Questions
Retraite & Solidarité

Caisse des Dépôts
GROUPE

Les études
Formation professionnelle

n°32 - Décembre 2020

Les enjeux de la neutralité du moteur de recherche de Mon compte formation et les travaux entrepris pour l'améliorer

Olivier Bertrand, Willie Drouhet, Abdou Fall

La loi du 5 septembre 2018 relative à la liberté de choisir son avenir professionnel a profondément modifié la physionomie du Compte personnel de formation (CPF), le transformant en une véritable place de marché publique vouée à faire se rencontrer les offres et les demandes de formation, et centralisant les financements publics comme privés mobilisés pour solvabiliser la demande de formation.
Le lancement du portail et de l'application Mon compte formation (MCF) le 21 novembre 2019 a ainsi permis de mettre en place un véritable parcours d'achat direct de la formation professionnelle. Les conséquences en sont visibles au bout d'un an : le nombre de formations demandées et financées via MCF s'avère très dynamique malgré la crise sanitaire, et la part des formations en langues vivantes a chuté au profit notamment des formations dans les domaines des transports (dopées par les formations au permis de conduire) et du développement des capacités d'orientation, d'insertion ou de réinsertion.
Par rapport à des équivalents commerciaux, la mise en œuvre d'une place de marché publique présente des exigences spécifiques en termes de neutralité, entendue comme l'équité de traitement aussi bien des demandeurs de formation que des organismes de formation. Premièrement, cela suppose d'être en mesure de gérer un catalogue exhaustif et transparent de l'offre de formation éligible aux financements publics. Deuxièmement, il importe de mettre à disposition des utilisateurs un moteur de recherche à la fois efficace en ce qu'il permet à un demandeur de repérer les offres relatives aux formations qui l'intéressent, et neutre en ce qu'il ne privilégie pas indûment dans l'affichage des résultats les offres de certains organismes par rapport à leurs concurrents. Ce qui suppose de contrecarrer les stratégies mises en œuvre par certains acteurs pour faire remonter artificiellement leurs offres dans les classements des réponses aux requêtes adressées au moteur de recherche.
Afin d'éclairer ces questions, la Caisse des Dépôts – gestionnaire de Mon compte formation – a développé des indicateurs s'efforçant d'objectiver le degré de neutralité du moteur de recherche. S'ils ne permettent pas de qualifier dans l'absolu le niveau de neutralité du moteur, ils ont en revanche permis d'identifier les requêtes pour lesquelles la performance du moteur en termes de neutralité est la moins satisfaisante, et ainsi d'engager des travaux d'amélioration de la neutralité du moteur sur ces requêtes, travaux qui ont commencé à porter leurs fruits.

La loi du 5 septembre 2018 relative à la liberté de choisir son avenir professionnel a transformé le Compte personnel de formation (CPF) en une véritable place de marché publique vouée à faire se rencontrer les offres et les demandes de formation professionnelle, et à centraliser les droits publics et privés mobilisables pour financer une demande. Le lancement du portail et de l'application Mon compte formation (MCF), le 21 novembre 2019, a ainsi permis de mettre en place un véritable parcours d'achat direct de la formation professionnelle. L'usager peut

directement contractualiser sans intermédiaire avec un organisme de formation (OF) pour suivre la formation de son choix en mobilisant ses droits à la formation (les droits CPF, les abondements de Pôle emploi à compter de l'été 2020, et le financement direct par le salarié ainsi que les abondements des employeurs à partir de l'automne 2020).

Les caractéristiques du moteur de recherche sont donc déterminantes. Non seulement celui-ci doit permettre à l'utilisateur d'identifier les offres de formation présentes dans le catalogue qui répondent à ses besoins, c'est-à-dire répondre à une exigence de qualité et de pertinence, mais le moteur de recherche doit également traiter chaque offreur et chaque demandeur de formation de façon équitable.

Ainsi, par rapport à des équivalents commerciaux, MCF revêt des exigences en termes de neutralité que la Caisse des Dépôts, opérateur de MCF, doit garantir. Les offres de formation proposées par un OF ne doivent pas être privilégiées par rapport à ses concurrents, dans le sens où elles ne doivent pas obtenir systématiquement de meilleurs rangs de classement dans les recherches effectuées par les utilisateurs. Pour estimer cette neutralité, deux indicateurs ont été développés.

Toutefois, avant de présenter ces indicateurs de neutralité, il est nécessaire dans un premier temps de revenir rapidement sur l'importance et la diversité des offres (donc sur le catalogue) et des demandes de formation. Dans un deuxième temps, la philosophie d'un moteur de recherche et son fonctionnement sont présentés. Enfin, dans un troisième temps, les indicateurs de neutralité sont décrits et de premières mesures de ces indicateurs sont proposées.

Les offres de formation et le catalogue

Le catalogue proposé par MCF repose sur une liste de titres et diplômes officiellement reconnus nommés certifications professionnelles. Concrètement, des institutions publiques ou privées (Ministères, organismes d'enseignement - tels que les universités, les grandes écoles, lycées professionnels - mais aussi des organismes de formations) proposent des certifications reposant sur un contenu pédagogique et des évaluations, permettant de valider les compétences acquises, et valorisables sur le marché du travail. Ces institutions sont des certificateurs. Les certifications sont soumises à France compétences, qui les instruit, les valide et les diffuse.

Pour être éligible à MCF, une formation professionnelle doit être certifiante (ou préparer à l'un des blocs de compétences qui composent une certification), dans le sens où elle doit préparer au passage d'un examen validant l'acquisition des compétences professionnelles propres à la formation proposée et constituant une reconnaissance professionnelle sur le marché du travail. De plus, pour avoir le droit de proposer des formations dans MCF, les OF doivent être habilités par le Ministère du travail et respecter des conditions générales d'utilisation précises de MCF. En d'autres termes, pour saisir une offre dans le catalogue des formations, un OF doit impérativement préciser la certification professionnelle qui sera délivrée et être dûment habilité par le certificateur qui en est dépositaire. Les offres de formation précisent le contenu de la formation et ses objectifs pédagogiques. Les formations se déclinent en actions de formation (AF) accessibles lors de différentes sessions, et dont les modalités d'organisation et d'enseignement, le prix, le lieu, le nombre d'heures… doivent être spécifiés. D'un point de vue pratique, les OF alimentent le catalogue des formations en se connectant au portail de MCF. Les AF saisies seront ensuite accessibles dès le lendemain pour les utilisateurs au travers du moteur de recherche et d'une fiche descriptive dédiée.

