, ces données peuvent en effet révéler des corrélations qui n'avaient pu être jusqu'alors appréhendées par les outils traditionnels du traitement statistique. L'ère du *big data* traduit ainsi une nouvelle façon d'appréhender les données, non plus uniquement comme des éléments passifs permettant le traitement de dossiers, mais comme des outils actifs pouvant avoir une influence sur la façon dont on traite les situations.

Pendant longtemps en effet, les entreprises et administrations ont stocké des informations sur leurs clients et administrés, sans pour autant en tirer tout leur potentiel. C'est notamment l'arrivée de nouveaux services internet fondés sur la gratuité (Facebook, Twitter, Google...) qui a poussé ces entreprises à donner de la valeur à leurs données pour s'assurer des revenus. En analysant les comportements des internautes, les pages consultées, les informations personnelles et d'autres variables, les entreprises ont ainsi pu personnaliser les campagnes publicitaires en ciblant davantage les potentiels clients.

Les services publics, notamment les organismes de Sécurité sociale, se sont plus récemment lancés dans l'analyse de leurs données de masse. La CNAF a par exemple été pionnière dans le développement d'outils de *data mining* dans le cadre de la lutte contre la fraude pour s'adapter à des circuits de contournement des règles de plus en plus complexes et innovants (I). Les technologies *big data* offrent toutefois d'autres champs d'action, pour simplifier les démarches des assurés et réduire le risque de non-recours aux droits mais aussi renforcer les outils d'évaluation et d'élaboration des politiques publiques (II). Toutefois, pour être à la hauteur

¹ Cet article s'appuie notamment sur des entretiens réalisés auprès de Stéphane DONNÉ, responsable du département des statistiques, système d'information et Big Data de la CNAF et de Fabien TARISSAN, chargé de recherches au CNRS.



de ces ambitions, des prérequis sont nécessaires : l'évolution des systèmes d'informations pour fluidifier les échanges d'informations inter-organismes et l'adaptation du management et de la gestion des ressources humaines des organismes (III). D'autres applications du *big data*, comme la personnalisation de la gestion des risques peuvent également être envisagées.

I- Concept renvoyant à des réalités diverses, le *big data* est aujourd'hui principalement utilisé dans le cadre de la lutte contre la fraude

I-1/ Comprendre et appréhender la notion de *big data* : un concept vaste offrant des opportunités pour la sphère des services publics

A/ Le *big data* est un concept qui se caractérise moins par une définition stricte que par le périmètre qu'il embrasse

Le concept de *big data* ne peut être défini précisément. On retrouve toutefois généralement trois dimensions, ou trois « V », proposés par Doug laney au début des années 2000, et qui permettent de le caractériser : le volume, la vitesse et la variété. Le volume renvoie à la quantité de données, qui est telle que les outils statistiques et informatiques traditionnels ne peuvent pas les traiter en l'état. La vitesse décrit quant à elle le besoin d'adaptation rapide des traitements aux données, au fil de leur arrivée. Enfin, la variété est liée au fait que l'on va traiter tout type de donnée sans nécessairement les présélectionner, toute donnée devenant potentiellement utilisable. Dans un domaine où les potentialités sont en pleine expansion, la définition du concept importe finalement peu, et c'est son périmètre qu'il s'agit donc de bien saisir pour éviter d'apposer un label « big data » à des méthodes statistiques traditionnelles.

L'un des grands apports des technologies *big data* par rapport aux outils statistiques classiques est leur dimension prédictive. De fait, les traitements statistiques traditionnels empruntaient une méthode d'analyse dite « hypothético-déductive » qui consistait à réfléchir en amont aux données à utiliser et aux variables qui permettraient de les exploiter, en proposant des hypothèses de travail à vérifier. Le *big data* tend à limiter les analyses préalables et à laisser les algorithmes extraire des régularités et des incongruités de la masse de données en présence, diminuant ainsi le biais humain des traitements statistiques traditionnels. Toutefois, selon les outils utilisés (*data mining*, *machine learning* etc.) la dimension prédictive est plus ou moins importante et le biais plus ou moins réduit. À l'inverse, plus l'analyse est dépourvue de cadre d'analyse *a priori*, plus l'interprétation des résultats est difficile pour l'humain. Le *big data* englobe donc des outils à géométrie variable et la position des curseurs a toute son importance dans leurs usages.

