RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Céreq WORKING PAPER
11
2022

Analyse du rapprochement expérimental entre le fichier historique des demandeurs d'emploi et l'enquête Génération 2004

Stéphane JUGNOT
Céreq > Direction scientifique

DOCUMENT DE TRAVAIL

ENQUÊTE GÉNÉRATION 2004
MÉTHODOLOGIE D'ENQUÊTE
ENQUÊTE
COMPARAISON
CHÔMEUR

Ce document vise à favoriser la diffusion et la discussion de résultats de travaux d'études et de recherches. Il propose un état d'avancement provisoire d'une réflexion pouvant déboucher sur une publication. Les hypothèses et points de vue n'engagent pas le Céreq et sont de la responsabilité des auteurs.

Sommaire

f

Introduction **4**

1. Objectifs de l'appariement **5**

1.1. Les informations disponibles dans les fichiers de l'enquête Génération 2004 5

1.2. Les informations disponibles dans les fichiers historiques de Pôle emploi 7

1.3. Les apports possibles d'un appariement de l'enquête Génération avec le fichier historique de Pôle emploi. 12

2. L'organisation de l'appariement **15**

2.1. L'organisation théorique de l'appariement 15

2.2. La mise en œuvre effective de l'appariement 17

3. Travaux exploratoires sur les tables d'identification **24**

3.1 Le contenu de la table TABLE_IDENTIFICATION (Céreq) 24

3.2 Le contenu de la table ID_UNIQUE (Pôle emploi) 24

3.3 Le contenu de la table ID_MULT (Pôle emploi) 25

3.4 Le contenu de la table ID_MULT_BNI (Pôle emploi) 26

3.5 Le contenu de la table CORRESP_IDENT (Pôle emploi) 27

3.6 Synthèse de la mise en œuvre de l'identification dans les fichiers de Pôle emploi 27

4. Caractéristiques des personnes retrouvées de manière univoque par rapport aux autres **30**

5. Quelques analyses sur les périodes d'inscriptions à partir du rapprochement de la table DE du FHS et de l'enquête Génération **33**

Conclusion **40**

Annexe 1 – Fiche programme n°402 ..... **41**

Annexe 2 - Liste des tables composant le fichier historique administratif en 2011 ..... **42**

Annexe 3 – Extraits du « rapport intermédiaire » relatif à l'« axe 1 : dimensions techniques et méthodologiques » du « groupe de travail sur l'avenir du dispositif Génération » **44**

Annexe 4 – Lettre du Cereq à Pôle emploi (février 2012) **46**

Annexe 5 – Complément d'informations transmises par le Céreq à la CNIL (décembre 2012) **52**

Annexe 6 – Contenu du fichier intitulé « Matching Cereq » daté du 11 avril 2014, créé le 14 janvier 2014 à Pôle emploi **62**

Annexe 7 – Durée moyenne d'inscription en catégorie 1, 2 ou 3 selon diverses caractéristiques **64**

Annexe 8 – Programme SAS **69**

La collection *Working Paper* publie des textes pour engager le débat avec d'autres chercheurs.
La publication n'engage que l'auteur.

Introduction

Depuis les années 1990, le Céreq réalise régulièrement des enquêtes auprès des jeunes sortis du système éducatif pour suivre leurs premiers pas dans la vie active, dans le cadre du dispositif Génération. Ces enquêtes relèvent de la Statistique publique au sens de la loi n°51-711 du 7 juin 1951. Sans entrer dans les détails, ces enquêtes interrogent un échantillon de jeunes sortis de formation initiale au cours ou à l'issue de la même année scolaire. Ils sont interrogés sur leur parcours professionnel. Ces enquêtes portent sur tous les niveaux de sortie, des sorties de l'enseignement secondaire sans diplôme aux sorties de l'enseignement supérieur avec un doctorat. Comme tous arrivent sur le marché du travail en même temps, dans le même contexte conjoncturel, ces enquêtes permettent d'analyser facilement les différences d'insertion dans l'emploi en fonction du niveau de sortie et de la nature de la formation initiale suivie, avec un recul significatif : chaque cohorte étudiée est interrogée trois ans après la sortie du système éducatif. Pour certaines cohortes, des réinterrogations permettent de suivre les parcours quelques années supplémentaires. Jusqu'à présent, huit cohortes de sortants ont été enquêtées.

En 2012, le Céreq a lancé une réflexion interne pour rénover ce dispositif selon deux axes de travail. Le premier portait sur les dimensions techniques et méthodologique ; le second, sur l'architecture de l'enquête et le contenu du questionnaire. Au sein du premier axe, deux sujets devaient être abordés : d'une part, l'introduction de la collecte par internet, en alternative à la passation traditionnelle du questionnaire par téléphone ; d'autre part, les apports possibles d'appariements avec diverses sources administratives, parmi lesquelles des extraits des fichiers sur les demandeurs d'emploi détenus par Pôle emploi. De façon générale, le rapprochement de données d'enquêtes avec des sources administratives peut permettre de disposer d'informations substituables à celles demandées par questionnaire afin de le simplifier. Il peut aussi permettre d'améliorer le traitement de la non-réponse et de l'attrition d'une interrogation à l'autre. Il peut enfin fournir des informations plus précises et complémentaires à celles collectées. Dans le cas des données de Pôle emploi, il s'agissait aussi de pouvoir étudier les effets sur les trajectoires du recours à cet important service public de l'emploi.

Après échange et conventionnement avec Pôle emploi, le Céreq a pu obtenir des extraits des fichiers historiques des demandeurs d'emploi relatifs aux 33 655 répondants de l'enquête Génération 2004[1]. Cette cohorte a fait l'objet de trois interrogations dans le cadre du dispositif du Céreq, trois ans, cinq ans et sept ans après leur sortie de formation initiale. Le parcours des 12 365 répondants à la dernière interrogation est donc observé de 2004 à 2011. Les données de Pôle emploi, livrées en 2014, permettent de couvrir cette même période. Une première exploitation préliminaire succincte a été réalisée pour le rapport intermédiaire du groupe de travail mis en place pour réfléchir à la rénovation du dispositif Génération, produit fin 2015.

Le chantier ouvert sur les appariements de données administratives a ensuite été suspendu, la priorité ayant été donnée à la mise en place de la collecte multimode et aux réflexions sur la structure du dispositif et le contenu du questionnaire. En 2020, le chantier des appariements de données a été réouvert. Dans ce cadre, les travaux amorcés en 2014 sur l'apport des données de Pôle emploi ont été relancés (fiche programme en annexe 1). Ce document de travail en présente les conclusions.

La première partie présente les données concernées par l'appariement et ses objectifs. La seconde partie aborde sa mise en œuvre pratique telle qu'elle a pu être reconstituée sur la base des documents disponibles et de l'exploration des fichiers. La troisième partie présente les résultats des travaux exploratoires réalisés sur les tables d'identification des personnes. La quatrième partie caractérise les personnes identifiées de manière univoque dans les fichiers de Pôle emploi à partir d'informations de l'enquête Génération 2004. La cinquième partie mobilise la table des demandes d'emploi du fichier historique statistique. La conclusion aborde des pistes pour l'avenir.

[1] Jeunes sortis du système éducatif au cours ou à l'issue de l'année scolaire 2003-2004.

La collection *Working Paper* publie des textes pour engager le débat avec d'autres chercheurs.
La publication n'engage que l'auteur.

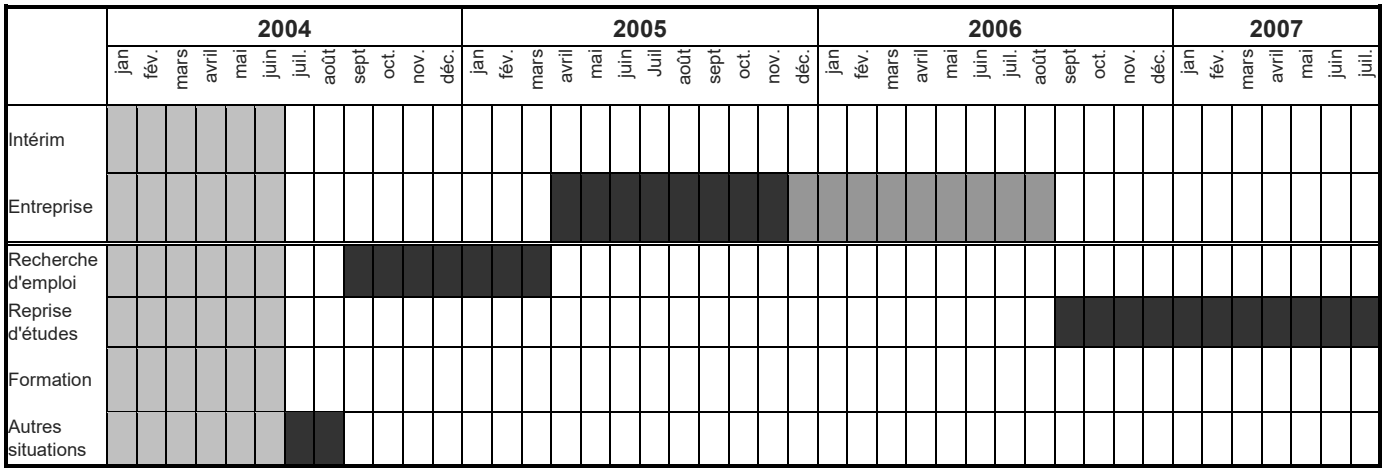
1. Objectifs de l'appariement

1.1. Les informations disponibles dans les fichiers de l'enquête Génération 2004

L'enquête Génération 2004 porte sur les 737 000 jeunes sortis de formation initiale au cours ou à l'issue de l'année scolaire 2003-2004. Une partie d'entre eux a été échantillonnée par le Céreq pour être interrogée en 2007 sur leur parcours scolaire et leurs premiers pas sur le marché du travail, de 2004 à 2007[2]. Cet échantillon global comportait un échantillon national et des échantillons complémentaires ciblés sur certaines populations pour répondre aux besoins de partenaires spécifiques. Dans l'échantillon national, 33 655 individus ont répondu à la première interrogation. Ils ont fait l'objet d'une deuxième interrogation en 2009, pour compléter le suivi de leur trajectoire professionnel jusqu'à cinq ans après leur sortie du système éducatif. 18 944 jeunes ont répondu à cette deuxième interrogation. Ils ont été interrogés à nouveau en 2011, sept ans après leur sortie du système éducatif. 12 365 jeunes ont répondu à cette troisième interrogation.

Sans entrer dans les détails, lors de la première interrogation, le questionnaire commence par aborder le parcours scolaire de l'enquêté. Il recueille ensuite son parcours professionnel au cours des trois dernières années à l'aide d'un calendrier mensuel qui permet d'identifier, à grosses mailles, les successions de séquences d'emploi et de non-emploi. Dans ce calendrier, les situations sont exclusives les unes des autres. Une seule situation est retenue chaque mois, avec un ordre de priorité, la situation d'emploi primant sur les autres. Pour les périodes non travaillées, la « recherche d'emploi » prime sur la reprise d'études, les autres formations (reconversion, stage…) et les autres situations.

Schéma 1 • Calendrier professionnel



Une fois que l'ensemble du calendrier mensuel est rempli et validé, le questionnaire aborde de façon détaillée les périodes d'emploi et certaines périodes de non-emploi. En particulier**, les situations de recherche d'emploi, de formation ou d'inactivité du passé ne font l'objet de questions complémentaires que lorsqu'elles ont duré quatre mois ou plus ou si elles correspondent à la situation de l'enquêté à la date de l'interrogation.** Ensuite, le questionnaire se termine par quelques

[2] Chacune des trois interrogations de la Génération 2004 fait l'objet d'un « Dictionnaire des codes » qui présente succinctement l'enquête et, de façon détaillée, le contenur des tables de diffusion. Les questionnaires utilisés sont également disponibles, qui prennent plutôt la forme de spécifications à l'intention du prestataire de collecte.
Voir aussi Aliaga C., Duplouy B., Jugnot S., Rouaud P., Ryk F., « Enquête Génération 2004 : méthodologie et bilan », *Net.Doc*, n° 63, Céreq, Mai 2010, 148 p. Ce document méthodologique présente les différentes étapes de production de l'enquête : la structuration du questionnaire, la construction de la base de sondage, le tirage de l'échantillon, le déroulement de la collecte et les traitements aval (codification, redressements, traitement de la non réponse).

La collection *Working Paper* publie des textes pour engager le débat avec d'autres chercheurs.
La publication n'engage que l'auteur.

modules thématiques (sur les origines sociales de l'enquêté, son sentiment de discrimination sur le marché du travail, etc.). Lors de la deuxième et de la troisième interrogation, le questionnaire part de la dernière situation décrite lors de l'interrogation précédente afin de l'actualiser. Il déroule ensuite le calendrier mensuel selon les mêmes principes généraux que lors de la première interrogation.

Dans ce questionnaire, l'inscription à l'ANPE/Pôle emploi n'est pas abordée de façon systématique. L'inscription « aux Assedics » n'est abordée que pour les séquences de recherche d'emploi, d'inactivité ou de formation en cours à la date d'enquête, quelle que soit la durée de ces séquences. Elle a aussi été abordée pour les séquences de recherche d'emploi, d'inactivité ou de formation du passé d'une durée de quatre mois ou plus lors de la première interrogation (donc uniquement pour les trois premières années suivant la sortie du système éducatif). Pour l'ensemble de ces séquences, l'information recueillie se limite à demander de façon binaire si l'enquêté était inscrit ou non « aux Assedics » pendant la période considérée, sans autre précision de dates ou de durée. Pour ces mêmes séquences, l'enquêté est interrogé pour savoir si, au cours de la période considérée, il « a été à l'ANPE [Pôle emploi] » au moins une fois[3].