Une forte progression des offres de formation

Lors du lancement de MCF en novembre 2019, le nombre de certifications professionnelles proposées était relativement faible (de l'ordre de 1 500, graphique 1). Ce nombre a augmenté fortement au cours de la première année de MCF, avec notamment l'arrivée progressive des formations longues et diplômantes se déroulant sur une année scolaire entière. Après une année d'existence, Mon compte formation propose des formations donnant accès à plus de 5 200 certifications (graphique 2).

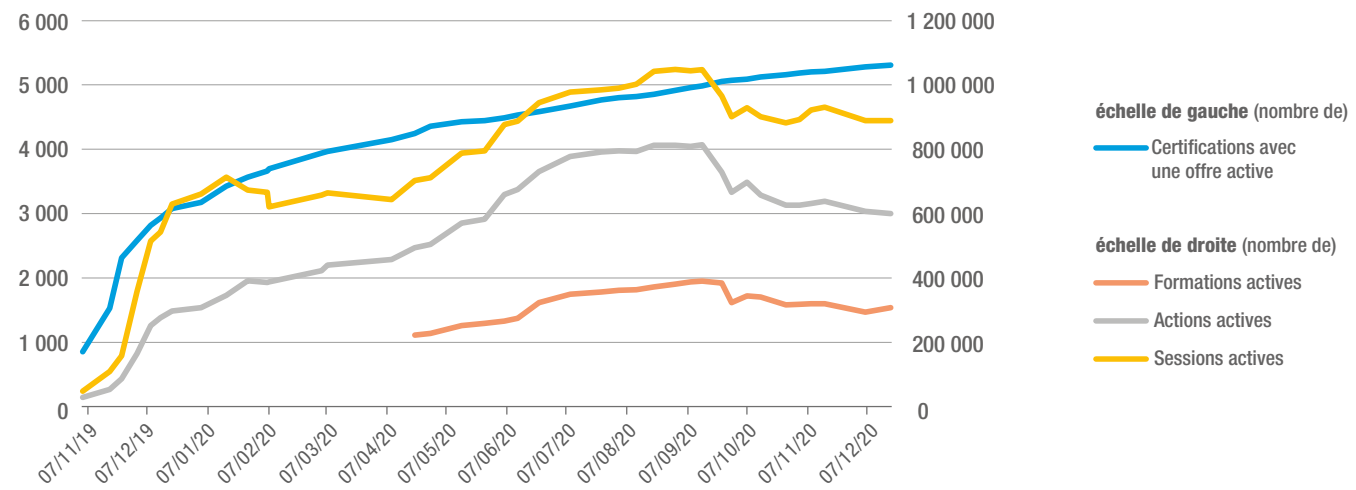
Le nombre d'OF présents tend également à augmenter régulièrement depuis 1 an même si MCF ne regroupe qu'un cinquième des organismes déclarés auprès du Ministère du travail[1]. Au total, 639 000 AF sont ainsi proposées sur MCF fin 2020, correspondant à 320 000 formations, illustrant à la fois le nombre important des formations proposées et leur grande diversité. Les formations en langues occupent une place prépondérante (tableau 1).

En première approche, la croissance assez régulière du nombre de formations et d'actions de formation présentes au catalogue suggère une évolution du catalogue allant dans le sens d'une plus grande diversité des formations proposées, et peut-être d'une plus grande pluralité d'offreurs de formation pour une certification donnée. La réalité est peut-être plus nuancée : la croissance de l'offre de formation a pu, dans certains cas, résulter en partie d'une stratégie de multiplication des offres par certains

[1] *Pour plus de détails, la liste publique des organismes de formation est disponible ici : https://www.data.gouv.fr/fr/datasets/liste-publique-des-organismes-de-formation-l-6351-7-1-du-code-du-travail/.*

Graphique 1

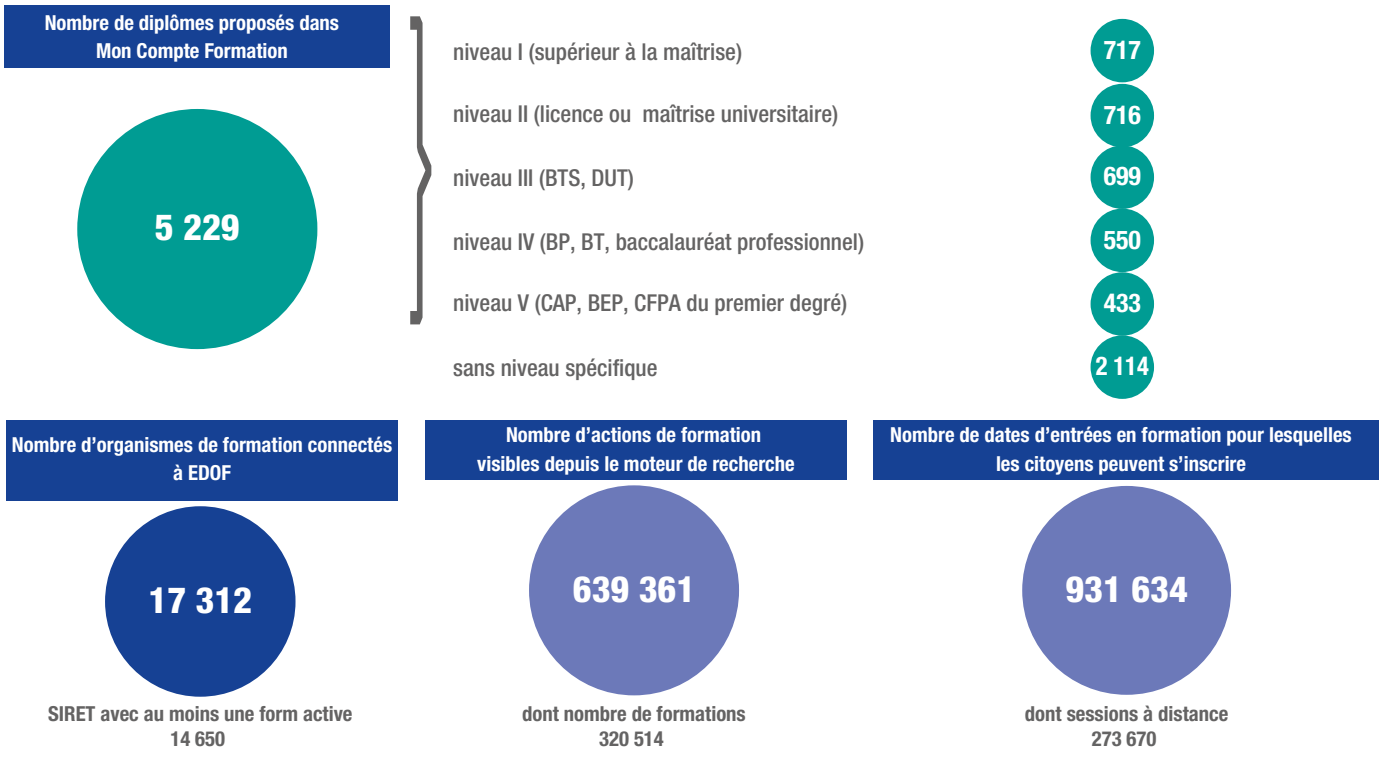
Évolution du contenu du catalogue de formation depuis novembre 2019