Petit lexique du *big data*

Data mining : le *data mining* renvoie aux méthodes permettant de traiter un très grand volume de données et d'en tirer des corrélations significatives afin, par la suite, d'analyser des données similaires à travers ces modèles.

Machine learning : contrairement aux algorithmes classiques, la logique du *machine learning* se veut auto-apprenante : c'est à partir des données que l'algorithme va se concevoir, permettant une adaptation rapide à leur évolution, particulièrement utile par exemple dans des processus de lutte contre la fraude. Le problème est que l'humain n'a plus la main sur le traitement et la difficulté réside dans la compréhension de l'analyse *a posteriori*.

Smart data : le concept de *smart data* se traduit par un usage plus rationalisé des données de masse, avec une sélection rigoureuse des données et de leurs usages principalement en amont de leur collecte, pour éviter la déperdition d'énergie.

B/ Les organismes gérant des services publics se sont engagé progressivement dans les technologies du *big data*

Les administrations, organismes et entreprises gérant des services publics disposent de bases de données pléthoriques qui permettent d'envisager des traitements sur des volumes suffisants, avec une variété d'informations particulièrement importante (données sociales, données fiscales, état civil, données de santé, données relatives à l'énergie, aux transports etc.). Ce gisement de données fait apparaître les services publics comme des acteurs incontournables dans le développement des technologies *big data*. Le Premier ministre Édouard Philippe a d'ailleurs rappelé, le 8 septembre 2017 les atouts des entreprises et administrations françaises en la matière, alors qu'il missionnait Cédric Villani pour concevoir une stratégie sur l'intelligence artificielle, intitulée #missionIA.

Les administrations publiques sont toutefois rentrées progressivement dans le *big data*, par le cadre de la lutte contre la fraude. Évoluant rapidement, les pratiques de fraude aux impôts ou aux prestations nécessitent une réactivité précoce des équipes chargées de les prévenir et de les limiter. De plus, le ciblage des contrôles favorisant une rationalisation de ceux-ci, les gains réalisés permettaient de justifier l'investissement dans ces technologies. Si le réseau des CAF a été le premier à déployer ces outils en 2012, d'autres entités ont exploité leurs données avec des méthodes en particulier basées sur le *data mining*, comme la Direction générale des Finances publiques à partir de 2014, l'administration des douanes depuis 2015 ou encore Pôle emploi et l'Agence centrale des organismes de Sécurité sociale (ACOSS) en 2014.²

² Délégation nationale à la lutte contre la fraude, 2017, Bilan 2015 – Lutte contre la fraude, Ministère de l'Économie et des Finances, pp. 28-30



I-2/ Lutter contre la fraude et verser le juste droit : l'exemple de la détection des anomalies grâce au *data mining* dans la branche Famille

A/ Depuis 2010, les CAF ont fait un usage de plus en plus performant du *big data* dans les dispositifs de contrôle au travers du *data mining*

Entre 2011 et 2013, le rendement des contrôles sur place effectués par la CAF s'est accru de 46 % grâce aux méthodes de ciblage fondées sur le *data mining*. En 2011, les contrôles sur place permettaient de détecter 150 millions d'euros d'anomalies contre 263 millions en 2013³. Cette hausse spectaculaire s'explique par l'introduction puis le déploiement d'outils de *data mining* depuis 2004 dans le réseau.

Dès 2004 en effet, la CAF de Dijon puis celle de Bordeaux se sont intéressées aux méthodes de *data mining* au travers de leur agence comptable, afin de combler une lacune en matière de lutte contre la fraude. Les contrôles visés portaient sur les fraudes en réseau aux aides au logement. Elles devaient donc être contrecarrées par des méthodes efficaces et réactives. S'appuyant sur une base de données ramassée de quelques milliers de dossiers, les modèles initiaux ont permis de dégager de premiers profils frauduleux, affinés ensuite grâce à l'appui d'autres bases de données et de la CNAF. C'est à partir de 2010 que le dispositif de *data mining* a été généralisé dans l'ensemble du réseau des Caf au travers des plans de contrôle, avant d'être consacré par une lettre circulaire en 2011 puis par la Convention d'objectifs et de gestion (COG) 2013-2017 signée entre la CNAF et l'État.