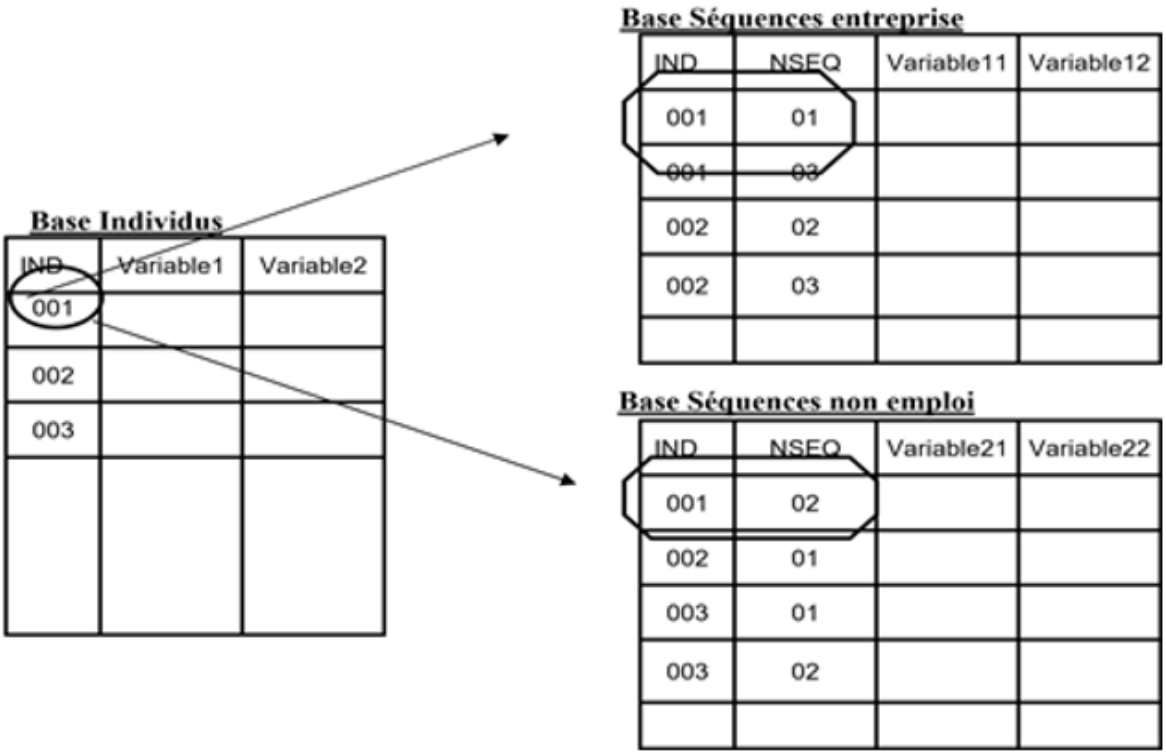
En résumé, **l'enquête Génération met l'accent sur la photographie des situations à la date d'enquête, trois ans, cinq ans et sept ans après la sortie du système éducatif, ainsi que sur la connaissance des périodes d'emploi. La connaissance de l'historique des séquences de non-emploi n'est pas approfondie. Elle permet surtout de construire des typologies de trajectoires. Le recours au service public de l'emploi n'est abordé que de façon secondaire. En particulier, l'enquête ne cherche pas et ne permet pas d'identifier l'ensemble des jeunes inscrits à l'ANPE, puis Pôle emploi, y compris à la date d'enquête, puisque pour les personnes inscrites qui exerce une activité réduite, c'est la description de l'emploi qui prévaut.** Dans leurs cas, aucune question n'est posée sur l'inscription aux Assedics ou un passage à Pôle emploi.

Concernant l'architecture des fichiers de diffusion de l'enquête, chacune des trois interrogations donne lieu à la mise à disposition d'un lot de trois fichiers :

- un fichier « Individus », avec les informations sur le parcours scolaire de l'individu, ses caractéristiques individuelles, son environnement familial ainsi que les autres informations recueillies dans les modules spécifiques. Dans ce fichier, il y a autant d'observations que d'individus, identifiés par une variable non signifiante IDENT.
- un fichier contenant les « séquences d'emploi ». Dans ce fichier, il y a autant d'observations que de séquences d'emploi. La variable IDENT permet de rattacher aux individus, les séquences qui les concernent. Pour un individu donné, un numéro d'ordre des séquences (NSEQ) permet d'ordonner les séquences de façon chronologique. Ce numéro est incrémenté en couvrant indifféremment l'ensemble des séquences d'emploi et des séquences de non-emploi successives.
- un fichier contenant les « séquences de non-emploi » décrit les séquences de non-emploi. Il comporte aussi les variables IDENT et NSEQ.

[3] La question prend la forme d'une liste de démarches possible : « Avez-vous au moins une fois : été à l'ANPE (Oui/Non), été dans un CIO (Oui/Non), fait une démarche auprès d'un employeur (Oui/Non), […] ».

Schéma 2 • Articulation des tables des fichiers de l'enquête Génération



Source : *Génération 2004, interrogation printemps 2007, dictionnaire des variables,* Céreq, juin 2008.
Note : Dans le schéma, la variable IND correspond à la variable IDENT évoquée dans le corps du texte.

1.2. Les informations disponibles dans les fichiers historiques de Pôle emploi

La présentation succincte des fichiers historiques de Pôle emploi, que nous proposons ici, s'appuie sur la documentation transmise au Céreq en 2011, à l'initialisation du projet de rapprochement expérimental[4]. Elle vaut pour les données exploitées dans le cadre de ces travaux. En revanche, elle ne tient pas compte des éventuelles évolutions qui seraient intervenues depuis dans l'organisation des systèmes d'information de Pôle emploi, ainsi que dans la production, la structuration et le contenu de ses fichiers historiques.

La gestion opérationnelle de Pôle emploi s'appuie sur un système d'information qui était déjà partagé par l'ANPE et l'Unédic[5] avant la création de Pôle emploi en décembre 2008. Dans ce système d'information, les informations évoluent en continu, au gré des nouvelles inscriptions, des radiations et de l'ensemble de l'activité de cette composante du service public de l'emploi.

À partir de ce système d'information opérationnel, un fichier AEDE[6] est produit chaque mois. Il donne une photographie du stock des demandeurs d'emploi présents à la fin du mois précédent. Ce fichier permet d'alimenter les statistiques mensuelles du marché du travail[7], publiées par la Dares et Pôle emploi. Il alimente également un fichier historique administratif (FHA), qui permet de suivre le parcours des demandeurs d'emploi au cours des dix dernières années, avec les caractéristiques du demandeur

[4] Cette partie s'appuie sur les informations contenues dans les trois documents suivants, produits par la direction des études, des statistiques et des prévisions de Pôle emploi : la version 1.4 du 10 juillet 2011 de la *Documentation générale du Fichier Historique Statistique* (29 pages) ; la version 3.0 du 12 aout 2011 *du Dictionnaire des données du Fichier historique statistique des demandeurs d'emploi* (46 pages) ; la version 3.7 du 10 février 2011 du *Dictionnaire des données du Fichier historique administratif des demandeurs d'emploi* (73 pages).
[5] Application GIDE : Gestion Informatisée des Demandeurs d'Emploi.
[6] AEDE : Amélioration des Études des Demandeurs d'Emploi.
[7] Ces statistiques sont parfois désignées par le sigle STMT.

d'emploi, ses périodes successives d'inscription, les caractéristiques des emploi qu'il a déclaré rechercher, l'exercice d'une activité réduite au cours de ses mois d'inscription, les entretiens qu'il a réalisé à Pôle emploi, les mises en relations effectuées, etc. Tous les demandeurs d'emploi ayant été inscrits au moins une fois au cours des dix dernières années sont présents dans ce fichier, en réalité structuré en plusieurs fichiers thématiques et découpé par région opérationnelle (ces fichiers sont listés dans l'annexe 2).

L'usage du fichier historique administratif est complexe. Les informations disponibles, utiles pour les travaux propres à Pôle emploi, sont nombreuses et peuvent être parfois sources de confusion pour des utilisateurs externes maîtrisant mal les processus de gestion interne. De plus, les informations enregistrées peuvent être modifiées rétroactivement en fonction des évènements de gestion survenus, mais aussi plus ponctuellement d'incidents de gestion. Par exemple, une personne qui n'a pas actualisée assez rapidement sa situation en fin de mois peut être sortie des listes et des fichiers en fin de mois mais réintégrée dès le mois suivant, sans que ce retour ne soit administrativement considéré comme une nouvelle inscription. Pour faciliter les usages statistiques, Pôle emploi produit donc chaque trimestre[8], un fichier historique statistique (FHS), à partir d'une partie des tables composant le fichier historique administratif, en opérant certains retraitements (encadré 1).

Encadré 1 • Règles de constitution du fichier historique statistique à partir du fichier historique administratif

Ces règles sont appliquées successivement :

1° Suppression des demandes dont la date d'effet d'annulation est de plus de 10 ans.

2° Suppression des doubles demandes en cours (cela arrive pour les demandeurs d'emploi qui changent d'agence au sein de la même zone Assedic).

3° Suppression des quelques demandes dont la date d'inscription est supérieure à la dernière date statistique du fichier historique (cela arrive en cas d'inscription anticipée, prenant effet après la date de fin d'observation du fichier historique produit)

4° Suppression des demandes dont la date d'effet d'inscription est strictement antérieure au 1er janvier 1976.

5° Suppression des demandes dont la différence entre la date d'effet d'annulation, quand celle-ci est renseignée, et la date d'effet d'inscription, est négative ou nulle.

6° Redressement des dates d'inscription et/ou d'annulation en cas de chevauchement de demandes. Le principe consiste à redresser la date d'annulation de la demande la plus ancienne pour la remplacer par la date d'inscription de la demande la plus récente.

7° Suppression des demandes de durée nulle ou négative à la suite des redressements de dates opérées au 6°.

8° Fusion des demandes lorsque le délai entre deux demandes consécutives est de 0, 1 ou deux jours et que leur catégorie est de même « type »[9]. La demande la plus ancienne est supprimée de la table Demandeurs/Demandes. La demande la plus récente est conservée en retenant comme motif et date d'inscription les valeurs de la demande la plus ancienne.

Dans la table des activités réduites, seules les activités réduites pour les demandes de catégorie opérationnelle 1, 2, ou 3 sont conservées. Si pour un même mois, il existe plusieurs demandes de catégorie 1, 2, ou 3, alors l'activité réduite sera affectée à la demande ayant le plus grand nombre de jours au cours de ce mois.

Source : Pôle emploi, version 1.4 du 10 juillet 2011 de *la Documentation générale du Fichier Historique Statistique* (29 pages).

[8] Le fichier historique statistique est constitué tous les trois mois : en février pour le FHS arrêté à décembre de l'année précédente ; en mai pour le FHS arrêté à mars de l'année courante ; en août pour un FHS arrêté à juin ; en novembre pour un FHS arrêté à septembre.

[9] Première possibilité pour une fusion des deux demandes : la catégorie des deux demandes est égale à 1, 2 ou 3 (personne immédiatement disponible et en recherche d'emploi, qu'il soit à temps complet ou à temps partiel, à durée indéterminée ou non). Deuxième possibilité : la catégorie des deux demandes est égale à 4 (personne non immédiatement disponible car en formation, en arrêt maladie, en congé de maternité). Troisième possibilité : la catégorie des deux demandes est égale à 5 (personne en emploi hors activité réduite et à la recherche d'un emploi). Cette typologie correspond aux catégories de demandes d'emploi utilisées dans les publications avant 1995.

Dans le fichier historique statistique, comme dans le fichier historique administratif, la table des demandes d'emploi (DE) est la table centrale. Nonobstant les retraitements opérés lors de la production du fichier historique statistique rappelés dans l'encadré 1, cette table comprend l'ensemble des demandes d'emploi successives survenues au cours des dix dernières années. Chaque observation correspond à une demande d'emploi, délimitée temporellement par une date d'inscription et une date d'annulation. Le motif d'inscription et le motif d'annulation sont renseignés selon les référentiels opérationnels de Pôle emploi. La nature et type d'emploi que le demandeur d'emploi déclare rechercher sont enregistrés dans cette table. Le métier recherché est codé selon la nomenclature du répertoire des métiers et des emplois de Pôle emploi (ROME)[10]. Le type d'emploi peut être caractérisé par la nature du contrat souhaité (contrat à durée indéterminée ou non) et la quotité de temps de travail souhaitée en distinguant juste le temps complet du temps partiel. Des caractéristiques du demandeur d'emploi sont aussi intégrés dans la table des demandes d'emploi, notamment son sexe, son âge, son niveau de formation, le fait qu'il bénéficie ou non d'une reconnaissance de handicap, le fait qu'il bénéficie ou non du RMI/RSA...

Pour comprendre la définition des types de demandes d'emploi dans la table DE du fichier historique statistique, il faut revenir à la définition des catégories opérationnelles des demandes ou demandeurs d'emploi, laquelle diffère depuis 1995 des catégories utilisées dans les publications des statistiques du marché du travail (voir tableau 1).

Lorsqu'une personne sans emploi et à la recherche d'un emploi s'inscrit à Pôle emploi, elle est inscrite en catégorie 1, 2 ou 3, selon la nature de l'emploi qu'elle indique rechercher : contrat à durée indéterminée à temps complet (catégorie 1), contrat à durée indéterminée à temps partiel (catégorie 2) ou autre type d'emploi : CDD, intérim (catégorie 3). Cette personne est ensuite soumise aux obligations des demandeurs d'emploi, notamment l'actualisation mensuelle de sa situation, qui se faisait autrefois par voie postale et désormais par voie dématérialisée.

Au cours du temps, ce demandeur d'emploi peut changer d'avis sur ce qu'il recherche. Par exemple, après avoir recherché un emploi à temps complet à durée indéterminée, il peut décider de viser un emploi en intérim. Il bascule alors de catégorie 1 en catégorie 3[11]. Dans le fichier historique administratif, ce changement se traduit par l'interruption de la demande de catégorie 1 et l'ouverture d'une nouvelle demande, en catégorie 3, soit deux observations. Dans le fichier historique statistique, ces deux demandes sont fusionnées en une seule. Dans la table DE, la catégorie associée à la demande d'emploi retenue est la plus récente, soit la catégorie 3, mais une table d'historisation de la catégorie permet de récupérer l'information qui indique que la demande d'emploi était d'abord enregistrée en catégorie 1 (table CATEGR). Ainsi, de façon générale, **tant que le demandeur d'emploi ne fait que basculer entre les catégories opérationnelles 1, 2 ou 3, il reste dans la même demande d'emploi pour la table DE du fichier historique statistique.**

Comme pour la catégorie de la demande d'emploi, d'autres informations sont susceptibles d'évoluer au cours d'une même demande d'emploi, comme le métier recherché, la situation par rapport au RMI/RSA ou la reconnaissance de situation de handicap. Chacune de ces informations font également l'objet d'une table d'historisation dédiée[12]. Dans tous les cas, la table DE ne comporte que les informations les plus récentes associées à chaque demande d'emploi.

[10] La nomenclature des métiers utilisée dans les fichiers de Pôle emploi ne s'articule pas sur la nomenclature des professions et catégories sociales utilisées dans la Statistique publique, donc dans l'enquête Génération. Le service statistique du ministère du travail a élaboré une nomenclature des familles professionnelles (FAP), qui permet d'articuler ces deux nomenclatures, le ROME et la PCS.

[11] Pour en savoir plus, voir Jugnot S. (2015), « Améliorer la publication mensuelle des statistiques du *chômage* pour faciliter le débat public. Quelques propositions », *Document de travail de l'IRES*, n°03.2015.

[12] Respectivement, les tables ROME, RMI, OBLIGEM.