Source : *données Mon compte formation, Caisse des Dépôts.*

Graphique 2

Le catalogue des formations au 21/11/2020



Source : *données Mon compte formation, Caisse des Dépôts.*

Note : *EDOF désigne l'espace des organismes de formation à destination des organismes de formation sur le portail Mon compte formation.*

Tableau 1

Part des principaux domaines dans l'offre globale d'actions de formation

Langues vivantes, civilisations étrangères et régionales	47,1 %
Informatique, traitement de l'information, réseaux de transmission des données	11,3 %
Transport, manutention, magasinage	9,8 %
Développement des capacités d'orientation, d'insertion ou de réinsertion	6,2 %
Sécurité des biens et des personnes, police, surveillance	2,6 %
Spécialités pluritechnologiques, génie civil, construction, bois	2,1 %
Formations générales	1,4 %
Enseignement, formation	1,2 %
Bâtiment, finitions	1,2 %
Commerce, vente	1,1 %

Source : *données Mon compte formation au 21 novembre 2020, Caisse des Dépôts.*

Note : *dans ce tableau, les actions de formations sont classées selon la nomenclature des spécialités de formation (NSF).*

OF, visant à monopoliser les premières places des classements des recherches, réduisant d'autant la visibilité des offres proposées par leurs concurrents. Ces stratégies s'avèrent problématiques pour le fonctionnement du moteur de recherche, en ce qu'elles risquent de conduire le moteur à donner un poids démesuré dans l'affichage des premiers résultats aux organismes adoptant ce type de comportement. Ces stratégies mettent donc à l'épreuve la « neutralité » du moteur de recherche (*cf*. infra) : pour cette raison, des actions ont été entreprises afin de limiter les phénomènes de saturation du catalogue.

Une demande de formation croissante, diverse et qui évolue

Depuis l'ouverture de MCF, plus d'un million de formations ont été validées, qui représentent environ un milliard d'euros de coût pédagogique. La hausse des demandes de formation a été particulièrement forte sur les formations de « préparation au passage du permis de conduire », et dans une moindre mesure sur les actions de formation dispensées aux créateurs et repreneurs d'entreprises, ou encore sur les bilans de compétences : il en a résulté une baisse de la part des formations en langues bien que leur nombre ait légèrement augmenté. Si les 5 domaines de formations comptant le plus d'inscriptions sont restés les mêmes qu'avant le lancement de MCF (Balmat et Corazza, 2020 ; Bousquet et Jaumont, 2020), l'ordre a en revanche changé, les transports devenant le domaine attirant le plus de demandes, suivi des domaines « développement des capacités d'orientation, d'insertion ou réinsertion sociale ou professionnelle » et « langues ». Malgré leur poids prépondérant dans les actions de formation proposées dans le catalogue, les langues n'arrivent donc maintenant plus qu'en 3e position des domaines de formations réalisées, alors qu'elles étaient auparavant en tête du classement.

Ce bouleversement du classement des domaines de formations a pour corollaire une évolution substantielle du profil des demandeurs de formations : 38 % des demandes de formations traitées par MCF proviennent désormais de personnes ayant un faible niveau de qualification (CAP, BEP ou inférieur), un chiffre significativement supérieur à ce qui était observé antérieurement[2].

Le moteur de recherche et son fonctionnement

MCF met à disposition l'ensemble du catalogue de formations à n'importe quel utilisateur. Le moteur de recherche de MCF joue par conséquent un rôle crucial pour les usagers en leur permettant d'identifier les formations susceptibles de les intéresser. Il tient également une place majeure dans l'activité des organismes de formation en facilitant l'accès à leur offre de formation très abondante. Cependant, il a été constaté que certains offreurs ont recours à des techniques d'optimisation de moteur de recherche (dites « SEO », *search engine optimization*) afin de se démarquer de leurs concurrents et améliorer leur visibilité au sein de la plateforme MCF. Or la question de l'efficacité du moteur de recherche, qui regroupe l'ensemble des techniques d'optimisation permettant de classer de manière pertinente les résultats de recherche, constitue un problème classique pour les moteurs de recherche mais qui se pose dans des termes particuliers lorsque la plateforme considérée a pour finalité de mettre en œuvre une politique publique.

Il existe dans la littérature de nombreux travaux réalisés sur les techniques d'optimisation des moteurs de recherche (Akram et al, 2010). Cependant la majorité d'entre eux portent sur les moteurs de recherche web (Chartier, 2013), et peu d'études traitent la problématique du SEO à laquelle sont confrontés les moteurs de recherche d'entreprise (plateforme e-commerce ou plateforme publique). A l'origine du SEO, les spammeurs ont développé et utilisé des techniques d'optimisation des moteurs de recherche pour propulser les sites Web de rangs éloignés vers les classements plus élevés. Parmi les techniques les plus utilisées figurent le placement rémunéré (« au clic »), la soumission d'annuaires, les

[2] *Pour plus de détails, voir notamment https://www.caissedesdepots.fr/blog/article/mon-compte-formation-111.*

échanges de liens, la duplication de contenu et la génération de mots clés. Ces dernières techniques sont très courantes dans le cas des plateformes populaires (Andrieu, 2016). Concrètement, des contenus sont démultipliés, des mots-clés illimités sont placés quotidiennement par les offreurs de contenus dans les moteurs de recherche pour obtenir un classement élevé et manipuler les résultats attendus par les nombreux usagers.

Pour lutter contre le SEO et identifier la solution la plus appropriée pour éviter la saturation des offres de contenu, les opérateurs de plateformes modifient en permanence leurs algorithmes de classement. Dans un cadre commercial, les méthodes de tri par audience (« une page bien classée se justifie par le nombre de visites/transactions ») et par popularité (Google, algorithme Page Rank) (Lardy, 2000) permettent de contrer efficacement les abus et manipulations constatés. Cependant, l'usage de ces méthodes, qui favorisent les contenus dits populaires sans s'astreindre à une quelconque règle de neutralité de marché, ne saurait prévaloir dans le cadre d'un service public. D'une part parce qu'il n'est pas envisageable qu'un opérateur public favorise indûment un offreur de service au détriment d'un autre, ce qui enfreindrait les règles élémentaires de la concurrence par l'introduction de distorsions[3]. D'autre part parce qu'il ne serait pas légitime de réduire le spectre des offres proposées aux usagers en fonction de l'idée très partielle que se fait l'opérateur public de leurs centres d'intérêt.