B/ Avec un objectif plus large de réduction des anomalies, le périmètre du *big data* ne s'arrête toutefois pas au *data mining*

Participant de la logique d'assurance raisonnable qui irrigue la comptabilité publique, le *data mining* permet de rationaliser les opérations de contrôle en les ciblant mieux, à travers une approche en trois étapes⁴ : dans un premier temps une base de données dédiée à l'étude des risques d'indus est créée. En second lieu il s'agit de procéder à l'analyse inductive de cette base de données, pour faire ressortir des régularités et créer des modèles. Enfin, dans une troisième étape les modèles sont intégrés dans les plans de contrôle afin de les appliquer à l'ensemble des données et de scorer les dossiers selon leur degré d'exposition aux risques.

Outre l'augmentation de la rentabilité des opérations de contrôle, cette méthode permet également de réduire le nombre d'opérations mal ciblées et donc le sentiment des allocataires d'être contrôlés à tort ou pour des raisons injustifiées. De fait, on ne cible plus *a priori* des populations à risque mais des profils de dossiers à risque, sur des critères plus objectifs, en diminuant le biais humain qui prévalait

3 DUBOIS, Vincent, PARIS, Morgane, WEILL, Pierre-Édouard, 2016, *Politique de contrôle et lutte contre la fraude dans la branche Famille*, Dossier d'étude de la CNAF n°183, CNAF, p. 79

4 Pierre Collinet, « Focus – Le *data mining* dans les CAF : une réalité, des perspectives », *Informations sociales* 2013/4 (n° 178), p. 129-132.

dans les contrôles traditionnels. Pour autant, le Défenseur des droits a récemment souligné les risques posés par les méthodes de *data mining* en termes de discrimination, lorsque ces outils viennent renforcer, voire légitimer la stigmatisation de profils à « risque ».⁵ Il soulève ici la question de la méthode retenue dans l'utilisation des technologies d'analyse des données de masse.

L'objectif n'est toutefois pas la seule augmentation du rendement des contrôles mais la réduction des risques d'anomalies de manière générale. Les premiers groupes de travail montés en 2004 visaient ainsi non seulement à prévenir les risques de fraudes, mais aussi à diminuer les risques « métier », reposant sur les erreurs des techniciens liés à une législation de plus en plus complexe et mouvante. Au-delà de l'aspect quantitatif, il y a donc également un objectif de qualité, qualité du contrôle et qualité du traitement des dossiers : la logique de versement du « juste droit » opère dans les deux sens, que l'anomalie ciblée soit en défaveur ou en faveur de l'assuré.

Pour autant, d'autres technologies du *big data*, notamment le *machine learning*, pourraient renforcer la logique prédictive. En effet, plus l'identification du risque d'anomalie arrive tôt dans le processus de contrôle, plus le risque d'avènement de l'erreur ou de la fraude peut être évité. Les opportunités offertes par le *big data* dépassent ainsi largement le cas de l'utilisation du *data mining* dans le cadre de la lutte contre la fraude. D'autres technologies et d'autres objectifs sont à rappeler : le ciblage de la communication sortante, les campagnes de prévention, l'action sociale. Et parmi eux le pendant de la lutte contre la fraude : la lutte contre le non-recours aux droits sociaux.

II- L'utilisation du *big data* contre le non-recours aux droits pourrait davantage outiller les politiques publiques face à ce phénomène

II-1/ Lutter contre le non-recours : une préoccupation des pouvoirs publics insuffisamment outillée

A/ L'accès aux droits a été progressivement identifié comme un enjeu majeur des politiques publiques

La lutte contre le non-recours émerge dans les années 1970, sous la pression de deux phénomènes : un poids croissant des prestations dans les ressources allocataires et la complexification des droits du fait notamment de la mise sous conditions de ressources. Cette préoccupation a été relancée lors de la mise en place du RMI puis du RSA. En effet, le RSA et ses variantes, le RSA « activité » et le RSA « jeunes » présentent ou ont présenté des taux de non-recours particulièrement élevés, estimés à 50 % en moyenne en 2011, le taux le plus fort étant celui du RSA « activité » qui s'élevait à 68 %.⁶

5 Défenseur des Droits, 2017, Rapport « Lutte contre la fraude aux prestations sociales : à quel prix pour les droits des usagers ? », Septembre 2017, pp. 19-20

6 WARIN, Philippe, 2011, Document de travail n°13 – « Le non-recours au RSA : des éléments de comparaison », ODENORE, décembre 2011, p. 2