Tableau 1 • Les catégories des demandes et demandeurs d'emploi

Catégorie opérationnelle (utilisée dans le fichier historique)	Exercice d'une activité réduite au cours du mois		
	Aucune	Moins de 78 heures	79 heures ou plus
Demandeur d'emploi inscrit, disponible, soumis aux obligations de recherche et d'actualisation, recherchant :			
- un CDI à temps complet	1	1	1
- un CDI à temps partiel	2	2	2
- un autre type d'emploi (CDD, intérim)	3	3	3
Autre demandeur d'emploi inscrit, non disponible immédiatement car :			
- en formation, en arrêt maladie, en congé maternité	4	4	4
- en emploi	5	5	5

Catégories "statistiques" utilisées dans la publication de la STMT de 1995 à 2009	Exercice d'une activité réduite au cours du mois		
	Aucune	Moins de 78 heures	79 heures ou plus
Demandeur d'emploi inscrit, disponible, soumis aux obligations de recherche et d'actualisation, recherchant :			
- un CDI à temps complet	1	1	6
- un CDI à temps partiel	2	2	7
- un autre type d'emploi (CDD, intérim)	3	3	8
Autre demandeur d'emploi inscrit, non disponible immédiatement car :			
- en formation, en arrêt maladie, en congé maternité	4	4	4
- en emploi	5	5	5

Catégories "statistiques" utilisées dans la publication de la STMT depuis début 2009	Exercice d'une activité réduite au cours du mois		
	Aucune	Moins de 78 heures	79 heures ou plus
Demandeur d'emploi inscrit, disponible, soumis aux obligations de recherche et d'actualisation, recherchant :			
- un CDI à temps complet	A	B	C
- un CDI à temps partiel	A	B	C
- un autre type d'emploi (CDD, intérim)	A	B	C
Autre demandeur d'emploi inscrit, non disponible immédiatement car :			
- en formation, en arrêt maladie, en congé maternité	D	D	D
- en emploi	E	E	E

Les demandeurs d'emploi doivent indiquer à l'occasion de leurs actualisations mensuelles s'ils ont effectué une activité réduite au cours du mois écoulé. Dans ce cas, ils doivent également indiquer le nombre d'heures travaillées. Cette réponse est prise en compte pour leur affectation dans les catégories « statistiques » mais ne change pas leur catégorie opérationnelle, ni dans les fichiers de gestion, ni dans le fichier historique. Une table spécifique (E0) permet d'identifier les mois d'inscription qui sont associés à l'exercice d'une activité réduite et le volume d'heures travaillées. Cette table ne donne pas d'autres informations sur les caractéristiques de l'emploi occupé. Il n'est donc pas possible de connaître la nature de l'emploi, le type de contrat de travail, etc.

En associant la table DE et la table E0, il est possible de revenir aux catégories « statistiques » telles que rappelées dans le tableau 1.

Lorsqu'il s'inscrit ou alors qu'il est déjà inscrit, le demandeur d'emploi peut ne pas être disponible immédiatement en raison d'une formation, d'un arrêt maladie ou d'un congé maternité. Il est alors exonéré de certaines obligations et inscrit dans la catégorie opérationnelle 4. Dans le fichier historique statistique, comme dans le fichier historique administratif, les périodes effectuées en catégorie 4 sont enregistrées comme des demandes d'emploi distinctes. Il en est de même pour les personnes qui s'inscrivent alors qu'elles sont encore en emploi : ces personnes sont affectées à la catégorie opérationnelle 5 et sont aussi exonérées des obligations pesant sur les demandeurs d'emploi de catégorie 1 à 3 (historiquement, bien avant l'époque d'internet, des personnes en emploi pouvaient s'inscrire pour accéder aux offres d'emploi de l'ANPE, qui en avait le monopole légal).

Généralement, les travaux mobilisant le fichier historique statistique se concentrent sur les demandes d'emploi de catégorie 1, 2 ou 3, qui correspondent aux catégories statistiques A, B et C. C'est l'agrégat le plus pertinent pour suivre les bénéficiaires du service public de Pôle emploi car tous sont soumis aux mêmes obligations de recherche et d'actualisation et peuvent bénéficier des services de Pôle emploi. Les indicateurs d'entrées et de sorties de Pôle emploi sont d'ailleurs calculés pour l'ensemble de ces catégories, sans distinction. Il en est de même pour les indicateurs relatifs au chômage de longue durée, qui est déterminé en agrègeant les durées passées dans l'ensemble de ces trois catégories.

Les fichiers historiques, administratifs et statistiques, sont produits par région opérationnelle. L'ensemble des tables composant le fichier historique est ainsi produit pour chacune des régions. L'affectation des demandes d'emploi à une région s'effectue sur la base du code département de l'agence locale pour l'emploi (ALE) d'inscription. Au sein du fichier historique administratif d'une région opérationnelle donnée, l'identifiant IDENT permet de relier les informations relatives à un même demandeur d'emploi. Cet identifiant est partiellement signifiant puisqu'il comporte le code de l'Assédic concernée, sachant que les régions opérationnelles ne sont pas identiques aux territoires que les Assédics couvraient[13].

En cas de changement d'Assédic, le demandeur d'emploi change d'identifiant. Depuis mai 2003, une table spécifique du fichier historique administratif (TRANSFER) conserve l'historique des changements d'Assédic des demandeurs d'emploi, avec l'identifiant de départ et l'identifiant d'arrivée.

Au sein du fichier historique statistique, l'identifiant IDENT est remplacé par un autre identifiant (IDX), totalement non signifiant. Il s'agit d'un numéro d'ordre incrémenté région par région. Le même identifiant peut donc être utilisé dans plusieurs fichiers régionaux et désigner des demandeurs d'emploi différents. Par construction, il y a par exemple autant de demandeur d'emploi ayant IDX=1 que de régions. L'incrémentation est renouvelée à chaque production du fichier historique statistique. Un même demandeur d'emploi change donc d'identifiant d'une version à l'autre du fichier historique statistique. Comme l'incrémentation ne mobilise pas la table TRANSFER évoquée plus haut, il n'y a pas de continuité dans le suivi d'un demandeur d'emploi qui change d'Assedic, même s'il ne change pas de région. En changeant d'Assedic, il change d'identifiant. Sa demande d'emploi du territoire de départ est interrompue et une nouvelle demande d'emploi débute dans sa zone d'arrivée. Dans ce cas, pour le fichier historique statistique, il y a deux demandes d'emploi associée à deux demandeurs d'emploi différents. D'après la documentation de Pôle emploi de 2011, environ 5 à 6 % des demandeurs auraient eu une demande tronquée suite à un changement de zone Assedic.

À côté du fichier historique statistique produit trimestriellement, Pôle emploi produit un « super fichier historique statistique », qui cette fois n'est pas produit par région mais au niveau de la France entière. Dans ce super FHS, le demandeur d'emploi est identifiée par son NIR anonymisé (IDX prend la valeur du NIR anonymisé). Dans ce cas, le demandeur d'emploi peut être suivi lorsqu'il change d'Assédic.

[13] Par exemple, l'Ile-de-France et Provence-Alpes-Côte d'Azur disposaient chacune d'une seule région opérationnelle de l'ANPE mais regroupaient plusieurs Assédics.

1.3. Les apports possibles d'un appariement de l'enquête Génération avec le fichier historique de Pôle emploi.

De façon synthétique, trois façons de mesurer le chômage peuvent être distinguées[14] :

- *le chômage administratif :* c'est le fait d'être demandeur d'emploi inscrit auprès de Pôle emploi. Ces données administratives permettent de disposer de chiffres rapidement et à des niveaux géographiques fins mais cette quantification ne porte que sur les bénéficiaires de ce service public de l'emploi. Or tous les chômeurs ne s'inscrivent pas. Par nature, la mesure est sensible aux évolutions des règles administratives relatives aux bénéficiaires de ce service public de l'emploi, ainsi qu'aux règles de gestion internes à Pöle emploi[15] ;
- *le chômage au sens du bureau international du travail (BIT) :* c'est une mesure statistique qui décline une définition internationalement harmonisée, selon laquelle le chômeur est une personne sans emploi, qui en recherche un et qui est disponible pour travailler. Seule cette mesure fait référence pour les statisticiens. Elle ne peut s'obtenir que par une enquête représentative. En France, c'est le rôle de l'enquête Emploi de l'INSEE. La définition du BIT implique de poser des conventions précises, qui se traduisent par une batterie de questions permettant de définir qui est une personne sans emploi, qui en recherche un et qui est disponible pour travailler. Ces conventions peuvent varier d'un pays à l'autre ou au cours du temps. Résultant d'une enquête par échantillon, la mesure du chômage au sens du BIT ne peut pas se décliner à des niveaux géographiques fins sans multiplier les hypothèses, donc les approximations ;
- *le chômage déclaratif* **:** c'est la mesure que donne le recensement de la population et la plupart des enquêtes auprès de personnes, dont les enquêtes Génération. Elle consiste à s'appuyer sur la simple déclaration de situation d'activité des enquêtés, car poser l'ensemble des questions utilisées dans l'enquête Emploi pour se caler sur sa définition, prendrait trop de place dans enquêtes dont l'objectif principal n'est pas de mesurer le chômage au sens du BIT.

Les écarts entre les trois principales mesures du chômage sont documentés depuis plusieurs décennies. En particulier, malgré la proximité de leur nombre et l'illusion que ces deux mesures seraient conceptuellement proches, les écarts sont importants entre les chômeurs au sens du BIT et les demandeurs d'emploi que Pôle emploi comptabilise dans la catégorie « statistique » des demandeurs d'emploi de catégorie A, même si une confusion s'est installée entre les deux notions depuis la mise en œuvre des nouvelles catégories statistiques sur les inscrits à Pole emploi en 2009. Ainsi, par exemple, selon une étude publiée par la Dares en 2019, à partir d'un rapprochement des réponses individuelles à l'enquête Emploi et des fichiers de Pôle emploi[16] :

- 56 % des demandeurs d'emploi inscrits en catégorie A en 2017 étaient au chômage au sens du BIT ; 20 % étaient dans le halo autour du chômage ; 16 % étaient inactifs et hors du halo et 9 % étaient en emploi.
- Inversement, 66% des chômeurs au sens du BIT étaient inscrits en catégorie A, 11 % étaient inscrits en catégorie B ou D (donc avec un emploi) et 22 % n'étaient pas inscrits.

Les données détaillées utilisées pour cette étude ne sont pas accessibles aux chercheurs. Il est donc difficile d'affiner ces résultats pour se centrer sur les plus jeunes et se rapprocher de la cible de l'enquête Génération, ou pour délimiter une catégorie de chômeur déclaratif plus souple que celle utilisée pour le chômage BIT. Cependant, certains des résultats globaux de l'étude apportent des enseignements utiles, en particulier ceux repris dans le tableau 2 ci-dessous. Ce tableau compare les situations identifiées dans l'enquête Emploi à celles enregistrées dans les fichiers de Pôle emploi par tranches d'âges, là encore de façon malheuresement insuffisamment détaillée puisque la catégorie « Autres cas » mélange les chômeurs au sens du BIT inscrits à Pôle emploi dans d'autres catégories que la catégorie A et les personnes en emploi selon l'enquête mais inscrites en catégorie A. Ce tableau confirme que les

[14] Sur la mise en place de ces trois mesures, voir Jugnot S., « Les mesures du chômage » in *Regards croisés sur l'économie*, n°2013/1 (n° 13), pages 31 à 44 (plusieurs articles de comparaison de sources sont cités).
[15] Voir Jugnot S. (2015), « Améliorer la publication mensuelle des statistiques du « chômage » pour faciliter le débat public. Quelques propositions », *Document de travail*, n°03.2015, Ires.
[16] Hameau A., Dix C., Coder Yohan et alii (2019), « Appariement entre l'enquête Emploi et le fichier Historique de Pôle emploi sur la période 2012-2017. Méthode et premiers résultats », *Document d'Etudes*, n°233, Dares.

jeunes chômeurs sont beaucoup plus fréquemment non inscrits à Pôle emploi que les plus âgés[17]. Il montre aussi qu'au moins un tiers des inscrits en catégorie A ne sont pas chômeurs au sens du BIT.

Tableau 2 • Situation comparée dans l'enquête Emploi et dans les fichiers de Pôle emploi, pour l'année 2017

Situation dans l'enquête Emploi	Situation dans les fichiers de Pôle emploi	15-29 ans	30-49 ans	50-64 ans
Inactif hors halo	Non-inscrits	80,1	81,7	86,9
En emploi	Non-inscrits			
En emploi	Catégorie B, C, E	4,3	6,1	4
Chômeurs	Non-inscrits	3,8	1,6	0,8
Chômeurs	Catégorie A	4,1	4,3	3
Inactif hors halo	Catégorie A	1	0,8	1,6
Dans le halo	Catégorie A	1,3	1,5	1,2
Dans le halo	Non-inscrits	2,8	1,7	1,2
Autres cas (dont les "En emploi / Catégorie A" et "Chômeurs / Catégorie B, C, D ou E")		2,6	2,3	1,3
Ensemble		100	100	100

Source : graphique 3.5 (page 44) de Hameau A., Dix C., Coder Yohan et alii (2019), « Appariement entre l'enquête Emploi et le fichier Historique de Pôle emploi sur la période 2012-2017. Méthode et premiers résultats », *Document d'Etudes*, n°233, Dares.

Ces résultats, combinés à la structure et au contenu du questionnaire de l'enquête Génération, conduisent aux principaux constats suivants :

1° Le questionnaire de l'enquête Génération n'aborde l'inscription à Pôle emploi que de façon très parcellaire, notamment sur les périodes du passé. Sur ce sujet, il y a donc peu à apprendre d'une comparaison entre les informations d'inscription recueillies dans l'enquête Génération et les données administratives. De ce fait, l'intégration des données de Pôle emploi constituerait un enrichissement certain, sous réserve que les opérations d'identification des personnes dans les deux sources permettent d'effectuer leur rapprochement dans de bonnes conditions, sans trop de perte, ni trop de biais.

2° Pour les mêmes raisons et dans les mêmes conditions, l'apport d'informations sur la nature de l'accompagnement proposé par Pôle emploi constituerait aussi un enrichissement. Cet apport n'a cependant d'utilité que si des études d'évaluation des effets de cet accompagnement sont envisagées. Dans ce cas, l'intérêt doit être nuancé par le fait que d'autres acteurs interviennent, à commencer par les missions locales, dont le rôle est centré sur l'insertion professionnelle et sociale des jeunes de moins de 25 ans. Sur ces sujets, il faudrait donc pouvoir tenir compte aussi de l'accompagnement proposé par les missions locales pour exploiter de façon pertinente les informations de Pôle emploi associées à l'enquête Génération.

3° La proportion importante des jeunes chômeurs qui ne s'inscrivent pas à Pôle emploi interdit d'envisager de substituer les données administratives aux données d'enquête pour calculer les indicateurs de situation sur le marché du travail à la date d'enquête, indicateurs qui constituent les résultats centraux et les plus utilisées des enquêtes Génération.

[17] Cette proportion aurait peu évolué entre 2013 et 2017. Le document de travail de la Dares présente également les données pour 2013 dans son graphique 3.5. Les chômeurs non inscrits représentaient alors 4,3 % des 15-29 ans et les chômeurs inscrits en catégorie A, 4,3 %. La catégorie fourre-tout « autres » rassemblaient 2,5 % des 15-29 ans.

La collection *Working Paper* publie des textes pour engager le débat avec d'autres chercheurs.
La publication n'engage que l'auteur.