En particulier pour la plateforme MCF, la qualité principale du moteur de recherche réside dans la neutralité de son système de classement des résultats pertinents (« ranking »). C'est un véritable facteur différenciant : si le moteur de recherche de MCF doit pouvoir trier, classer ou ordonner les informations selon certains principes de pertinence (*i.e.* répondre au mieux à la question posée par l'utilisateur), il doit aussi répondre à des objectifs d'intérêt général en étant neutre, loyal et équitable (Bertail, Bounie, Clémençon et Waelbroeck, 2019). En d'autres termes, les résultats proposés ne doivent pas opérer de discriminations et de distinctions à la fois entre les personnes en fonction de leur localisation géographique, ou encore d'attributs protégés par la loi (genre, ou encore la situation de famille), et entre les organismes de formation en fonction de leur taille, leur notoriété, etc., tout en luttant contre les manipulations (SEO) côté offreurs de formations (*cf.* encadré 1 pour une présentation simplifiée de la définition, des enjeux, des paramètres d'optimisation d'un moteur de recherche).

[3] *Réglementation P2B (Plateforme to business), applicable depuis 12/07/2020* https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32019R1150&from=FR.

Encadré 1

La définition, les enjeux et les paramètres d'un moteur de recherche

Prenons l'exemple d'une plateforme de diffusion de contenus (par exemple de contenu vidéo). Cette plateforme fait se rencontrer différents groupes d'acteurs :

- des offreurs de contenus : des personnes qui développent et déposent des contenus ;
- des demandeurs de contenus : il s'agit d'utilisateurs de la plateforme qui consomment du contenu par exemple qui regardent des vidéos ;
- potentiellement (mais pas forcément) des acheteurs d'espace publicitaire.

La question à laquelle va devoir répondre l'opérateur de la plateforme est de savoir comment il va prioriser les contenus dans l'affichage des recherches des utilisateurs. Va-t-il privilégier certains offreurs de contenus en les mettant en avant dans le classement ; va-t-il privilégier les préférences des utilisateurs en recherchant la meilleure adéquation entre leurs souhaits et les offres ; ou encore va-t-il chercher à maximiser le prix des espaces publicitaires, par exemple en orientant les classements vers les contenus les plus regardés ? Ainsi, le programme d'optimisation de l'opérateur ne sera pas le même en fonction de ce qu'il va choisir de privilégier et l'ordre des contenus soumis à l'utilisateur lors d'une recherche en sera très différent.

Dans le cadre de MCF, lorsqu'un utilisateur effectue une recherche de formation, deux possibilités se présentent. Dans le premier cas, l'utilisateur va rechercher la formation qu'il souhaite faire à partir des listes de domaines de formations et des certifications. Les AF qui sortent en réponse appartiennent alors à la liste des formations rattachées au domaine ou à la certification sélectionnés. Dans le second cas, l'utilisateur saisit un intitulé libre et les AF qui sortent en réponse appartiennent à la liste des formations dont la proximité textuelle avec le libellé saisi est importante. Dans les deux cas, ces listes d'AF sont mélangées de façon aléatoire pour chaque utilisateur et pondérées par la distance entre son lieu de résidence et le lieu de la formation, définissant ainsi des rangs différents de résultats de requête pour chaque individu.

Dans ce cadre, la Caisse des Dépôts doit se concentrer sur deux objectifs assignés au moteur de recherche MCF : proposer les réponses les plus pertinentes aux utilisateurs et assurer la neutralité entre des OF. Pour cela, les réponses aux recherches effectuées ne doivent pas systématiquement fournir de meilleurs rangs de classements à certains OF : cela suppose de tenir compte des stratégies concurrentielles de certains OF (par exemple proposer un très grand nombre d'AF relativement à ses concurrents ce qui pourrait garantir d'avoir toujours des AF bien classées). Deux indicateurs de neutralité ont été développés afin de repérer les domaines de formations et certifications dans lesquels le niveau de neutralité serait relativement plus réduit par rapport aux autres domaines.

Le moteur de recherche de MCF répond à environ un million de recherches par jour, que ce soit à travers l'application mobile ou le portail internet. Les formations proposées en résultats d'une interrogation de MCF par un usager visent à répondre à deux exigences : donner des réponses pertinentes aux mots clés saisis et ne pas privilégier dans l'affichage des réponses un OF par rapport à un autre OF dont les formations seraient tout aussi pertinentes au regard de l'interrogation soumise. Pour répondre au mieux à la question de la pertinence, le moteur de recherche s'appuie sur des algorithmes et des pondérations liées à la proximité textuelle entre le champ saisi et les intitulés et domaines des formations, à la proximité géographique entre l'utilisateur et le lieu de la formation, ainsi qu'à un aléa propre à l'utilisateur pour renforcer la neutralité : deux utilisateurs distincts localisés au même endroit et soumettant la même requête ne recevront donc pas les résultats exactement dans le même ordre de classement. Lorsque l'utilisateur effectue une recherche à partir de la liste des certifications et des domaines de formations qui lui est proposée plutôt que par saisie libre, alors seules les formations rattachées à cette certification sont affichées, toujours avec des pondérations liées à la proximité géographique et à l'aléa propre à l'utilisateur.

L'objectif premier du dispositif de formation professionnelle étant de permettre l'acquisition de compétences valorisables sur le marché du travail, le moteur de recherche a été conçu autour des référentiels de certifications professionnelles. Les domaines de formations utilisés dans MCF sont donc ceux du référentiel Formacode[4] (maintenu par Centre Inffo). Ce dernier constitue la base de l'indexation fine des certifications professionnelles et le vecteur principal du score du moteur de recherche sur lequel s'opèrent les rapprochements soit entre le champ de saisie libre et les formations proposées via les domaines de rattachement, soit directement lors d'une recherche à partir des listes de domaines et de certifications.

Le poids largement majoritaire attribué aux données des certifications apporte le premier élément constitutif de la neutralité : deux organismes de formation préparant à une même certification sont traités de façon identique, la distance géographique venant seulement pondérer le score. L'intitulé de la formation donné par l'organisme permet avec un poids relativement faible une recherche plus fine sur des mots non référencés. Afin qu'un résultat ne soit pas toujours présenté en haut de liste, un poids aléatoire de quelques pourcents est ajouté à chaque élément mélangeant ainsi ceux bénéficiant d'une pertinence proche. C'est à ce titre que certaines formations équivalentes mais parfois plus éloignées de l'utilisateur de quelques kilomètres de leurs voisines peuvent apparaître en amont de celles-ci.