Les premières analyses des motifs de non-recours ont permis de distinguer deux logiques. D'abord une logique du non-recours involontaire : dans ce cas, il n'est pas le fruit d'un choix car l'utilisateur n'a pas connaissance de ses droits ou ne s'estime pas éligible. À l'inverse, dans la logique de non-recours volontaire, le bénéficiaire potentiel sait qu'il pourrait être destinataire d'une prestation sociale. Son non-recours vient alors d'un ressenti négatif vis-à-vis de cette prestation : il peut craindre une stigmatisation, il peut estimer que la trop grande complexité administrative n'en vaut pas l'enjeu, ou encore il peut craindre un risque d'indu lié à une erreur de gestion des organismes de protection sociale.⁷

Dans une étude de 2014, le Secrétariat Général pour la Modernisation de l'Action Publique (SGMAP) a quant à lui identifié une série de facteurs explicatifs du non-recours. Figurent parmi ceux-ci une mauvaise image de la prestation, un manque d'information, l'absence de proactivité de la part de l'utilisateur, la non-connaissance de l'éligibilité, le déficit de médiation et d'accompagnement, la complexité de la prestation, la rupture du processus d'instruction ou encore la concurrence d'une autre prestation.⁸

Cependant, bien que les motifs de non-recours soient de mieux en mieux identifiés, les politiques d'accès aux droits peinent à augmenter significativement la plénitude des droits en particulier pour les publics les plus précaires.

B/ Les outils restent encore insuffisants pour lutter contre le non-recours aux droits sociaux

Malgré les nombreuses études et actions⁹, le phénomène de non-recours reste une réalité insuffisamment outillée. Pour autant, des pistes émergent. Les préconisations du rapport de l'Assemblée nationale d'octobre 2016¹⁰ sur l'évaluation des politiques publiques en faveur de l'accès aux droits sociaux portent ainsi sur une meilleure identification des bénéficiaires potentiels au travers du *data mining*, la détection du non-recours par échanges de données entre organismes et des actions proposant notamment des rendez-vous lors d'événements marquants pour les assurés.

Dans la Branche Famille, l'accès aux droits s'est traduit, entre autres actions, par la mise en place des « Rendez-vous des droits » sur la période conventionnelle 2013-2017. Le dispositif consiste à proposer à un allocataire, suite à un événement

7 Pour une typologie plus détaillée des motifs de non-recours, se reporter à WARIN, Philippe, 2017, *Le non-recours aux politiques sociales*, PUG

8 KESTEMAN, Nadia, « Le fondement des politiques d'accès aux droits dans la branche famille du régime général : le non-recours aux prestations », Regards n°46, septembre 2014, EN3S.

9 Notamment le volet d'accès aux droits du Plan pluriannuel de lutte contre la pauvreté, qui s'articulait autour de trois axes : mieux connaître les phénomènes de non-recours, mieux détecter et mieux informer (au travers des « Rendez-vous des droits » pour la branche Famille et du dispositif « PLANIR » pour la branche Maladie).

10 BIÉMOURET, Gisèle, COSTES, Jean-Louis, 2016, *Rapport d'information sur l'évaluation des politiques publiques en faveur de l'accès aux droits sociaux*, n°4158, 26 octobre 2016, Assemblée nationale, pp. 116-134, notamment proposition n°8 p. 134.

de la vie, ou la survenue d'un « risque » (premières demandes de RSA, rupture, grossesse...), un entretien pour faire un état des lieux de sa situation et de ses droits potentiels aux prestations sociales. À titre d'exemple, dans le cas d'un primo-demandeur de RSA, le conseiller l'informe sur la prestation mais peut également effectuer une préouverture du droit à la CMU-C.

Les premières opérations de ciblage et les parcours attentionnés ont porté leurs fruits. En 2016, la moitié des rendez-vous avaient donné lieu à une ouverture de droits nouveaux¹¹. Pour autant, on peut imaginer à l'avenir un renforcement de ces dispositifs par davantage de proactivité des caisses pour aller chercher les bénéficiaires potentiels, une segmentation plus fine des publics, une simplification de l'accès aux droits (et pas forcément une simplification du droit) voire une augmentation des pré-liquidations ou des liquidations automatiques. À cet égard, les technologies du *big data* ont une place de choix dans l'avenir des politiques d'accès aux droits.