4° En revanche, le constat précédent ne vaut pas forcément pour le calendrier professionnel. Dans l'enquête Génération, celui-ci n'est recueilli qu'à grosse maille, avec une priorité à l'emploi et sans doute des biais de mémoire, voire de collecte, pour les trajectoires les plus heurtées. Ce calendrier n'est pas central pour la plupart des exploitations réalisées. Son utilisation la plus fréquente est indirecte *via* l'utilisation de la typologie des trajectoires (variable TYPTRAJ). De ce fait, il pourrait être intéressant de regarder si une typologie élaborée à partir des données administratives de Pôle emploi et des périodes d'emploi du calendrier professionnel permettrait d'aboutir à des analyses convergentes avec la typologie « Céreq » en termes de différences de parcours selon la formation initiale suivie. Deux types d'informations pourraient être plus particulièrement mobilisés dans les données de Pôle emploi : d'une part, la durée et la récurrence des périodes d'inscription sans activité réduite ; d'autre part, l'importance et la récurrence de l'activité réduite.

Outre ces points relatifs aux informations exploitables, les données de Pôle emploi pourraient, peut-être, permettre d'améliorer le traitement de la non-réponse ou de l'attrition en fournissant des informations disponibles sur les répondants et les non-répondants.

Seule une partie des apports possible d'un rapprochement des fichiers historiques de Pôle emploi avec l'enquête Génération a été étudiée. Les travaux exploratoires réalisés se sont d'abord centrés sur la qualité de l'appariement : combien de personnes sont retrouvées dans les fichiers de Pole Emploi ? Quelle proportion selon les niveaux de sortie ? Le sont-elles de façon univoque ou plus complexe ?

Une exploitation de la table des demandes d'emploi du fichier historique statistique a ensuite été réalisée pour examiner la cohérence des informations issues des données administratives avec les résultats de l'enquête Génération, concernant les difficultés d'accès au marché du travail selon les niveaux de sortie. Pour les raisons expliquées plus haut, il n'y aurait pas eu de sens à chercher une cohérence entre les réponses individuelles des jeunes dans l'enquête Génération et les réalités enregistrées dans les données administratives de Pole Emploi. En revanche, il était important de s'assurer que les messages globaux sur les difficultés différentielles d'accès au marché du travail étaient cohérents entre les deux sources.

2. L'organisation de l'appariement

Le projet de rapprochement expérimental de l'enquête Génération 2004 avec le fichier historique statistique a été initié en 2012 dans le cadre d'une réflexion plus large portant sur la rénovation des enquêtes Génération. Ce projet a fait l'objet :

- d'une saisine du comité du secret le 5 novembre 2012, examinée le 6 décembre 2012,
- d'une déclaration dématérialisée auprès de la Cnil le 4 décembre 2012 (sous le numéro 1636140), complétée d'un dossier transmis par lettre du directeur du Céreq[18],
- d'une convention entre le Céreq et Pôle emploi, signée le 26 mai 2014.

Peu de documents sont aujourd'hui disponibles pour expliquer la genèse de l'opération, la nature et la motivation des choix faits et les opérations concrètes réalisées. De plus, une partie des personnes en charge de ce dossier à l'époque a quitté le Céreq depuis plusieurs années. Cette partie s'appuie donc essentiellement sur l'examen des fichiers et documents disponibles. Certains de ces documents figurent en annexes :

- la partie consacrée à cet appariement dans un « rapport intermédiaire » du « groupe de travail sur l'avenir du dispositif Génération » (annexe 3),
- la lettre formelle adressée par le Céreq à Pôle emploi en février 2012 pour initier le projet[19] (annexe 4),
- les informations transmises à la Cnil en décembre 2012 en complément du formulaire standard (annexe 5),
- le contenu d'un fichier sans en tête intitulé « Matching Céreq » et daté du 11 avril 2014, qui semble émaner de Pôle emploi et résumer succinctement ses opérations d'identification (annexe 6).

2.1. L'organisation théorique de l'appariement

Selon les documents disponibles, l'appariement devait se faire en six étapes :

Étape 1 - Création par le Céreq d'une table d'identification contenant les informations suivantes : nom, prénoms, date de naissance, identifiant non signifiant différent de l'identifiant non signifiant utilisé dans les fichiers de résultats des enquêtes de la Génération 2004.

Étape 2 – Transmission à Pôle emploi de la table d'identification selon les modalités prescrites par Pôle emploi et validées par la CNIL.

Étape 3a[20] – Recherche des identifiants Pôle emploi correspondant aux identités transmises par le Céreq et traitement des doublons.

Étape 3b – Récupération par Pôle emploi des informations destinées à enrichir les fichiers de résultats des enquêtes de la Génération 2004 pour les individus retrouvés (voir encadré 2).

Étape 4 – Transmission au Céreq des tables issues de Pôle emploi indexées sur l'identifiant non signifiant du Céreq, selon des modalités validées par la CNIL.

Étape 5 – Création par le Céreq d'une table « Pôle emploi » substituant à l'identifiant non signifiant utilisé pour les échanges avec Pôle emploi l'identifiant non signifiant utilisé dans les fichiers de résultats des enquêtes de la Génération 2004.

[18] Un projet de lettre daté du 30 novembre 2012 a été conservé. Selon les attendus de la convention passée entre le Céreq et Pôle emploi, l'autorisation de la Cnil aurait été donnée le 3 décembre 2013.
[19] Une réponse positive formelle a été adressée au Céreq par Pôle emploi, par lettre de Stéphane Ducatez, sous-directeur de l'évaluation et des prévisions, en date du 20 mars 2012
[20] Les documents produits ne distinguent pas les étapes 3a et 3b mais il nous semble nécessaire de le faire pour plus de clareté sur le processus.

La collection *Working Paper* publie des textes pour engager le débat avec d'autres chercheurs.
La publication n'engage que l'auteur.

Étape 6 – Destruction par Pôle emploi de la table d'identification et de la table transmise en retour au Céreq, au plus tard un mois après le rapprochement. Destruction de la copie de la table d'identification par le Céreq au plus tard un mois après le rapprochement.

Encadré 2 • Les tables de Pôle emploi demandées par le Céreq dans le cadre du projet

Le projet de rapprochement initié en 2012 ne se limitait pas aux quelques tables essentielles du fichier historique statistique nécessaires pour réaliser les travaux expérimentaux tels que nous les proposons dans la partie 1.4. Il s'inscrivait dans une perspective plus large en cherchant à couvrir tous les sujets potentiellement intéressants à des fins d'études et de recherche. Il portait donc, non seulement sur le fichier historique statistique, mais aussi sur certaines des tables du fichier historique administratif habituellement non intégrées au fichier historique statistique. Ainsi, les tables suivantes étaient demandées :

Extraction du fichier historique statistique (FHS) :

- Table DE : Table des demandeurs – demandes (sauf variable DEPCOM)
- Table D2 : table des indemnisations
- Table E0 : Table des activités réduites (rattachées à des demandes en catégorie 123678 uniquement)
- Table PAP : Table des entretiens PAP (depuis juillet 2001 et uniquement pour des entretiens rattachés à des demandes 123678)
- Table PARCOURS : Tables des parcours des demandeurs d'emploi
- Table STATDE : Table de statistiques individuelles sur le demandeur (sauf variable REGION)
- Table RSA : table des droits RSA (sauf variable REGION)
- Tables des variables historicisées (sauf variables REGION et NUMZUS[21])

Extraction du fichier historique administrative (FHA) :

- Table E3 : table de toutes les actions (conseillées, réalisées ou non)
- Table M0 : table des mises en relations (positives ou non)
- Table P2 : table des formations (sauf variable SIRET)

Les documents écrits disponibles passent sous silence plusieurs aspects pratiques, importants pour comprendre la mise en œuvre concrète de l'appariement. Nous verrons par la suite comment certains d'entre eux semblent avoir été traités. D'autres restent des points aveugles à ce stade.

P1. Dans l'enquête Génération, deux dates de naissance sont parfois disponibles, l'une provenant des fichiers administratifs ayant servi à constituer la base de sondage, l'autre collectée lors de la première interrogation, lors du questionnement introductif visant à s'assurer de la bonne identité de la personne contactée. Généralement, les deux dates sont identiques ou la date collectée complète une information manquante dans les fichiers administratifs. Dans quelques cas, la date de naissance collectée diffère cependant de la date disponible dans la base de sondage. Il faut donc prévoir et gérer cette situation.

P2. Les données identifiantes mobilisées pour procéder à l'identification des personnes et au rapprochement des données ne correspondent qu'à une partie tronquée de l'état-civil des personnes : le nom, le prénom, le mois de naissance et l'année de naissance. En revanche, le sexe, le jour de naissance et le lieu de naissance n'ont pas été utilisés. Le jour et le lieu de naissance ne sont pas collectés dans l'enquête Génération. Le sexe était disponible dans les deux fichiers mais l'information n'a pas été utilisée pour des raisons inconnues. Un fichier Excel se présentant comme les résultats d'un « test » réalisé sur 2 592 individus pourrait suggérer que cette information n'a pas été jugée cruciale pour l'identification même si la nature du test réalisé demeure inconnue (voir les « résultats » dans le tableau 3). Enfin, il faut noter que les patronymes enregistrés par Pôle emploi sont les patronymes donnés à la naissance alors que le mode de collecte de l'information conduirait plutôt aux patronymes d'usage pour l'enquête Génération. Il n'est pas impossible que cet écart conduise à une capacité à l'identification des personnes biaisée, même si le risque est limité par la tranche d'âges des

[21] L'information sur la résidence en ZUS est jugée non fiable par Pôle emploi, car alors souvent non renseignée.

personnes relevant du champ de l'enquête Génération (dans le cas où le Céreq souhaiterait procéder à des rapprochements plus systématiques, il pourrait être utile de s'assurer que l'enquête collecte les informations les plus efficaces pour procéder à l'identification des personnes dans les fichiers administratifs désirés).

Tableau 3 • Les résultats du « test » (fichier « Base31mai.xls »)

	Nombre de correspondances SID - Pôle emploi (âge, sexe, an et mois de naissance obtenus à partir du NIR)						
Correspondances selon les critères :	0	1	x (multiple)	% corresp 1-1	% corresp 1-x	% corresp 1-x ou 1-1	Total
Nom, Prénom, Sexe, Année et mois de naissance	538	2054	0	79,2%	0,0%	79,2%	2592
Nom, Prénom, Sexe, Année de naissance	433	1765	394	68,1%	15,2%	83,3%	2592
Nom, Prénom, Sexe	313	1140	1139	44,0%	43,9%	87,9%	2592
Nom, Prénom	303	1143	1146	44,1%	44,2%	88,3%	2592
Nom, Prénom, Année et mois de naissance	528	2059	5	79,4%	0,2%	79,6%	2592

P3. Aucune information n'a été retrouvée sur la méthode précise d'identification des personnes dans les fichiers de Pôle emploi, en particulier pour savoir comment sont pris en compte ou non les risques d'erreurs d'orthographe sur les patronymes ou les cas de prénoms multiples. Comme nous le verrons, il est question de prénoms « ressemblants », sans que cette notion soit définie. C'est un point qu'il conviendrait de conserver dans la documentation si une nouvelle opération devait être réalisée.

2.2. La mise en œuvre effective de l'appariement

La description de la mise en œuvre effective de l'appariement s'appuie sur les étapes théoriques décrites dans la partie précédente.

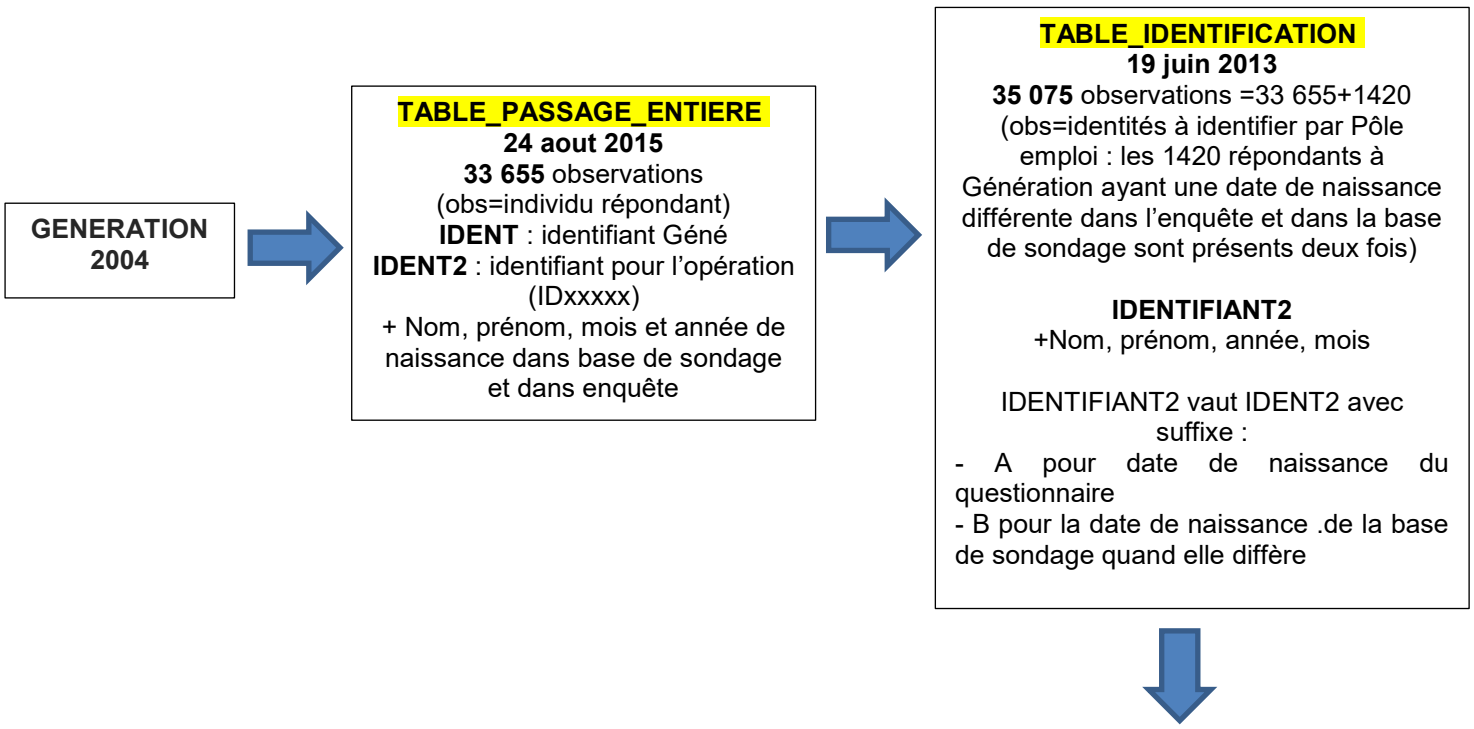
Étape 1 : production par le Céreq de la table des individus à identifier

À partir de l'enquête Génération 2004, le Céreq a produit la **table TABLE_PASSAGE_ENTIERE**, qui comprend 33 655 observations, soit une observation par individu ayant répondu à la première interrogation de l'enquête et appartenant à l'échantillon national concerné par la deuxième interrogation.