Neutralité du moteur de recherche MCF : deux mesures complémentaires

La Caisse des Dépôts s'est donc engagée dans des travaux visant à opérationnaliser le concept de neutralité dans le moteur de recherche de MCF. Ceci suppose de développer des indicateurs quantifiant le degré de neutralité du moteur de recherche.

A cet effet, les résultats des réponses du moteur à différentes requêtes ont été analysés. Formellement, une requête correspond à la sollicitation du moteur de recherche par un individu donné (champ « qui »), situé dans un lieu donné (champ « où » : par exemple « Paris »), avec une question donnée (champ « quoi », correspondant à une saisie libre de mots clés ou via le choix dans les listes proposées des certifications et domaines : par exemple « Anglais »).

Pour les besoins de cette étude sur la neutralité du moteur, 254 champs « quoi » distincts tirés de l'ensemble des requêtes effectuées par les usagers début 2020 ont été testés, illustrant la grande variété des domaines de formations recherchés ainsi que la forme des requêtes formulées par les utilisateurs. Les réponses du moteur à ces 254 champs « quoi » ont été analysées pour 6 individus distincts tirés aléatoirement et deux localisations possibles pour chaque individu (Paris ou Biarritz). Soit au total 3 048 requêtes différentes soumises au moteur de recherche, chacune étant ensuite dédoublée selon que l'on traite la requête telle quelle ou que l'on neutralise l'effet de la localisation de l'utilisateur.

L'idée étant de mesurer la neutralité, la modélisation retenue consiste à comparer statistiquement le comportement du moteur de recherche MCF à celui d'un moteur de recherche neutre. Par moteur de recherche neutre, on entend simplement une approximation empirique du moteur de recherche neutre obtenue à partir de *n* rééchantillonnages aléatoires des résultats des requêtes de plusieurs individus, avec *n* tendant vers l'infini.

Concrètement, pour chaque requête, le moteur de recherche a renvoyé un nombre variable d'AF classées par ordre d'apparition et dont seules les 1 000 premières ont été retenues pour les mesures de neutralité (lorsque le nombre d'AF classées excédait 1 000). Plusieurs AF peuvent naturellement émaner d'un même OF. Si l'on fait l'hypothèse que les AF ainsi classées (au maximum de 1 000 pour chaque requête) sont toutes également pertinentes au regard de la requête considérée, il convient de vérifier, pour s'assurer de la neutralité du moteur de recherche, que l'ordre de classement des OF, observé sur

[4] *Le référentiel Formacode est maintenu par Centre Inffo ; pour plus de détails, voir https://formacode.centre-inffo.fr/.*

plusieurs requêtes (deux requêtes étant différentes dès lors que soit le champ « quoi », soit l'individu, soit la localisation de ce dernier diffèrent), n'est pas systématiquement meilleur pour certains OF et moins bon pour d'autres OF dans les résultats renvoyés par le moteur de recherche que dans le rééchantillonnage aléatoire des résultats renvoyés par le moteur.

Deux indicateurs ont été développés sur ces bases. Chacun de ces deux indicateurs évalue la neutralité du moteur pour les requêtes correspondant à un champ « quoi » et une localisation donnés.

Le premier indicateur (*I1*) se fonde sur une comparaison, OF par OF, du rang moyen dans le classement renvoyé par le moteur et dans les classements aléatoires, qui est ensuite agrégée sur l'ensemble des OF et pour les 6 individus testés[5]. L'indicateur est positif ou nul : plus il s'éloigne de 0, et moins le classement renvoyé par le moteur est aléatoire (et donc moins l'hypothèse de neutralité est vérifiée).

Le second indicateur (*I2*) teste directement l'hypothèse selon laquelle, pour une requête donnée, la distribution des rangs moyens des OF serait la même dans le classement renvoyé par le moteur et dans les classements aléatoires, puis agrège le résultat sur les 6 individus testés. L'indicateur est positif ou nul : comme le premier indicateur, plus il s'éloigne de 0, et moins le classement renvoyé par le moteur est aléatoire (et donc moins l'hypothèse de neutralité est vérifiée). L'encadré 2 présente de façon plus technique les deux indicateurs calculés et la méthode d'agrégation.

[5] *Le calcul des indicateurs sur la base de seulement 6 individus implique évidemment certaines limites en termes de performance statistique. Ce choix est motivé à ce stade par de pures contraintes techniques (temps de calcul). Dans une phase ultérieure les indicateurs seront calculés sur la base de simulations reposant sur un nombre beaucoup plus grand d'individus.*

Encadré 2

Construction des deux indicateurs permettant de mesurer le niveau de neutralité du moteur de recherche

Pour chaque requête, le classement des AF renvoyé par le moteur de recherche a été remélangé aléatoirement 100 fois, et la position des OF a été comparée dans le classement renvoyé par le moteur de recherche et dans les 100 classements aléatoires. La position d'un OF dans un classement (qu'il s'agisse du classement renvoyé par le moteur de recherche ou les classements aléatoires) a été mesurée comme le rang moyen dans le classement des AF proposées par l'OF considéré. A partir de ces rangs moyens des OF dans les différents classements, deux indicateurs ont été développés.

Dans le premier indicateur (*I1*), pour chaque requête et chaque OF, a été calculée la différence entre le rang moyen de l'OF dans le classement renvoyé par le moteur et chacun des 100 classements aléatoires, puis la moyenne de ces 100 différences. Cette moyenne a été comparée à la moyenne des différences de rang moyen de l'OF entre les 100 classements aléatoires pris deux à deux, et un indicateur de significativité statistique de l'écart entre ces deux moyennes (statistique de Student) a été produit. L'indicateur final *I1* (*champ « quoi », localisation*) correspond à la moyenne quadratique (calculée sur les 6 individus testés) de cette statistique de Student calculée sur l'ensemble des OF présents dans le classement. Ce même indicateur est aussi calculé en redressant de la distance entre les usagers et les lieux de formation, donnant finalement 4 indicateurs pour un même champ de saisie (l'indicateur pour la requête pour les deux lieux d'habitations - Paris et Biarritz - et pour ces deux sorties en neutralisant le paramètre de distance entre le lieu d'habitation et le lieu de formation*). Cet indicateur évalue la neutralité du moteur pour les requêtes correspondant à un champ de saisie et une localisation donnés.