II-2/ Aller au plus près des situations : l'exploitation des données du *big data* permet un ciblage plus précis tout en interrogeant les orientations des politiques sociales

A/ L'analyse plus fine des situations individuelles est rendue possible grâce à la micro-segmentation

Le *big data* pourrait en premier lieu améliorer le recours aux droits par une meilleure connaissance des comportements des usagers. L'intérêt de cette démarche serait d'avoir une vision exhaustive, une analyse globale des situations, qui ne serait pas nécessairement orientée par des choix politiques. La phase d'analyse et d'étude se détacherait ainsi d'une lecture politique du phénomène du non-recours et permettrait de l'objectiver. Ce n'est qu'ensuite que l'exploitation de ces analyses et surtout leur transformation en actions de lutte contre le non-recours aux droits laisserait la place aux choix politiques.

À partir d'une exploration large des bases de données, il devient ainsi possible d'étudier plus précisément des segments de population et de droits. Une des méthodes d'analyse *big data* consiste ainsi en l'étude de micro-segments de population : l'exploitation de bases de données de masse permet d'analyser des populations à des niveaux plus fins en ce que si elles avaient été étudiées au travers d'un échantillonnage classique, par exemple avec des méthodes de régression logistique, elles auraient été exclues du champ car non-représentatives. Cette méthode permettrait dès lors d'accroître encore la personnalisation du droit.

Une fois la connaissance des segments de population affinée, plusieurs actions peuvent être mises en œuvre. Cela va de campagnes d'informations collectives ou personnalisées jusqu'à des liquidations automatiques, ce qui correspondrait à une forme d'internalisation de la complexité. Des études parlementaires¹² défendaient cette option en demandant la création de liens automatiques entre prestations : par exemple, un lien automatique entre demande de prime d'activité et l'obtention du RSA socle ; une automaticité du renouvellement de l'ACS pour

11 CNAF, 2017, « Le rendez-vous des droits des caisses d'Allocations familiales », L'e-ssentiel n°168 – 2017, CNAF

12 BIÉMOURET, Gisèle, COSTES, Jean-Louis, 2016, *Rapport d'information sur l'évaluation des politiques publiques en faveur de l'accès aux droits sociaux*, n°4158, 26 octobre 2016, Assemblée nationale, proposition n°14 p. 160.



les bénéficiaires du RSA ; ou encore l'octroi automatique de l'ASPA pour les bénéficiaires du RSA, de l'ASS ou de l'AAH qui atteignent l'âge requis.

En ce qu'elles touchent à certains principes de notre système de protection sociale basé sur la démarche de l'usager, ces différentes propositions de simplification amènent toutefois des interrogations.

B/ Cette évolution questionne certains principes de notre système de protection sociale et l'élaboration des politiques publiques en général

Cette voie, qui implique une proactivité plus importante des organismes et non plus des bénéficiaires potentiels, remet en cause le principe actuel de droit quérable. À cet égard, en parallèle des réflexions sur l'exploitation des données, la branche Famille s'est dotée d'un comité éthique pour garantir la prise en compte de cette dimension dans les démarches *big data*. La montée en puissance des techniques prédictives vient ainsi réinterroger ces problématiques qui renvoient à l'équilibre entre la responsabilisation des assurés et les actions d'accès aux droits. Une solution intermédiaire, qui s'inscrirait dans la continuité des actions d'accès aux droits déjà utilisées, pourrait être envisagée. Il s'agirait de cibler les rendez-vous des droits en fonction du *scoring* issus d'un *data mining* d'accès aux droits.

L'analyse de ces opportunités doit par ailleurs s'accompagner d'une réflexion sur les risques que pose la personnalisation croissante des prestations, notamment le renforcement des inégalités ou l'augmentation de la complexité administrative, y compris pour les techniciens en charge de traiter les dossiers, dans la mesure où cela peut augmenter le risque d'erreurs.

Plus globalement, l'amélioration de l'accès aux droits implique un arbitrage entre l'abondement de financement supplémentaire, la diminution des aides ou encore le renforcement du ciblage des prestations. Ces questionnements renvoient à un des paradoxes des organismes de Sécurité sociale où l'équilibre budgétaire est possible pour certaines prestations grâce à des taux de recours faibles.

Au-delà de ces enjeux, une utilisation du *big data* dans les politiques d'accès aux droits nécessite également des prérequis, en particulier eu égard aux systèmes d'information des organismes, mais aussi concernant les compétences nécessaires pour les faire évoluer.