Cette table comprend les informations suivantes :
- le nom (variable NOM),
- les prénoms (variable PREN),
- le mois et l'année de naissance dans la base de sondage[22] (variables MOISA et ANNAI),
- le mois et l'année de naissance dans le questionnaire (Q2 et Q3),
- l'identifiant non signifiant dans l'enquête Génération (IDENT),
- l'identifiant non signifiant créé pour l'opération de rapprochement (IDENT2)
- des variables construites (pour séparer les prénoms, repérer l'existence de caractères spéciaux dans la variable NOM et PREN, pour repérer les écarts possibles sur le mois ou l'année de naissance).

À partir de cette table, la table TABLE_IDENTIFICATION a été créée, à destination de Pôle emploi. Celle-ci ne comporte que 5 variables :
- le nom (variable NOM),
- les prénoms (variable PREN),
- le mois et l'année de naissance (apriori tantôt alimenté par MOISA et ANNAI, tantôt par Q2 et Q3),
- l'identifiant non signifiant destiné à Pôle emploi (variable IDENTIFIANT2).

[22] Hypothèse de notre part.



TRAITEMENTS D'DENTIFICATION DANS LES FICHIERS PÔLE EMPLOI

Étape de traitement préliminaire : Suppression des espaces et des tirets dans les noms et prénoms. Année et mois de naissance mis en format numérique.

Création du fichier Individus : à partir de l'axe Individus de SID, filtrage sur les années de naissance des individus pour ne retenir que les nés sur la période 1969-1989 afin de faciliter le matching. Récupération de la date de mise à jour, de l'identifiant de l'individu (**IDTIDV**), l'identifiant BNI de l'individu (**IDENT_BNI**), son nom de naissance, son prénom de naissance, son année et mois de naissance. Suppression des espaces et des tirets dans les noms et prénoms.

Matching de cette table avec celle transmis par le CEREQ :
Étape 1 : jointure des deux tables à partir du **nom, de l'année et du mois de naissance**.
Étape 2 : conservation des individus dont le **prénom Cereq « ressemble » au prénom Pôle emploi**.
Étape 3 : pour éliminer les doublons, utilisation de la notion d'identifiant unique rattachée au numéro BNI ainsi que la date de mise à jour. Plus précisément :

⇒ ***1°*** *Si pour un individu « Cereq », on a un seul écho* dans la table Pôle emploi sur la base du même nom + même mois et année de naissance + prénom « ressemblant »

⇒ table **ID_UNIQUE** (16 789 « individus », 16 789 lignes)

Remarques : 16 « individus » proposés par le Céreq avec deux dates de naissance différentes sont identifiés par Pôle emploi comme des personnes différentes. Il y a donc 16 773 IDENT2 différents. L'IDENT_BNI n'est pas connu pour 179 lignes et 11 IDENT_BNI sont en doublon car associés à deux IDENT2 du CEREQ différents.

⇒ ***2°*** *S'il y a deux échos ou plus, alors :*

Cas 1 *:* l'identifiant BNI renseigné sur chaque ligne est identique. On suppose qu''il s'agit d'un seul et même individu qui a eu plusieurs identifiants Sigma. On conserve une ligne pour cet individu avec le dernier identifiant connu (date de mise à jour maximale) et on conserve les anciens identifiants Sigma **classés du plus récent au plus ancien.** Concrètement, création de 20 colonnes IDTIDV1 à IDTIDV20, pour conserver ces identifiants du plus récent au plus ancien au cas où ces identifiants seraient utiles

⇒ table **ID_MUL** (3 359 « individus », 3 359 lignes)

Remarque : 3 IDENT_BNI sont en doublon car associés à deux IDENT2 du CEREQ différents.

Cas 2 : l'identifiant BNI n'est pas renseigné sur chaque ligne ou bien il diffère d'une ligne à l'autre. On considèrera alors qu'il peut s'agir de plusieurs individus distincts. On conserve autant de lignes qu'il y a d'identifiants BNI distincts. Pour chaque identifiant BNI, on conserve l'historique des identifiants Sigma se rapportant à cet identifiant BNI, comme dans le cas 1.

⇒ table **ID_MULT_BNI** (1228 individus, 2742 lignes)

TABLES D'IDENTIFICATION DE PÔLE

ID_UNIQUE
14 janvier 2014
16 789 observations
(obs=individu Cereq à identifier)
IDENT_BNI : Id. individu à BNI
IDTIDV : Id. individu dans Sigma
IDENTIFIANT2 : Id. « Céreq » pour l'opération
+ Nom, prénom, mois et année de naissance

ID_MULT
14 janvier 2014
3 359 observations
(obs=individu Cereq à identifier)
IDENT_BNI : identifiant indiv. à BNI
IDTIDV1 à IDTIV20 : identifiants Indiv. Sigma successifs du plus récent au plus ancien
IDENTIFIANT2 : identifiant « Céreq » pour l'opération
+ Nom, prénom, mois et année de naissance

ID_MULT_BNI
14 janvier 2014
2 742 observations
IDENT_BNI : identifiant indiv. à BNI
IDTIDV1 à IDTIV20 : identifiants Indiv. Sigma successifs du plus récent au plus ancien
IDENTIFIANT2 : identifiant « Céreq » pour l'opération
+ Nom, prénom, mois et année de naissance

COMPIL
14 janvier 2014
22 890 observations
= empilement des trois précédents

Fichier CORRESP_IDENT
25 juillet 2014
28 262 observations
IDX : « identifiant » (FHS ?)
IDTIDV : « identifiants de l'individu » (Sigma ?)
IDENT: « identifiant du demandeur » (FHA ?)
Remarques :
1° IDENT=IDTIDV et jamais vides.
2° IDX non renseigné pour 2 853 observations (10% des observations)
3° IDX unique et non vide dans 24 884 cas, IDX en double (avec 2 IDTIDV différents) dans 261 cas et en tripe (avec 3 IDTIDV) dans 1 cas
4° Tous les IDTIDV sont différents
5° IDX différent de IDENT_BNI

TABLES DU FHS

(dans fichier zippé "avFHS201403" daté du 28 juillet 2014 ; les fichiers sont datés du 23 mai 2014)

DE : Table des demandeurs – demandes (sauf variable DEPCOM). Identifiants **IDX, IDTIDV, IDENT, NDEM**

D2 : table des indemnisations

E0 : Table des activités réduites (rattachées à des demandes en catégorie opérationnelle 1, 2 ou 3). Identifiants **IDX, IDTIDV, IDENT, NDEM**

PAP : Table des entretiens PAP (depuis juillet 2001). Identifiants **IDX, IDTIDV, IDENT, NDEM, NUMPAP**

PARCOURS : Tables des parcours des demandeurs d'emploi

STATDE : Table de statistiques individuelles sur le demandeur (sauf variable REGION). Identifiants **IDX, IDTIDV, IDENT**

RSA : table des droits RSA (sauf variable REGION)

Tables des variables historicisées (sauf variables REGION et NUMZUS)

TABLES DU FHA

(dans fichier zippé "avFHA201403" daté du 28 juillet 2014 ; les fichiers sont datés des 20 et 21 mai 2014)

Identifiants **IDTIDV, IDENT**

E3 : table de toutes les actions (conseillées, réalisées ou non)

M0 : table des mises en relations (positives ou non)

P2 : table des formations (sauf variable SIRET)

La collection *Working Paper* publie des textes pour engager le débat avec d'autres chercheurs.
La publication n'engage que l'auteur.

Au total, TABLE_IDENTIFICATION comprend 35 075 observations. Les individus pour lesquels deux dates de naissance différentes sont disponibles dans TABLE_PASSAGE_ENTIERE sont dupliqués et apparaissent donc deux fois dans cette table. Les 35 075 observations correspondent donc aux 33 655 individus initiaux dont 1 420 ayant deux dates de naissances.

La variable d'identification IDENTIFIANT2, destinée à Pôle emploi correspond à la variable IDENT2 de TABLE_PASSAGE_ENTIERE complétée d'un suffixe qui prend la valeur "A" pour les individus présents avec une seule date de naissance. Il prend la valeur "A" ou "B" pour les individus qui sont présents deux fois parce qu'ils ont deux dates de naissance disponibles.

La date du fichier TABLE_IDENTIFICATION (19 juin 2013) est cohérente avec la chronologie des opérations (ce qui n'est pas le cas du fichier TABLE_PASSAGE_ENTIERE récupéré).

Étape 2 : Transmission de la table du Céreq à Pôle emploi

D'après la convention, la transmission se serait faite par échange de données cryptées sur serveur FTP.

Étape 3a : Identification par Pôle emploi des individus présents dans ses bases pouvant correspondre à ceux enquêtés par le Céreq

Selon la note intitulée « Matching Céreq » datée du 11 avril 2014, qui semble expliquer les opérations effectuées à Pôle emploi, les opérations d'identification ont utilisé en entrée un fichier qui comporte les variables suivantes : l'identifiant Céreq, le nom, le prénom, l'année de naissance et mois de naissance. Cette description correspond au contenu du fichier TABLE_IDENTIFICATION.

Toujours selon cette même note, les opérations suivantes auraient été effectuées :

1° Dans une étape préliminaire, Pôle emploi formate la table transmise par le Céreq en supprimant les espaces et les tirets dans les noms et prénoms. Il transforme en format numérique l'année et le mois de naissance.

2° Pôle emploi créé un fichier des Individus présents dans ses fichiers de gestion (à partir de "l'axe individu de SID"). Pour limiter le volume de ce fichier, Pôle emploi filtre ses données sur l'année de naissance pour la limiter aux années 1969 à 1989, soit les personnes atteignant un âge compris entre 15 ans et 35 ans en 2004. Ce fichier conserve les variables suivantes : le nom de naissance, le prénom de naissance, l'année et le mois de naissance, la date de mise à jour, l'identifiant "Sigma" de l'individu (IDTIDV) et l'identifiant BNI de l'individu (IDENT_BNI) - une même personne est apriori identifiée de manière unique dans la BNI alors qu'elle peut être présente dans les fichiers de gestion avec plusieurs identifiants IDTIDV[23]. Là encore, les espaces et les tirets sont retirés des noms et prénoms.

3° La table des individus transmise par le Céreq est rapprochée de la table des individus de Pôle emploi sur la base du nom, de l'année et du mois de naissance. En sortie, ce type de rapprochement propose tous les couples possibles entre un individu de la table du Céreq et les individus de la table de Pôle emploi qui ont les mêmes valeurs pour le nom, l'année et le mois de naissance.

[23] L'indentifiant « Tranche de vie » change quand le demandeur d'emploi change de région opérationnelle. Il peut aussi changer quand, sans changer de région, le demandeur d'emploi change de sphère, par exemple en passant du statut de non salarié au statut de salarié.

Issues de la table CEREQ		Issues de la table Pole Emploi				
Nom-Année-Mois	**Prénom**	**Nom-Année-Mois**	**Prénom**	**IDENT_BNI**	**IDTIDV**	**Mise à Jour**
MACREAU-1985-04	Stéphane	MACREAU-1985-04	Stephane	AZ1928	1234	12/12/2005
MACREAU-1985-04	Stéphane	MACREAU-1985-04	Stephan	AZ1928	5679	20/04/2008
MACREAU-1985-04	Stéphane	MACREAU-1985-04	Stephan	BY2300	9875	03/05/2004
MACREAU-1985-04	Stéphane	MACREAU-1985-04	Judith	ZT8453	3948	02/01/2009
ROYAL-1988-05	Ségolène	ROYAL-1988-05	Ségolène	QQ9876	4389	22/09/2006
ROYAL-1988-05	Ségolène	ROYAL-1988-05	Ségolène	QQ9876	2987	11/07/2009
HOLLANDE-1981-05	François	HOLLANDE-1981-05	François	HJ7893	2735	14/07/2012
HOLLANDE-1981-05	François	HOLLANDE-1981-05	Cunéguonde	FT7623	5372	27/03/2014

Note : Illustration schématique, ne respectant pas le format des variables des vrais fichiers

4° Parmi tous les couples obtenus, seuls sont conservés ceux pour lesquels le prénom de Pôle emploi « ressemble » au prénom du Céreq (la notion de ressemblance n'est pas précisée).

Issues de la table CEREQ		Issues de la table Pole Emploi				
Nom-Année-Mois	**Prénom**	**Nom-Année-Mois**	**Prénom**	**IDENT_BNI**	**IDTIDV**	**Mise à Jour**
MACREAU-1985-04	Stéphane	MACREAU-1985-04	Stephane	AZ1928	1234	12/12/2005
MACREAU-1985-04	Stéphane	MACREAU-1985-04	Stephan	AZ1928	5679	20/04/2008
MACREAU-1985-04	Stéphane	MACREAU-1985-04	Stephan	BY2300	9875	03/05/2004
~~MACREAU-1985-04~~	~~Stéphane~~	~~MACREAU-1985-04~~	~~Judith~~	~~ZT8453~~	~~3948~~	~~02/01/2009~~
ROYAL-1988-05	Ségolène	ROYAL-1988-05	Ségolène	QQ9876	4389	22/09/2006
ROYAL-1988-05	Ségolène	ROYAL-1988-05	Ségolène	QQ9876	2987	11/07/2009
HOLLANDE-1981-05	François	HOLLANDE-1981-05	François	HJ7893	2735	14/07/2012
~~HOLLANDE-1981-05~~	~~François~~	~~HOLLANDE-1981-05~~	~~Cunéguonde~~	~~FT7623~~	~~5372~~	~~27/03/2014~~

Note : Illustration schématique, ne respectant pas le format des variables des vrais fichiers

5° Si, à l'issue de l'étape précédente, un individu du Céreq reste identifié de manière unique dans la table Pôle emploi, c'est-à-dire s'il ne reste qu'un seul couple pour cet individu Céreq, l'observation alimente la table ID_UNIQUE. Cette table comprend 16 789 observations, qui correspondent à autant d'individus différents.

Dans ID_UNIQUE				
Nom-Année-Mois	**Prénom**	**IDENT_BNI**	**IDTIDV**	**Mise à Jour**
HOLLANDE-1981-05	François	HJ7893	2735	14/07/2012

Note : Illustration schématique, ne respectant pas le format des variables des vrais fichiers

6° Pour les individus du Céreq pour lesquels il reste plusieurs couples, c'est-à-dire si l'individu du Céreq correspond à plusieurs individus de la table de Pôle emploi, qui ont le même nom, le même mois de naissance, la même année de naissance et un prénom « ressemblant », alors les doublons sont traités en considérant qu'un même identifiant à la BNI correspond à la même personne dans les fichiers de Pôle emploi. Deux cas sont donc possibles.

Dans le premier cas, l'identifiant à la BNI est renseigné et il est identique pour tous les individus de Pôle emploi en doublon. On considère alors qu'il s'agit du même individu, qui dispose de plusieurs identifiants Sigma. On ne conserve donc qu'une observation pour cet individu mais on conserve aussi l'ensemble des identifiants Sigma successifs dont il dispose. L'identifiant sigma le plus récent est conservé dans la variable IDTIDV. Les autres alimentent les variables IDTIDV1 à IDTID20, en allant du plus récent au plus ancien. L'observation est intégrée à la table ID_MULT. Cette table comprend 3 359 observations, qui correspondent à autant d'individus.