Dans le second indicateur (*I2*), pour chaque requête, a été calculé un indicateur (de type p-value) testant l'hypothèse que la distribution des rangs moyens des OF est la même dans le classement renvoyé par le moteur et dans chacun des 100 classements aléatoires. Pour chaque requête, une distribution de 100 estimateurs p-value a donc été produite. Le même type de calcul a été effectué en testant cette fois l'identité des distributions des rangs moyens des OF entre les classements aléatoires pris deux à deux, générant une seconde distribution d'estimateurs de type p-value. L'identité de ces deux distributions a ensuite été testée, via une nouvelle statistique de p-value correspondant à une requête donnée. L'indicateur de neutralité *I2* (*champ « quoi », localisation*) correspond à l'opposé du logarithme décimal de la moyenne de cette p-value sur les 6 individus testés. Ce même indicateur est aussi calculé en redressant de la distance entre les usagers et les lieux de formation, donnant finalement 4 indicateurs pour un même champ de saisie « quoi ». Ce second indicateur évalue donc également la neutralité du moteur pour les requêtes correspondant à un champ de saisie « quoi » et une localisation donnés.

Les deux indicateurs sont positifs ou nuls : plus ils s'éloignent de 0, et moins le classement renvoyé par le moteur est aléatoire (et donc moins l'hypothèse de neutralité est vérifiée).

** En pratique, les AF proposées en réponse à une interrogation sont celles qui se déroulent dans un rayon maximum de 500 kilomètres du lieu de résidence de l'utilisateur. Paris et Biarritz étant éloignées de plus de 500 km, le fait de neutraliser l'impact de la distance dans la pondération des actions de formation n'empêche pas d'avoir des listes de réponses différentes pour un champ de saisie et un individu donnés, selon que cet individu réside à Paris ou à Biarritz.*

Des indicateurs de neutralité relativement meilleurs pour les formations de commerce et vente ainsi que pour les formations en finance, banque, assurance

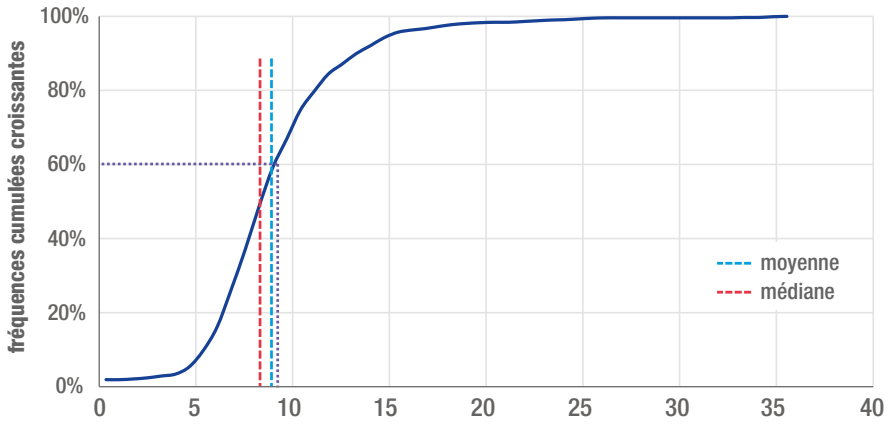
Les graphiques 3 et 4 décrivent la distribution des deux indicateurs calculés sur les 254 champs de saisie, les 2 localisations testées et la neutralisation des 2 localisations. Les deux distributions ne présentent pas la même allure. La distribution de l'indicateur 1 est unimodale : ce mode (correspondant au point d'inflexion sur le graphique 3), d'environ 7 unités d'*I1*, est proche de la moyenne et de la médiane de la distribution. En revanche la distribution de l'indicateur 2 est bimodale (avec deux points d'inflexion locaux sur le graphique 4), avec un premier mode proche de zéro et un second proche de 30 unités d'*I2*. Les valeurs moyenne et médiane de la distribution de ce second indicateur se situent entre ces deux modes.

Au-delà d'un classement relatif de neutralité, l'interprétation de ces indicateurs ne va toutefois pas de soi : si une valeur nulle correspond en tout état de cause à une situation de neutralité absolue, comment juger si une valeur de (par exemple) 10 pour l'indicateur 1 ou l'indicateur 2 est ou non acceptable du point de vue des exigences de neutralité qui s'imposent à l'opérateur public qu'est la Caisse des Dépôts ? Faute de pouvoir fournir une réponse claire à cette question délicate, nous pouvons étudier dans quelle mesure les valeurs prises par les indicateurs sont comparables ou au contraire diffèrent selon les domaines de formations considérés.

On constate en pratique des écarts substantiels entre les grands domaines de formations (tableau 2). Globalement l'indicateur *I1* donne en moyenne des résultats relativement homogènes d'un domaine à l'autre, à l'exception notable de l'informatique, domaine pour lequel la performance du moteur de

Graphique 3

Fonction de répartition de l'indicateur I1



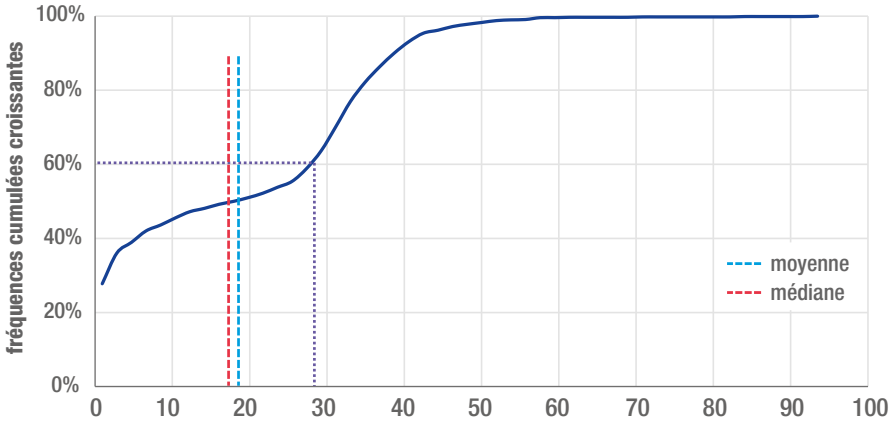
Source : *données Mon compte formation, Caisse des Dépôts.*

Note : *l'indicateur I1 est calculé sur les 3 048 requêtes testées ; 1 016 valeurs sont calculées pour l'indicateur I1, correspondant à 254 champs de saisie, 2 localisations et 2 neutralisations de la localisation.*