III- Les modèles statistiques du *big data* nécessitent des prérequis en matière d'interopérabilité des bases et de gestion des compétences

III-1/ Exploiter les données de la protection sociale : une base de connaissance des publics et de l'usage des prestations partiellement exploitable en l'état

A/ Plusieurs niveaux d'analyse sont envisageables selon le périmètre des bases de données que l'on mobilise

Ce qu'offrent principalement les techniques d'analyse du *big data* est une recherche exhaustive et non plus ciblée dans les données disponibles. Or, pour le seul domaine de la protection sociale, cela peut recouvrir un champ très vaste. En effet, l'ensemble des branches et régimes de la Sécurité sociale (régime général, régimes alignés et régimes spéciaux), mais aussi Pôle emploi, les organismes de formation, les organismes complémentaires, etc. disposent de données pouvant faire l'objet de requêtes informatiques. Il apparaît dès lors souhaitable d'exploiter le maximum de données disponibles si l'on souhaite affiner les analyses effectuées.

On peut toutefois envisager plusieurs objectifs d'études en fonction de l'étendue des bases consultées :

- ▼ En premier lieu l'identification de la complétude des droits appelés pour un risque se baserait sur les données déclarées par les assurés dans les bases nationales des organismes de Sécurité sociale, au niveau de chaque branche.
- ▼ Dans un champ plus étendu, on peut envisager d'étudier les modalités de recours à l'ensemble des droits aux prestations de la Sécurité sociale, grâce au croisement de plusieurs bases de données des différents risques : cette lecture pourrait se rapprocher des dynamiques actuellement à l'œuvre autour notamment du Portail numérique des droits sociaux (PNDS).
- ▼ Enfin, dans un périmètre plus large, il s'agirait d'être capables de décrire les phénomènes de renoncement aux droits sociaux de manière globale : à partir de données disponibles auprès d'autres services publics voire des réseaux sociaux, d'identifier les usagers ne recourant pas à leurs droits et comprendre les motifs de renoncement.

B/ Un niveau fin d'analyse implique des prérequis en termes d'interopérabilité des systèmes voire de fusion des bases de données

Actuellement les recherches sont centrées sur les fichiers par risque. Plusieurs raisons expliquent ce choix stratégique. D'une part, toutes les données présentes dans les systèmes de liquidation ne sont pas forcément requêtables, du fait du développement partiel du système d'information décisionnel ou de la multiplicité des applications sur lesquelles les techniciens traitent les demandes de prestations. D'un point de vue technique, la standardisation des systèmes d'information entre les organismes de Sécurité sociale et les administrations n'est pas encore effective. Aucune norme commune n'est établie ; ce sont des systèmes qui n'identifient pas les personnes de la même manière (le numéro allocataire du foyer pour les CAF, le NIR individuel pour le RNCPS, l'INE pour le ministère de l'Éducation nationale, le numéro fiscal des impôts etc.).

D'autre part, l'ouverture des bases de données en interbranches n'est pas totalement établie à l'heure actuelle même si des projets sont en cours (par exemple le RNCPS). Sur ce point, le travail mené sur la normalisation et l'harmonisation des données en interbranche lors de la mise en œuvre de la déclaration sociale nominative (DSN) reste précurseur de par son impact.



L'ensemble des entreprises du secteur privé est, à terme, concerné et la plupart d'entre elles transmet déjà aujourd'hui des données exploitables pour l'ensemble des partenaires de la protection sociale. Cet enjeu majeur d'interconnexion entre les fichiers des organismes de protection sociale était par ailleurs inscrit dans le schéma stratégique des systèmes d'information piloté par la Direction de la Sécurité sociale (DSS). Il sera sûrement repris sur la prochaine période conventionnelle tant les avancées sur ce domaine sont encore fortement attendues.

Pour aller plus loin, l'horizon d'une base de données multilatérale de Sécurité sociale, à l'instar de la Banque carrefour de la Sécurité sociale belge, ne doit pas être totalement édulcoré. Quoiqu'il en soit, l'évolution à venir des systèmes d'information et l'utilisation accrue de technologies comme celles du *big data* nécessiteront un plan d'action cohérent et englobant l'ensemble des fonctions support des organismes, des directions des systèmes d'informations à celles de la maîtrise des risques, en passant par les ressources humaines ou l'agence comptable.