Dans ID_MULT						
Nom-Année-Mois	**Prénom**	**IDENT_BNI**	**IDTIDV**	**IDTIDV1**	**IDTIDV2**	**IDTIDV3**
ROYAL-1988-05	Ségolène	QQ9876	2987	4389		

Note : Illustration schématique, ne respectant pas le format des variables des vrais fichiers.

Dans le second cas, l'identifiant à la BNI n'est pas toujours renseigné ou n'est pas le même pour les deux individus de Pôle emploi qui sont associés au même individu du Céreq. On considère alors qu'il peut s'agir de plusieurs individus distincts et on conservera autant de lignes qu'il y a d'identifiants BNI distincts. Pour chaque identifiant BNI, l'historique des identifiants Sigma est conservé selon le même schéma que précédemment. Les observations intègrent la table ID_MULT_BNI. Cette table comprend 2 742 lignes, correspondant à 1 228 individus du Céreq différents.

Dans ID_MULT_BNI						
Nom-Année-Mois	**Prénom**	**IDENT_BNI**	**IDTIDV**	**IDTIDV1**	**IDTIDV2**	**IDTIDV3**
MACREAU-1985-04	Stéphane	AZ1928	5679	1234		
MACREAU-1985-04	Stéphane	BY2300	9875			

Note : Illustration schématique, ne respectant pas le format des variables des vrais fichiers.

Étape 3b : Récupération par Pôle emploi des extraits de ses fichiers historiques à transmettre au Céreq

La récupération des informations extraites pour le Céreq des fichiers historiques de Pôle emploi a été effectuée par Pôle emploi. Elle n'est pas documentée dans les notes conservées au Céreq.

L'examen des données transmises via la tabulation de la variable MOIS de la table E0 sur l'activité réduite permet juste de savoir que le fichier historique statistique utilisé est celui s'arrêtant en mars 2014, qui a dû être produit en mai 2014. Cette datation est aussi cohérente avec la date des fichiers transmis (20 et 21 mai 2014 pour les tables extraites du FHA et 23 mai 2014 pour les tables extraites du FHS). La fenêtre temporelle des dix ans couvre donc la période allant d'avril 2004 à mars 2014.

Étape 4 : Transmission des données de Pôle emploi au Céreq

Les modalités techniques de cette transmission ne sont pas connues.

Sur la base des fichiers récupérés, Pôle emploi a transmis au Céreq **les tables ID_UNIQUE, ID_MULT et ID_MULT_BNI issus de ses travaux d'identification**. Ces tables comportent les informations suivantes :

- l'identifiant non signifiant transmis par le Céreq (IDENTIFIANT2)
- une variable avec le nom reformaté (sans tiret, ni espace) ;
- deux variables de prénoms reformatés (souvent les deux variables sont identiques mais elles peuvent différer légèrement sur l'orthographe, voire comporter plusieurs prénoms accolés pour l'une d'elle, l'autre accolant cependant parfois aussi deux prénoms) ;
- le mois et l'année de naissance,
- l'identifiant à la BNI (IDENT_BNI),
- le ou les identifiants "Sigma" (IDTIDV, IDTIDV1 à IDTIDV20).

Un fichier COMPIL semble regrouper par empilement ces trois fichiers. Comme eux, il date du 14 janvier 2014. Il regroupe 22 890 observations, soit la somme des observations des trois fichiers. Il comporte les mêmes variables, augmentées d'une variable supplémentaire (FIC) qui indique le fichier source de l'observation, comme le suggère sa tabulation :

Tableau 4 • Tabulation de la variable FIC de la table COMPIL

FIC	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
Id_mult	3 359	14.67	3 359	14.67
Id_mult_BNI	2 742	11.98	6 101	26.65
Id_unique	16 789	73.35	22 890	100.00

Deux fichiers zippés datés du 28 juillet 2014 ont été également transmis, **l'un avec les tables issues du FHS, l'autre avec les tables issues du FHA :**

- Dans les fichiers du FHA récupéré, deux variables destinées à identifier les individus sont disponibles : une variable IDENT et une variable IDTIDV. Selon la documentation du FHA, seule une variable IDENT est présente normalement dans le FHA, comme « identifiant unique au sein de chaque Assédic ». La variable IDENT présente dans les fichiers a le même format et le même libellé que ceux mentionnés dans la documentation du FHA. Compte tenu de son appellation, la variable IDTIDV correspondrait à l'identifiant « Sigma » déjà évoqué, qui peut changer pour un même individu.

- Dans les fichiers du FHS récupérés, trois variables destinées à identifier les individus sont disponibles : une variable IDX et, comme pour le FHA, une variable IDTIDV et une variable IDENT, qui sont là encore strictement identiques. Selon la documentation du FHS, seule la variable IDX est normalement présente dans le FHS. La présence des deux autres variables implique donc un traitement spécifique, sans doute effectué lors de l'extraction des seules observations concernant le Céreq.

- La variable IDX est normalement incrémentée par région et peut donc changer pour un individu qui change de région sauf dans le super fichier historique où chaque individu dispose d'un unique IDX.

Enfin, une dernière **table CORRESP_IDENT**, datée du 25 juillet 2014, a été récupérée (nous faisons l'hypothèse qu'elle provient aussi de Pôle emploi). Elle comprend 28 262 observations et trois variables : IDX, qui a pour libellé « Identifiant » ; IDTIDV, qui a pour libellé « identifiant de l'individu » et « IDENT », qui a pour libellé « identifiant du demandeur ».

3. Travaux exploratoires sur les tables d'identification

3.1 Le contenu de la table TABLE_IDENTIFICATION (Céreq)

La table TABLE_IDENTIFICATION correspond à la table des individus que le Céreq semble avoir transmis à Pôle emploi pour procéder aux opérations de recherche et d'identification de ces personnes dans les fichiers de Pôle emploi.

Cette table comprend 35 075 observations, qui correspondent aux 33 655 répondants de l'échantillon national de la première interrogation de l'enquête Génération 2004, dont 1 420 sont présents en double, car avec deux dates de naissance disponibles dans la table initiale TABLE_PASSAGE_ENTIERE. 4,2 % des personnes à identifier sont donc concernés. Pour ces personnes, la date de naissance issue de la base de sondage (ANNAI MOISA) est différente de celle collectée à la première interrogation (questions Q2 et Q3). L'écart porte essentiellement sur le mois de naissance (82 % des cas), puis sur l'année de naissance uniquement (12 % des cas) et rarement sur les deux (6 %).

3.2 Le contenu de la table ID_UNIQUE (Pôle emploi)

La table ID_UNIQUE comprend les individus de la table d'identification du Céreq qui ont été identifiés de façon unique par Pôle emploi dans ses fichiers. Elle comporte 16 789 observations, à rapprocher des 35 075 observations de la table transmise par le Céreq. 47,9 % de ces « individus » ont donc été retrouvés par Pôle emploi de façon univoque, avec le même nom, le même mois de naissance, la même année de naissance et un prénom "ressemblant".

La table transmise par le Céreq comprend cependant 33 655 individus *a priori* différents, dont 1 420 pour lesquelles deux dates de naissance possible ont été transmises à Pôle emploi, l'une issue de la base de sondage, l'autre issue du questionnaire. L'examen de la variable IDENT2**,** qui identifie de manière unique les individus de l'enquête Génération montre que sur 16 789 observations identifiées dans la table ID_UNIQUE :

- 16 757 identifiants sont présents une seule fois, correspondant à autant d'observations,
- 16 identifiants sont présents en doublon, soit 32 observations. Il s'agit d'individus figurant deux fois dans la table transmise par le Céreq parce que deux dates de naissance différentes étaient disponibles pour eux (la variable IDENTIFIANT2 est donc présente pour eux avec le suffixe "A" et avec le suffixe "B"). Pour chacune des deux dates de naissance, Pôle emploi a retrouvé une personne unique différente (les variables IDENT_BNI et IDTIDV prennent deux valeurs différentes pour les deux observations en doublon sur IDENT2).

Ainsi, en faisant abstraction de ces identifications en doublon, 16 757 individus auraient été repérés de façon unique dans les fichiers de Pôle emploi, soit 49,8 % des 33 655 individus de la partie nationale de l'enquête Génération 2004.

L'examen de la variable IDENT_BNI, qui doit permettre d'identifier de manière unique les individus dans les fichiers de Pôle emploi lorsqu'ils changent d'identifiant « Sigma » montre que sur 16 789 observations :

- l'identifiant est vide pour 80 observations,
- l'identifiant vaut YYYY pour 97 observations,
- l'identifiant vaut 0000000000 pour deux observations,
- l'identifiant est présent en doublon pour 11 cas, soit 22 observations. Ces cas semblent correspondre à des personnes présentes en doublon dans l'enquête Génération : deux identifiants dans l'enquête sont associés à une même identité (ce point mériterait d'être investigué pour voir si, effectivement, une même personne a pu être enquêtée deux fois par le Céreq, ou si ce constat résulte d'erreurs dans la production du fichier d'identification transmis à Pôle emploi).

L'examen de la variable IDTIDV, qui correspond à l'identifiant "Sigma" susceptible de changer pour un même individu au cours de sa trajectoire d'inscription à Pôle emploi, montre que sur 16 789 observations de ID_UNIQUE:

- 16 763 identifiants ne sont présents qu'une fois, correspondant à autant d'observations,
- 13 identifiants sont en doublon, correspondant à 26 observations. Parmi eux, 11 doublons, soit 22 observations, correspondent aux doublons sur la variable IDENT_BNI évoqué juste auparavant. Les deux derniers doublons, qui correspondent à quatre observations, s'expliquent de la même façon. Ils ne sont pas évoqués parmi les doublons pour IDENT_BNI parce que dans leur cas la variable IDENT_BNI n'est pas renseignée.

Au total, en combinant les différentes "anomalies", sur 16 789 observations de la table ID_UNIQUE :

- 16 556 observations disposent de tous leurs identifiants renseignés, sans que ceux-ci soient présents en doublon dans la table,
- 179 observations n'ont pas de valeur renseignée pour IDENT_BNI mais dispose d'un identifiant Sigma renseigné unique (dont 4 sont aussi comptabilisées *infra* dans les 58 observations nécessitant des investigations complémentaires),
- 58 observations nécessiteraient des investigations complémentaires pour arbitrer (32 observations correspondent aux 16 individus de l'enquête Génération identifiés sur leurs deux dates de naissance et 26 observations correspondent aux 13 individus de l'enquête Génération présents avec deux identifiant Céreq différents).

Se limiter aux 16 731 observations univoques et non ambigües revient à conserver 99,7 % des 16 789 identifiants Céreq (IDENT2) présents dans la table ID_UNIQUE[24]. C'est donc le choix que nous ferons plus loin pour examiner les caractéristiques des personnes appariées.

3.3 Le contenu de la table ID_MULT (Pôle emploi)

La table ID_MULT comprend les individus de la table d'identification du Céreq qui ont été identifiés de façon unique par Pôle emploi dans ses fichiers sur la base d'un IDENT_BNI renseigné, mais pour lesquels Pôle emploi a récupéré plusieurs identifiants Sigma (IDTIDV). Cette situation est attendue puisqu'un même demandeur d'emploi peut changer d'identifiant Sigma, notamment lorsqu'il change d'Assédic, l'identifiant à la BNI étant sensé permettre le suivi de ce demandeur d'emploi dans les fichiers de gestion.

Cette table comporte 3 359 observations, à rapprocher des 35 075 observations de la table transmise par le Céreq, soit 9,6 % de ses « individus », qui ont été retrouvés par Pôle emploi avec le même nom, le même mois de naissance, la même année de naissance et un prénom "ressemblant" à ceux transmis par le Céreq.

71 % des observations sont présentes avec deux identifiants Sigma, 21 % avec trois identifiants et 8 % avec davantage (tableau 5). De façon anormale, une observation n'a qu'un seul identifiant Sigma.

Comme pour la table ID_UNIQUE, un examen sur les éventuels identifiants en doublon a été effectué. Il en ressort qu'il existe :

- trois doublons sur la variable IDENT_BNI, lié à trois individus de la table du Céreq présents avec la même identité mais avec deux identifiants IDENT2 différents,
- un doublon sur la variable IDENT2, pour une personne pour laquelle deux dates de naissance ont été transmises par le Céreq et qui aboutissent chacune à l'identification d'une personne différente dans les fichiers de Pôle emploi, donc à deux IDENT_BNI différents.

Là encore, nous ne retiendrons que les observations univoques et non ambigües pour examiner les caractéristiques des personnes appariées.

[24] 16 789 observations – 16 observations pour lesquelles IDENT2 est présent deux fois, avec des dates de naissance différentes.

Tableau 5 • Nombre d'identifiants IDTIDV récupérés par observation de la table ID_MULT

Nb_idtidv	Fréquence	Pourcentage
1	1	0.03
2	2393	71.24
3	700	20.84
4	177	5.27
5	57	1.70
6	22	0.65
7	7	0.21
8	1	0.03
9	1	0.03

3.4 Le contenu de la table ID_MULT_BNI (Pôle emploi)

La table ID_MULT_BNI comprend les individus de la table d'identification du Céreq qui ont été identifiés avec plusieurs IDENT_BNI différents par Pôle emploi dans ses fichiers. Dans leur cas, les variables utilisées pour effectuer l'appariement sont insuffisantes pour identifier les personnes de façon univoque dans les fichiers de Pôle emploi.

Il faut souligner que si cette difficulté d'identification univoque apparait explicitement pour les individus de cette table, elle peut aussi exister pour les personnes identifiées de façon univoque dans les fichiers de Pôle emploi, donc retenues dans les tables ID_UNIQUE et ID_MULT. Dans leur cas, l'identification univoque pourrait parfois résulter d'un rapprochement erroné de l'identité d'un répondant de l'enquête Génération qui n'a jamais été inscrit à Pôle emploi avec une identité ressemblante d'un inscrit à Pôle emploi.

La table ID_BNI_MULT comprend 2 742 observations, correspondant à 1 224 individus différents de la table du Céreq (IDENT2). Parmi eux, 1 039 individus de la table Céreq sont présents deux fois, dont 6 correspondent à un doublon dans la table du Céreq. Les 1 033 autres enquêtés du Céreq présents deux fois sont donc associés à deux personnes différentes dans les fichiers de Pôle emploi. Par ailleurs, 120 individus de la table Céreq sont présents trois fois ; 39 sont présents quatre fois et 26 sont présents cinq fois ou plus.

Dans cette table des personnes identifiées avec plusieurs identifiants BNI différents :

- 203 observations sont en réalité sans identifiants BNI et 9 observations ont un identifiant BNI valant "0000000000".
- Par ailleurs, 11 identifiants BNI renseignés sont en doublon, soit 22 observations (tableau 9). Dans six cas, le doublon résulte de la présence d'une même personne présente avec deux identifiants différents dans la table du Céreq. Dans un septième cas, il semble plutôt s'agir du rapprochement de deux personnes différentes dans la table du Céreq avec une même identité dans les fichiers de Pôle emploi, en raison, sans doute, d'un prénom jugé « ressemblant » (Christelle et Christopher).