Lecture : *60 % des valeurs prises par l'indicateur I1 sont inférieures à 8,3.*

Graphique 4

Fonction de répartition de l'indicateur I2



Source : *données Mon compte formation, Caisse des Dépôts.*

Note : *l'indicateur I2 est calculé sur les 3 048 requêtes testées ; 1 016 valeurs sont calculées pour l'indicateur I2, correspondant à 254 champs de saisie, 2 localisations et 2 neutralisations de la localisation.*

Lecture : *60 % des valeurs prises par l'indicateur I2 sont inférieures à 26.*

recherche en termes de neutralité est sensiblement moins bonne que pour les autres domaines, et dans une certaine mesure le « développement personnel et professionnel », domaine dans lequel les champs « quoi » servant aux mesures de neutralités obtiennent des résultats plus dispersés (écart-type relativement plus élevé). L'indicateur 2 fournit des résultats plus hétérogènes entre grands domaines de formations, avec notamment une performance en termes de neutralité bien meilleure sur le domaine du commerce et de la vente que sur les autres domaines de formations. Surtout, les deux indicateurs ne fournissent pas toujours des diagnostics similaires. Par exemple, le moteur de recherche apparaît comparativement performant en termes de neutralité pour le domaine défense, prévention et sécurité avec l'indicateur 1, mais c'est l'inverse si l'on considère l'indicateur 2.

Cela étant, ces valeurs moyennes et médianes par grand domaine de formation peuvent masquer des écarts très substantiels au sein d'un domaine de formation entre deux requêtes reposant sur des champs de saisie « quoi » distincts. C'est par exemple le cas du domaine de formation des langues vivantes, le premier en termes de nombre de formations suivies en 2018 (Bousquet et Jaumont, 2020). Sur l'ensemble des champs de saisie testés, l'indicateur 1 s'établit en moyenne à 8,80 et l'indicateur 2 à 16,61. Mais, pour le champ de saisie très général « Anglais » et la localisation « Biarritz », l'indicateur 1 s'établit à 4,54 et l'indicateur 2 à 0,64. Une requête portant également sur les formations en langue anglaise, avec la même localisation mais un champ de saisie beaucoup plus précis (« Anglais-cours individuel-Débutant test Bright Anglais Level A-50h ») donne une valeur de 16,46 pour l'indicateur 1 et de 29,93 pour l'indicateur 2.

L'écart à la neutralité est donc beaucoup plus marqué avec le second champ de saisie « quoi » qu'avec le premier. Cela ne doit en fait rien au hasard : la première requête, très fréquente, a fait l'objet prioritairement de travaux visant à améliorer la performance du moteur de recherche en termes de neutralité. Les valeurs plus faibles des deux indicateurs pour le premier champ de saisie « quoi » sont donc en partie le fruit de ces travaux.

Par ailleurs, les graphiques 3 et 4 représentent la distribution des indicateurs sur 1 016 combinaisons (correspondant à 254 champs de saisie, 2 localisations et 2 neutralisations de la localisation) sans pondérer la fréquence de ces combinaisons dans les requêtes effectivement soumises au moteur.

Tableau 2

Valeurs moyennes et médianes des indicateurs de neutralité déclinées pour les principaux domaines de formations

	Indicateur 1			Indicateur 2		
	Moyenne	Écart-type	Médiane	Moyenne	Écart-type	Médiane
Tous domaines confondus	**8,78**	**3,80**	**8,16**	**17,86**	**16,31**	**17,52**
Langues	8,80	3,67	9,11	16,61	15,83	16,14
Informatique	13,25	2,66	13,97	20,77	9,55	26,31
Transports	8,66	0,63	8,46	31,14	11,86	34,50
Développement personnel et professionnel	9,93	5,78	7,21	21,47	15,96	26,36
Défense, prévention et sécurité	6,84	0,79	6,60	29,46	14,71	38,92
Échanges et gestion	8,61	1,05	8,52	34,34	4,12	34,35
Commerce, vente	6,53	0,62	6,35	5,06	5,93	2,59
Finance, banque, assurance	6,89	0,72	7,18	16,59	13,84	16,46
Ressources humaines	9,05	1,69	8,51	24,02	14,29	28,21
Ingénierie, formation et pédagogie	7,87	1,94	7,35	32,00	19,56	37,15

Source : *données Mon compte formation, Caisse des Dépôts.*

Lecture : *la moyenne des indicateurs de neutralité I1 (resp. I2) calculés pour les formations relevant du domaine des langues est de 8,80 (resp. 16,61) pour un écart-type de 3,67 (resp. 15,83) et une médiane de 9,11 (resp. 16,14).*

Or le constat d'une faible performance du moteur en termes de neutralité sur une combinaison donnée est d'autant plus problématique que le champ de saisie de la combinaison en question est fréquemment présent dans les requêtes des internautes. Le graphique 6 illustre cette question sur le premier indicateur de neutralité à partir des requêtes effectuées sur la sous-période du 21 au 26 janvier 2020 (sous-période sélectionnée au hasard pour la représentation graphique sur l'ensemble de la période de début 2020 utilisée pour l'étude).

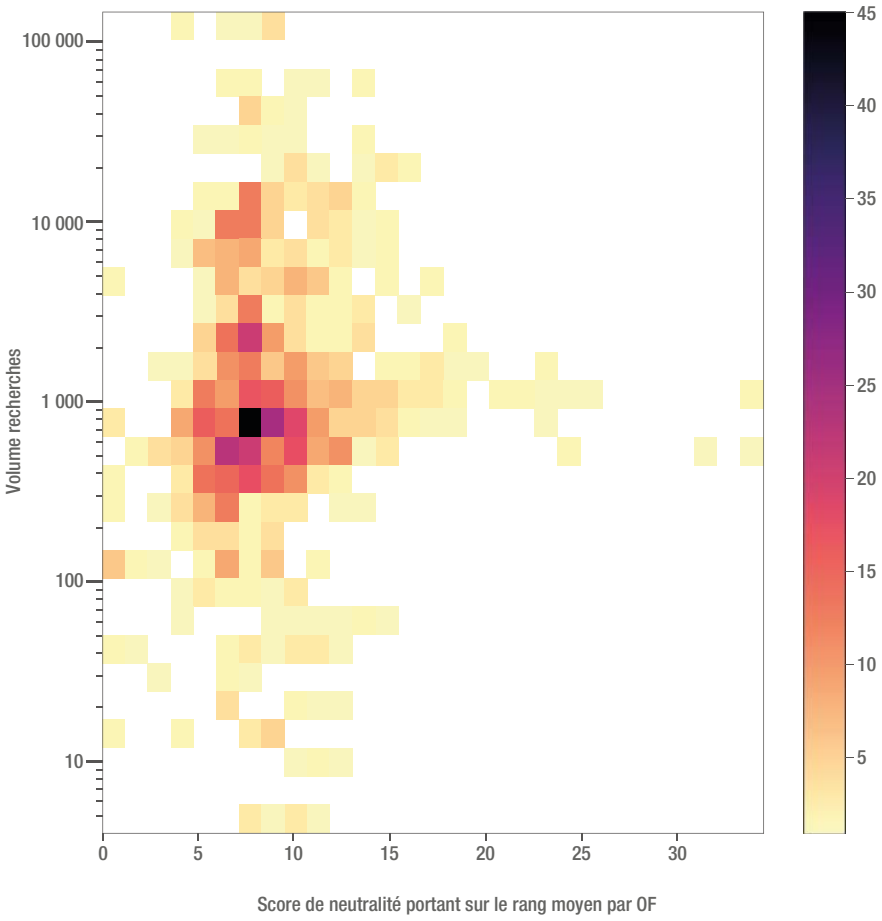
On remarque que les requêtes pour lesquelles le moteur de recherche est le moins neutre (zone du graphique 5 la plus à droite avec des valeurs de l'indicateur 1 supérieures à 20) ont fait l'objet de l'ordre de 1 000 recherches pendant la sous-période utilisée pour l'illustration. Ces requêtes sont loin d'être celles qui font l'objet le plus fréquemment de recherches (zone du graphique 5 la plus en haut). Certaines requêtes, comme « Permis B » ou « Anglais », ont ainsi fait l'objet de plus de 100 000 recherches pendant la même période, et leur indicateur de neutralité (*I1* pour le graphique 5) est meilleur puisqu'il est systématiquement inférieur à 10.