III-2/ Accompagner le *big data* : les problématiques en matière de sécurité et d'accompagnement des équipes doivent être investies en profondeur

A/ Un usage massif des données nécessite de solides garanties vis-à-vis de leur sécurité et de leur finalité

L'usage accru des données sociales va tout d'abord nécessiter un accompagnement en termes de sécurisation. De fait, plus les systèmes d'informations sont interconnectés, plus une attaque informatique à leur égard est dangereuse. Les récents cas de cybercriminalité qui ont paralysé de nombreux secteurs d'activité, et notamment le système de santé britannique en mai 2017, ont montré la vulnérabilité de systèmes d'information et de parcs informatiques insuffisamment entretenus et mis à jour.

De plus, l'utilisation de ce type de donnée sensible peut être détournée. Aujourd'hui, toute utilisation des données à des fins de contrôle par exemple doit être au préalable notifiée à la Commission Nationale de l'Informatique et des Libertés. De solides garanties doivent être conservées en la matière, d'autant plus que, comme le soulignait Pierre Colinet en 2013¹³, cette utilisation soulève des enjeux éthiques en termes de respect de la vie privée et de la liberté des allocataires.

B/ L'introduction de ces outils amène d'importants enjeux en termes de ressources humaines et de gestion des compétences

Outre la question de la sécurité des données, les enjeux financiers et humains restent conséquents. De fait, les méthodes actuelles de *data mining* impliquent des coûts d'entrée importants, en particulier si l'on cherche à les appliquer à l'accès aux droits. Il faut en effet créer des données d'apprentissage pouvant être consti-

¹³ COLLINET, Pierre, 2013, « Focus – Le *data mining* dans les CAF : une réalité, des perspectives », *Informations sociales* 2013/4 (n° 178), p. 129-132.

tuées à partir d'enquêtes auprès d'allocataires, ce afin de faire une évaluation des droits potentiels par rapport aux droits demandés. L'enjeu est de recueillir des données suffisamment importantes pour faire une base d'apprentissage performante. Il faut donc accepter d'investir, de mobiliser des ressources supplémentaires financières, humaines et techniques, afin de construire des modèles pour l'avenir, ce alors qu'il existe un risque de non atteinte des objectifs : il n'est pas possible de prédire la qualité ou la pertinence des résultats obtenus.

À plus longue échéance, et au-delà du *data mining*, les méthodes dites de « *deep learning* » permettraient de brasser l'ensemble de la base de données. Pour la branche Famille, cela représenterait l'analyse d'informations pour 32 millions d'ayants droit. Outre le coût de mise en œuvre de ces outils, les résultats obtenus grâce au *deep learning* restent difficilement interprétables, les algorithmes étant complexes à reconstituer et à traduire pour les équipes et les usagers. Pour une mise en œuvre, il conviendra donc d'abord de développer les compétences pour être en mesure d'interpréter les données puis de voir comment les restituer de manière lisible auprès des usagers « ciblés ».

Jusqu'à présent, les compétences statistiques et informatiques orientées *big data* étaient peu recherchées dans les organismes. Une montée en charge des équipes gérant les systèmes d'information est donc à prévoir, voire l'appel à des compétences extérieures. Par ailleurs, un besoin d'acculturation de l'ensemble des collaborateurs pour comprendre la méthode et les gains pouvant en découler apparaît nécessaire à terme. Ces approches étant nouvelles, il s'agit ainsi de les expliciter, de démontrer leurs intérêts, présenter leur plus-value. Pour exploiter au maximum les opportunités offertes par le *big data*, il convient par ailleurs que toutes les directions (statistiques, informatiques et métiers) se mobilisent.

Ces besoins humains ne constituent pas seulement des freins, mais aussi des opportunités : l'exploitation du *big data* offre par exemple la possibilité de mobiliser des équipes pluridisciplinaires à travers une approche transversale. De même, une mutualisation des moyens techniques et humains à un niveau interbranche et inter-régime, voire inter-organismes de protection sociale apparaît souhaitable, si ce n'est primordial.

Conclusion

Le *big data* n'en est qu'à ses débuts dans le domaine de la protection sociale. Toutefois, face à l'ampleur des chantiers à venir, il semble urgent d'investir dès aujourd'hui une réflexion poussée sur cette question et sur l'avenir de nos systèmes d'information en règle générale. Les enjeux qui accompagnent le déploiement de ces technologies impliquent une mobilisation forte et coordonnée des organismes pour éviter la déperdition des énergies. Si la lutte contre la fraude a été un point d'entrée pour ces nouveaux outils, le non-recours au droit permettrait de leur donner une légitimité nouvelle. D'autres champs restent également à explorer.