Dans 83,4 % des cas, les observations de la table ID_MULT_BNI ne sont associées qu'à un seul identifiant Sigma, dans 11,1 % des cas à deux identifiants Sigma et dans 5,5 % des cas à trois identifiants Sigma ou plus (tableau 6). Ces proportions sont analogues à celles obtenues pour les personnes identifiées avec un seul identifiant à la BNI puisque si l'on compile la table ID_UNIQUE et ID_MULT, 83,3 % des observations sont associées à un seul identifiant Sigma, 11,9 % à deux identifiants Sigma et 4,8 % à trois identifiants Sigma ou plus.

Tableau 6 • Nombre d'identifiants IDTIDV récupérés par observation de la table ID_MULT_BNI

Nb_idtidv	Fréquence	Pourcentage
1	2 286	83.37
2	304	11.09
3	101	3.68
4	38	1.39
5	5	0.18
6	6	0.22
7	2	0.07

3.5 Le contenu de la table CORRESP_IDENT (Pôle emploi)

La table CORRESP_IDENT comprend 28 262 observations et trois variables uniquement :
- IDTIDV, correspondant apriori à l'identifiant "Sigma"
- IDENT, nom de la variable de l'identifiant du FHA
- IDX, nom de la variable de l'identifiant du FHS.

Sans que l'on ait cherché à s'en assurer effectivement par une analyse appropriée, le nombre d'observations s'approche du nombre d'IDTIDV différents repérés dans les trois tables d'identification, soit : 16 789 dans ID_UNIQUE, 8 078 dans ID_MULT et 3 424 dans ID_MULT_BNI. La somme de ces trois nombres vaut 28 291, soit un écart de 29, ce qui correspond sans doute au cas des personnes présentes dans deux des trois tables (voir infra).

Par ailleurs, l'analyse de la table CORRESP_IDENT montre que :
- les observations ont toutes un IDTIDV different,
- IDX est non renseigné pour 2 853 observations, soit 10% des observations,
- IDX est unique et non vide pour 24 884 observations. Il est en doublon, donc associé à deux IDTIDV différents, dans 261 cas (522 observations) et en triple, donc avec trois IDTIDV, dans un cas (3 observations).

L'articulation des tables d'identification pour associer l'enquête Génération aux données de Pole Emploi s'effectue donc ainsi :
- la TABLE_PASSAGE_ENTIERE du Céreq permet d'associer l'identifiant dans l'enquête Génération (IDENT) à l'identifiant créé par le Céreq pour l'opération d'appariement (IDENTIFIANT2, IDENT2) ;
- les tables ID_UNIQUE, ID_MULT et ID_MULT_BNI permettent d'associer l'identifiant créé par le Céreq pour l'opération d'appariement (IDENTIFIANT2, IDENT2) à l'identifiant Sigma de Pôle emploi (IDTIDV) ;
- la table CORRESP_IDENT permet d'associer l'identifiant Sigma de Pôle emploi (IDTIDV) aux identifiants non signifiants des tables du FHS (IDX) et deu FHA (IDENT).

3.6 Synthèse de la mise en œuvre de l'identification dans les fichiers de Pôle emploi

En première analyse, par simple juxtaposition des analyses effectuées pour chacune des trois tables d'identification transmises par Pôle emploi, sur 33 655 répondants à l'enquête Génération 2004 identifiés par la variable IDENT2 (tableau 7) :
- **20 082 auraient été retrouvés de façon univoque dans les fichiers de Pôle emploi, soit 59,7%.** Parmi eux, 16 732 disposent d'un seul identifiant Sigma (IDTIDV) répondants et 3 350 de deux identifiants[25] ou plus ;

[25] Une observation de ID_MULT ne comporte qu'une valeur d'IDTIDV.

La collection *Working Paper* publie des textes pour engager le débat avec d'autres chercheurs.
La publication n'engage que l'auteur.

- **1 273, soit moins de 5%, auraient été retrouvés avec des ambiguïtés repérées sur l'identification**. Parmi les personnes concernées, 1 218 personnes sont retrouvées comme étant associées à plusieurs personnes à la BNI par Pôle emploi[26]. 17 personnes ont été identifiées par Pôle emploi comme deux personnes différentes pour chacune des deux dates de naissance transmises par le Céreq[27]. 19 personnes ayant apparemment la même identité sont présentes deux fois dans la table transmise par le Céreq, avec deux identifiants IDENT2 différents[28].

Tableau 7 • Synthèse de la phase d'identification par Pôle emploi par table transmise

	Nombre d'obs.		1 IDTIDV	2	3	4 ou plus
ID_UNIQUE	**16 789**					
	16 731	personnes différentes avec IDENT2 unique	16 731			
	32	16 IDENT2 présents deux fois car identifié par Pôle emploi avec leurs deux dates de naissance A et B	32			
	26	13 personnes ayant apparemment la même identité mais disposant de deux IDENT2 différents dans la table du Céreq	26			
ID_MULT	**3 359**					
	3 351	Personnes différentes avec IDENT2 unique	1	2 386	699	265
	2	1 IDENT2 présent deux fois car identifié avec ses deux dates de naissance A et B		1	1	
	6	3 personnes ayant apparemment la même identité mais disposant de deux IDENT2 différents dans la table du Céreq		6		
ID_MULT_BNI	**2 742**	1224 IDENT2 différents				
	2 078	1039 IDENT2 présents deux fois				
	360	120 IDENT2 présents trois fois				
	156	39 IDENT2 présents 4 fois (dont 3 personnes ayant apparemment la même identité mais disposant de deux IDENT2 différents dans la table du Céreq)				
	148	26 IDENT2 présents 5 fois ou plus				

La réalité est un peu plus complexe que cela car, même si cela ne joue qu'à la marge, il faut aussi regarder si des répondants ont été identifiés dans plusieurs des trois tables transmises par Pôle emploi. Après analyse, c'est le cas de 11 d'entre eux :

- quatre individus de l'enquête Génération, sur la base de la variable IDENT2, sont communs à la table ID_UNIQUE et à la table ID_MULT ;
- quatre individus de l'enquête Génération, sur la base de la variable IDENT2, sont communs à la table ID_UNIQUE et à la table ID_MULT_BNI ;
- trois individus de l'enquête Génération, sur la base de la variable IDENT2, sont communs à la table ID_MULTet à la table ID_MULT_BNI.

Les 11 individus concernés sont présents dans deux tables parce qu'ils disposent de deux dates de naissance. Ils figurent donc deux fois, avec deux valeurs différentes de la variable IDENTIFIANT2, dans la table d'identification transmis par le Céreq à Pôle emploi et chacune de ces « identités » a donné lieu à une identification par Pôle emploi, dans deux tables différentes.

[26] Soit les 1224 IDENT2 différents de la table ID_MULT_BNI diminués des 6 IDENT2 liés aux 3 personnes ayant apparemment la même identité mais disposant de deux IDENT2 différents dans la table du Céreq.
[27] Soit 16 personnes dans ID_UNIQUE et une personne dans ID_MULT.
[28] Soit 13 personnes dans ID_UNIQUE, 3 personnes dans ID_MULT et 3 personnes dans ID_MULT_BNI, ce qui correspond donc à 38 valeurs différentes pour IDENT2.

En tenant compte de ces onze individus, ce sont donc 20 067 individus[29] parmi les 33 655 répondants à l'enquête Génération qui ont été retrouvés de façon univoque dans les fichiers de Pôle emploi, soit 59,6 %, tandis que moins de 5 %, auraient été retrouvés avec des ambiguïtés repérées sur l'identification (tableau 8).

Tableau 8 • Synthèse globale de la phase d'identification par Pôle emploi

	Suffixe dans IDENTIFIANT2			
	A	**B**	**AB**	**Total**
Identification simple	**19 344**	**723**	**0**	**20 067**
- dans ID_UNIQUE	*16 125*	*598*	*0*	*16 723*
- dans ID_MULT	*3 219*	*125*	*0*	*3 344*
identification complexe	**1 205**	**40**	**32**	**1 277**
Non identifié par Pôle emploi				**12 311**
Ensemble				33 655

Pour traiter les cas ambigus, il faudrait mobiliser des informations complémentaires du fichier historique pour les rapprocher d'informations théoriquement proches disponibles dans l'enquête Génération comme le niveau de formation, le lieu de résidence à la date d'enquête, la compatibilité du calendrier professionnel et des séquences d'inscription à Pôle emploi, avec ou sans activité réduite. Cependant, vu le faible nombre d'identifications ambigües relativement aux autres, il pourrait y avoir intérêt à se limiter aux seules personnes retrouvées de façon univoque pour s'économiser des traitements manuels couteux en temps et d'un apport incertain.

Par ailleurs, comme nous l'avons déjà indiqué, certains cas d'identifications ambigües constatées suggèrent qu'il existe aussi des identifications faussement univoques qui résulteraient du rapprochement d'une personne de l'enquête Génération n'ayant jamais été inscrite avec une personne inscrite ayant une identité proche sur les critères retenues pour l'identification. Par construction, il est difficile de repérer ces cas. Seule l'analyse comparée des situations d'emploi et des trajectoires dans l'enquête Génération relativement à celles fournies par les données de Pôle emploi pourrait permettre de quantifier statistiquement des situations « anormales » pouvant résulter, en partie, de ces erreurs d'identification.

Concernant les 1 420 individus qui ont été doublonnés par le Céreq en raison de la disponibilité de deux dates de naissance :

- 509, soit 36 %, n'ont pas été retrouvés par Pôle emploi,
- 723[30], soit 51 %, ont été retrouvés de manière unique grâce à la ligne suffixée « B », qui semble correspondre pour les doublons, aux informations collectées dans l'enquête (questions Q2 et Q3),
- 105[31], soit 7 %, sont retrouvés de manière unique par Pôle emploi grâce à la ligne « A », qui correspond aux données des bases de sondage,
- 40, soit 3 %, sont retrouvés avec des échos multiples à la BNI[32] grâce à la ligne « B »,
- 11, soit moins de 1 %, sont retrouvés avec des échos multiples à la BNI grâce à la ligne « A »,
- 32, soit 2 %, aboutissent des échos pour Pôle emploi pour les deux lignes « A » et « B ».

L'intérêt de transmettre deux dates de naissance pourrait être questionné. Il pourrait être plus simple de n'en transmettre qu'une en donnant la priorité à la date de naissance validée lors de l'enquête.

[29] Au 20 082 comptabilisés d'abord, il faut retirer 15 observations : les 4 observations comptées à la fois dans ID_UNIQUE et ID_MULT (soit 8 observations sur les deux tables), les 4 observations comptabilisées dans ID_UNIQUE qui sont en doublon dans ID_MULT_BNI et les 3 observations comptabilisées dans ID_MULT, qui sont en doublon dans ID_MULT_BNI.
[30] Parmi eux, 598 sont dans la table ID_UNIQUE et 125 dans la table ID_MULT.
[31] Parmi eux, 89 sont dans la table ID_UNIQUE et 16 dans la table ID_MULT.
[32] Ils sont donc dans la table ID_MULT_BNI.

4. Caractéristiques des personnes retrouvées de manière univoque par rapport aux autres

Cette partie propose d'examiner si les caractéristiques des personnes que l'on retrouve dans les fichiers de Pôle emploi sont cohérents avec ce que l'on sait des difficultés différentielles d'insertion sur la marché du travail dans l'enquête Génération. Cette première analyse n'exploite pas les tables du FHS et du FHA. Elle se limite à exploiter un rapprochement de l'enquête Génération 2004 à sept ans et d'un fichier synthétisant les résultats de la phase d'identification de Pôle emploi. Pour mémoire, parmi les 33 655 individus ayant répondu à la première interrogation dans le tronc commun national, seuls 12 365 jeunes ont répondu à la troisième interrogation. Les statistiques présentées se limitent à ces derniers. Ce choix est cohérent avec le fait que le travail d'identification de Pôle emploi a été réalisé début 2014 et couvre donc des personnes inscrites au moins une fois sur les dix années précédentes.

Tableau 9 • Situation d'identification selon le sexe (variable Q1)

	Identification simple	Non identifié par PE	identification complexe	Effectifs non pondérés
Hommes	59 %	37 %	4 %	5 936
Femmes	56 %	41 %	3 %	6 429
Ensemble	57 %	39 %	3 %	12 365

Tableau 10 • Situation d'identification selon le type de trajectoire d'insertion (variable TYPOTRAJ_11)

	Identification simple	Non identifié par PE	identification complexe	Effectifs non pondérés
1 = ils se sont stabilisé rapidement en EDI	41%	56%	3%	4753
2 = ils se sont stabilisé en EDI de façon différée	63%	33%	3%	2440
3 = ils sont restés durablement en EDD	68%	27%	5%	1310
4 = ils se sont stabilisé tardivement en EDI	70%	27%	4%	2254
5 = ils ont décroché de l'emploi	71%	25%	4%	832
6 = ils ont connu des épisodes de chômage persistants ou récurrents	75%	21%	4%	432
7 = ils ont passé une ou plusieurs longues périodes en dehors du marché du travail	65%	32%	3%	344

Tableau 11 • Situation d'identification selon le niveau de sortie (variable NIVSOR9)

	Identification simple	Non identifié par PE	identification complexe	Effectifs non pondérés
Non diplômé	70 %	26 %	4 %	1 056
CAP-BEP-MC	62 %	33 %	4 %	1 887
Bac	61 %	36 %	4 %	2 938
Deug	53 %	43 %	4 %	265
BTS-DUT-Santé social	46 %	52 %	3 %	2 549
Licence, L3	50 %	47 %	4 %	1 186
Maitrise, M1,…	64 %	33 %	3 %	683
DEA, DESS, M2	62 %	35 %	3 %	1 609
Doctorat	42 %	56 %	2 %	192

Tableau 12 • Situation d'identification selon le plus haut diplôme (variable PHDIP)

	Identification simple	Non identifié par PE	identification complexe	Effectifs non pondérés
Non diplômé	69 %	27 %	4 %	961
CAP-BEP-MC Tertiaire	67 %	29 %	4 %	820
CAP-BEP-MC Industriel	60 %	35 %	5 %	961
Bac pro./techno. Tertiaire	65 %	31 %	4 %	1 438
Bac pro./techno. Industriel	55 %	41 %	4 %	951
Bac. Général	58 %	39 %	3 %	634
Bac.+2 Santé-Social	28 %	70 %	2 %	1 292
Bac.+2 Tertiaire	64 %	32 %	4 %	951
Bac.+2 Industriel	61 %	37 %	2 %	610
Licence pro.	60 %	36 %	5 %	345
L3 LSH, Gestion, Droit	49 %	48 %	3 %	520
L3 Science, Santé, STAPS	42 %	56 %	2 %	210
M1	59 %	37 %	4 %	742
M2 LSH, gestion, Droit	63 %	33 %	4 %	696
Ecoles de commerce bac.+5	58 %	41 %	1 %	120
M2 Science, Santé, STAPS	65 %	32 %	3 %	433
Ecoles d'ingénieurs	58 %	39 %	3 %	475
Doctorat	41 %	57 %	2 %	206

Tableau 13 • Situation d'identification selon le nombre total de périodes de non-emploi déclarées dans l'enquête Génération à sept ans (variable NSCHO_TOT)

	Identification simple	Non identifié par PE	identification complexe	Effectifs non pondérés
Aucune	35 %	63 %	2 %	5 516
1	67 %	29 %	4 %	295
2	72 %	23 %	6 %	394
3	75 %	21 %	4 %	2 720
4	73 %	22 %	5 %	366
5	77 %	17 %	5 %	419
6	78 %	17 %	5 %	950
7	76 %	20 %	3 %	279
8	74 %	21 %	5 %	290
9	82 %	16 %	2 %	377
10	81 %	14 %	6 %	154
11	79 %	18 %	3 %	153
12	82 %	14 %	4 %	141
13 ou plus	71 %	26 %	3 %	311

La collection *Working Paper* publie des textes pour engager le débat avec d'autres chercheurs.
La publication n'engage que l'auteur.