La zone la plus foncée (à gauche au centre du graphique) regroupe la majeure partie des requêtes. Le seul carré noir rassemble 45 requêtes différentes. On y retrouve des recherches comme « CACES 1 » (certificat d'aptitude à la conduite en sécurité relatif au maniement de chariots élévateurs), « CAP Petite enfance », « Chauffeur VTC », « Développement durable » ou « Langage Java » par exemple. Cette zone correspond à des volumes de recherche pour chaque champ « quoi » distinct légèrement inférieurs à 1 000 sur la période (entre 640 et 920 en ordonnée) pour un niveau de neutralité compris entre 7,1 et 8,3 (en abscisse).

Ces premiers résultats sont donc encourageants en ce qu'ils mettent en évidence que les requêtes présentant les plus mauvaises performances en termes de neutralité sont loin d'être celles qui font l'objet des recherches les plus fréquentes, et que les champs de saisie sur lesquels le moteur a été corrigé pour en améliorer la neutralité présentent effectivement de meilleurs indicateurs de neutralité. Ainsi, la partie supérieure droite du graphique 5 (correspondant à des champs de saisie « quoi » qui seraient fréquemment recherchés par les utilisateurs de MCF et qui auraient des scores indiquant une neutralité relativement faible) est vide : aucune des requêtes étudiées n'est dans cette situation. Ces illustrations permettent par ailleurs d'identifier les paramètres qui influent sur la neutralité (la localisation, le volume de recherche, la distance, le nombre d'AF par OF, la certification, le domaine de formation) et de mettre en place à court terme des améliorations concrètes pour ces requêtes moins neutres. Pour autant, les travaux visant à améliorer l'algorithme vont se poursuivre, afin d'améliorer la neutralité (l'aspect examiné ici) sans dégrader la pertinence. Cette dernière s'entend comme la capacité du moteur de recherche à renvoyer des offres de formation en adéquation avec les besoins et les recherches des usagers.

Graphique 5

Fréquence des recherches dans MCF de l'indicateur 1 de neutralité



Source : *Caisse des Dépôts, données Mon compte formation.*

Lecture : *chaque case rassemble les requêtes dont l'indicateur de neutralité 1 correspond à un certain niveau de neutralité figurant dans un intervalle donné (abscisses) et qui ont fait l'objet pendant la période du 21 au 26 janvier 2020 d'un nombre de recherches compris dans un intervalle donné (ordonnées, échelle logarithmique). Le dégradé de couleurs correspond au nombre de requêtes dans la case considérée.*

Bibliographie

Akram, M., I. Sohail, S. Hayat, M.I. Shafi et U. Saeed (2010), « Search Engine Optimization Techniques Practiced in Organizations: A Study of Four Organizations", *Journal of Computing*, Vol.2, n°6, juin.

Andrieu, O. (2016), « *Réussir son référencement web : stratégie et techniques SEO* », Eyrolles.

Balmat, C. et E. Corazza (2020), « Le compte personnel de formation en 2018 : 900 000 formations suivies par les salariés du secteur privé entre 2015 et 2018 », *Dares résultats*, février.

Bertail, P., D. Bounie, S. Clémençon et P. Waelbroeck (2019), « *Algorithmes : biais, discrimination et équité* », Rapport de recherche.

Bousquet, G. et L. Jaumont (2020), « Le compte personnel de formation pour les salariés : un retour sur les coûts de formation 2018 », *Questions retraite et solidarité-Les études,* n°29, février.

Chartier, M. (2013), « Le guide du référencement web ». *First interactive*.

Lardy J.-P. (2000), « Méthodes de tri des résultats des moteurs de recherche », *La lettre de l'URFIST*.

Les collections Questions retraite et solidarité : QRS – Les cahiers, QRS – Les études et QRS – Les brèves

QRS – Les études est une publication de la direction des retraites et de la solidarité de la Caisse des Dépôts. Elle a vocation à faire connaître les résultats des travaux d'études dans les domaines de la retraite, de la protection sociale et de la formation professionnelle. Elle est complétée par ***QRS - Les cahiers*** qui est une série de documents de travail diffusant des études approfondies et ***QRS – Les brèves*** qui propose des éclairages statistiques. L'ensemble des numéros est disponible sur le site https://retraitesolidarite.caissedesdepots.fr/ à la rubrique ***Études et événements***.

retraitesolidarite.caissedesdepots.fr Consultez les publications ou abonnez-vous à leur diffusion sur le site : retraitesolidarite.caissedesdepots.fr à la rubrique Études

Une publication de la direction des retraites et de la solidarité de la Caisse des Dépôts
Directeur de la publication : Michel Yahiel – Rédacteur en chef : Laurent Soulat
Réalisation : direction de la Communication - Retraites et Solidarité
Impression : Imprimerie CDC (75) – Dépôt légal : 4e trimestre 2020 – ISSN : 2264-0029
Contact : etudesdrs@caissedesdepots.fr – 12, avenue Pierre Mendès-France – 75914 Paris cedex 13