Les tableaux 9 à 13 montrent que les jeunes ayant connu des parcours d'insertion difficiles sont plus souvent identifiés dans les fichiers de Pôle emploi que les autres. Ainsi, par exemple, 70 % des jeunes qui se sont stabilisés tardivement dans un emploi à durée indéterminée sont retrouvés de façon simple dans les fichiers de Pôle emploi et 75 % des jeunes ayant connu des épisodes persistants et récurrents de chômage. *A contrario*, ce taux n'est que de 41 % pour les jeunes ayant une trajectoire de stabilisation rapide en emploi à durée indéterminée. Ce résultat est plutôt cohérent avec ce à quoi l'on pouvait s'attendre.

On retrouve également une certaine cohérence quand on examine les variables corrélées aux difficultés d'insertion, comme le niveau de sortie, le plus haut diplôme ou le nombre de périodes de non-emploi déclaré. Ainsi, par exemple, 69 % des non-diplômés sont retrouvés de façon simple dans les fichiers de Pôle emploi, contre 58 % des diplômés des écoles d'ingénieurs ou de commerce. De même, les deux tiers des jeunes qui ont déclaré une seule période de chômage au cours des sept années ayant suivi leur sortie de formation initiale sont retrouvés dans les fichiers de Pôle emploi et les trois quarts des jeunes ayant déclaré deux périodes de chômage ou plus, contre 35 % pour ceux qui n'ont déclaré aucune période de non-emploi dans l'enquête Génération.

Enfin, quelles que soient les caractéristiques des jeunes, la proportion de personnes identifiées dans les fichiers de Pôle emploi, donc inscrites au moins une fois sur l'ensemble de la période, est élevée. Comme nous l'avons dit, c'est par exemple le cas de 35 % des jeunes ne déclarant aucune période de chômage sur sept ans à l'enquête Génération, 41 % des jeunes ayant une trajectoire de stabilisation rapide en emploi à durée indéterminée et 58 % des diplômés des écoles d'ingénieurs ou de commerce. Le passage par Pôle emploi est donc une étape fréquente pour tous au cours des premières années dans la vie active.

Pour aller plus loin, il faut examiner si ces écarts dans la propension à être retrouvé dans les fichiers de Pôle emploi se renforce quant aux caractéristiques des périodes d'inscription, en examinant le nombre de mois d'inscription à Pôle emploi, avec ou sans activité réduite, au cours des 98 mois couverts par l'enquête Génération.

5. Quelques analyses sur les périodes d'inscriptions à partir du rapprochement de la table DE du FHS et de l'enquête Génération

Afin de prolonger les travaux présentés dans la partie précédente, nous avons rapproché la table des demandes d'emploi du fichier historique statistique (table DE) et les résultats de la troisième interrogation de l'enquête Génération 2004. Ce rapprochement permet de relier les durées d'inscription à Pôle emploi à des caractéristiques des répondants à l'enquête, en particulier le niveau de sortie des jeunes, leur plus haut diplôme et leur type de trajectoires sur sept ans selon la typologie élaborée par le Céreq. Comme nous le verrons, les exploitations réalisées aboutissent à des résultats qui semblent relativement cohérents avec les résultats de l'enquête Génération.

La table DE a été privilégiée parce qu'elle a un rôle central dans le dispositif du fichier historique statistique. Les autres tables du fichier ne s'exploitent qu'en articulation avec elle et elle permet de connaître :

- les durées totales d'inscription dans les catégories opérationnelles 1, 2 ou 3 (personnes sans emploi astreinte aux obligations de recherche d'emploi), grâce à la date d'inscription et la date d'annulation de la demande (DATINS, DATANN) ;
- le nombre de mois civils en activité réduite courte (NBAR78) ;
- le nombre de mois civils en activité réduite longue (NBAR79).

Elle permet aussi de disposer du motif d'inscription et du motif d'annulation de la demande d'emploi, du niveau de formation, du sexe et de la nationalité du demandeur d'emploi, ainsi que le dernier métier recherché au cours de la demande d'emploi. Les demandes d'emploi de catégorie 4 et de catégorie 5 sont également disponibles (voir tableau 14).

Tableau 14 • Catégorie des demandes d'emplois contenues dans la table DE pour les répondants à la troisième interrogation de l'enquête Génération identifiés de façon univoque dans les fichiers de Pôle emploi

CATREGR	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
1	51 838	74.52	51 838	74.52
2	2 411	3.47	54 249	77.98
3	5 500	7.91	59 749	85.89
4	6 697	9.63	66 446	95.51
5	3 121	4.49	69 567	100.00

Fréquence manquante = 2 242

Nous nous sommes plus particulièrement interressé à la distribution de trois indicateurs quantitatifs calculés sur l'ensemble des dix années couvertes par le FHS, jusqu'en mars 2014 inclus : la durée totale d'inscription dans les categories opérationnelles 1, 2 ou 3 sur ces dix années, le nombre de mois sans activité réduite et le nombre de mois avec une activité réduite. Les résultats détaillés des exploitations réalisées sont présentés dans l'annexe 7 et illustrés par les graphiques 1 et 2 *infra*.

Nous avons ensuite élaboré un calendrier mensuel des mois d'inscription, sur le modèle de celui construit dans l'enquête Génération, sans distinguer l'exercice ou non d'une activité réduite (ce qui aurait pu être fait si nous avions mobilisé la table E0). Les graphiques 3 et 4 *infra* sont issus des exploitations réalisées.

Encadré 3 • Quelques precisions méthodologiques

Comme pour les exploitations présentées dans la partie 4, l'analyse est restreinte au champ des répondants à la troisième interrogation de l'enquête Génération 2004.

Seules les personnes retrouvées de manière univoque dans les fichiers de Pôle emploi sont considérées comme ayant été inscrites au moins une fois sur la fenêtre d'observation couverte par l'extrait du FHS transmis au Céreq. Pour les autres, non identifiées ou identifiées de façon non univoque, la durée d'inscription est mise à zero (l'ensemble des personnes présentes dans la table ID_MULT_BNI est donc considéré comme non inscrit, de même qu'un nombre limité de personnes identifiées dans les tables ID_UNIQUE et ID_MULT).

Par construction, plusieurs identifiants IDTIDV sont associées aux personnes issues de la table ID_MULT. Dans ce cas, tous les identifiants IDTIDV sont pris en compte pour le calcul des durées.

Seules les demandes d'emploi de catégorie 1, 2 ou 3 sont prises en compte pour calculer la durée totale d'inscription. Celle-ci est calculé en « mois civils ». Si une personne s'inscrit au cours du mois de janvier et sort des listes au cours du mois suivant, sa durée d'inscription est alors de deux mois (janvier et février). La durée d'inscription est donc surestimée. Ce choix a été fait parce qu'en cas d'exercice d'une activité réduite, la seule information disponible est le nombre de mois concernés par l'exercice d'une activité réduite. La durée d'inscription sans activité réduite est calculée par soustraction du nombre de mois d'exercice d'une activité réduite à la durée totale d'inscription.

Les résultats n'utilisent pas les pondérations disponibles dans l'enquête Génération 2004.

Les durées moyennes sont d'abord calculées sur l'ensemble des répondants à la troisième interrogation de l'enquête Génération 2004, avec une durée nulle pour les répondants absents des fichiers de Pôle emploi ou retrouvés de façon non univoque (tableaux à gauche dans les pages suivantes). Elles sont ensuite calculées pour les seules personnes inscrites au moins une fois en catégorie 1, 2 ou 3 (tableaux à droite dans les pages suivantes). Les premiers tableaux permettent d'évaluer le degré d'exposition au « chômage administratif » de l'ensemble du groupe de personnes étudiées. Les seconds tableaux donnent la durée du « chômage administratif » réellement subie par les personnes concernées au sein de chaque groupe.

Sans entrer dans le detail de l'ensemble des résultats, les durées d'inscription à Pôle emploi selon le type de trajectoires sur sept ans (dans la typologie élaborée par le Céreq) varient fortement selon le type de trajectoires et cela, de façon conforme à ce à quoi l'on pourrait s'attendre (graphique 1). En particulier :

- les jeunes en situation de chômage persistant et recurrent ont la durée moyenne d'inscription de loin la plus forte (42 mois sur dix ans). Ils sont suivis par les jeunes qui ont décroché de l'emploi (30 mois) ;
- à l'opposé, ceux qui ont accédé à un emploi durable ont des durées moyennes d'inscription beaucoup plus courtes et d'autant plus courtes que l'accès à l'emploi durable a été rapide (4 mois d'inscription en cas d'accès rapide, 9 mois en cas d'accès différé, 15 mois en cas d'accès tardif) ;
- les personnes éloignées durablement du marché du travail ont aussi une durée d'inscription plus réduite (14 mois) ;
- les personnes durablement en emploi à durée déterminée sont dans une situation intermédiaire avec une durée moyenne d'inscription de 25 mois sur dix ans. Ce sont aussi eux qui, en moyenne, ont exercé le plus d'activité réduite.

Le calendrier mensuel d'inscription permet de produire un profil moyen d'inscription pour chacun de ces types de trajectoire définis par le Céreq à partir de l'enquête Génération. Les profils des calendriers mensuels moyens obtenus confortent les résultats déjà évoqués (graphique 3 et 4). En particulier, pour les personnes ayant accédé à un emploi durable, le profil mensuel présente une bosse initiale, dont

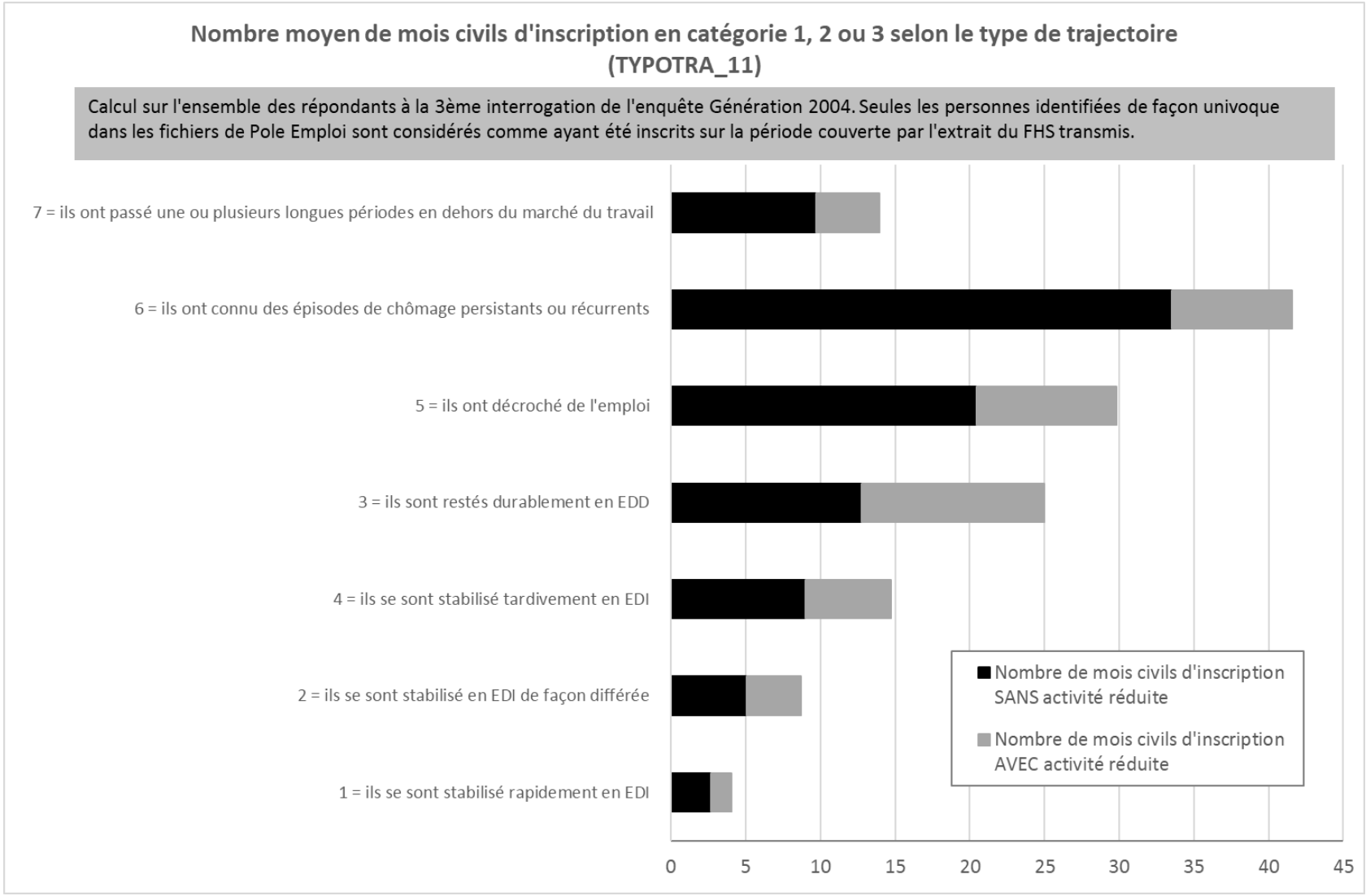
l'ampleur et la largeur est corrélée avec la vitesse de l'accès à l'emploi durable. Pour ceux qui ont décroché de l'emploi, le profil mensuel montre un niveau élevé d'inscription sur toute la période, qui s'accentue en fin de période.

Si l'on s'intéresse aux durées moyenne d'inscription selon le plus haut diplôme (graphique 2), les écarts sont moindres entre les différentes categories mais sont là aussi cohérentes avec les résultats connus par ailleurs. En particulier :

- ce sont les non diplômés qui ont la durée moyenne d'inscription la plus longue (24 mois sur dix ans).
- à l'opposé, les diplômés des écoles d'ingénieurs ou de commerce font partis de ceux qui ont les durées moyennes les plus faibles (7 mois) ;
- entre les deux, les durées moyennes sont *grosso modo* décroissantes avec le niveau de diplôme ;
- à niveau de diplôme professionnel equivalent (CAP, bac professionnel ou licence professionnelle), les diplômés des formations industrielles ont une durée moyenne d'inscription un peu plus faible que les diplômés des formations tertiaires.

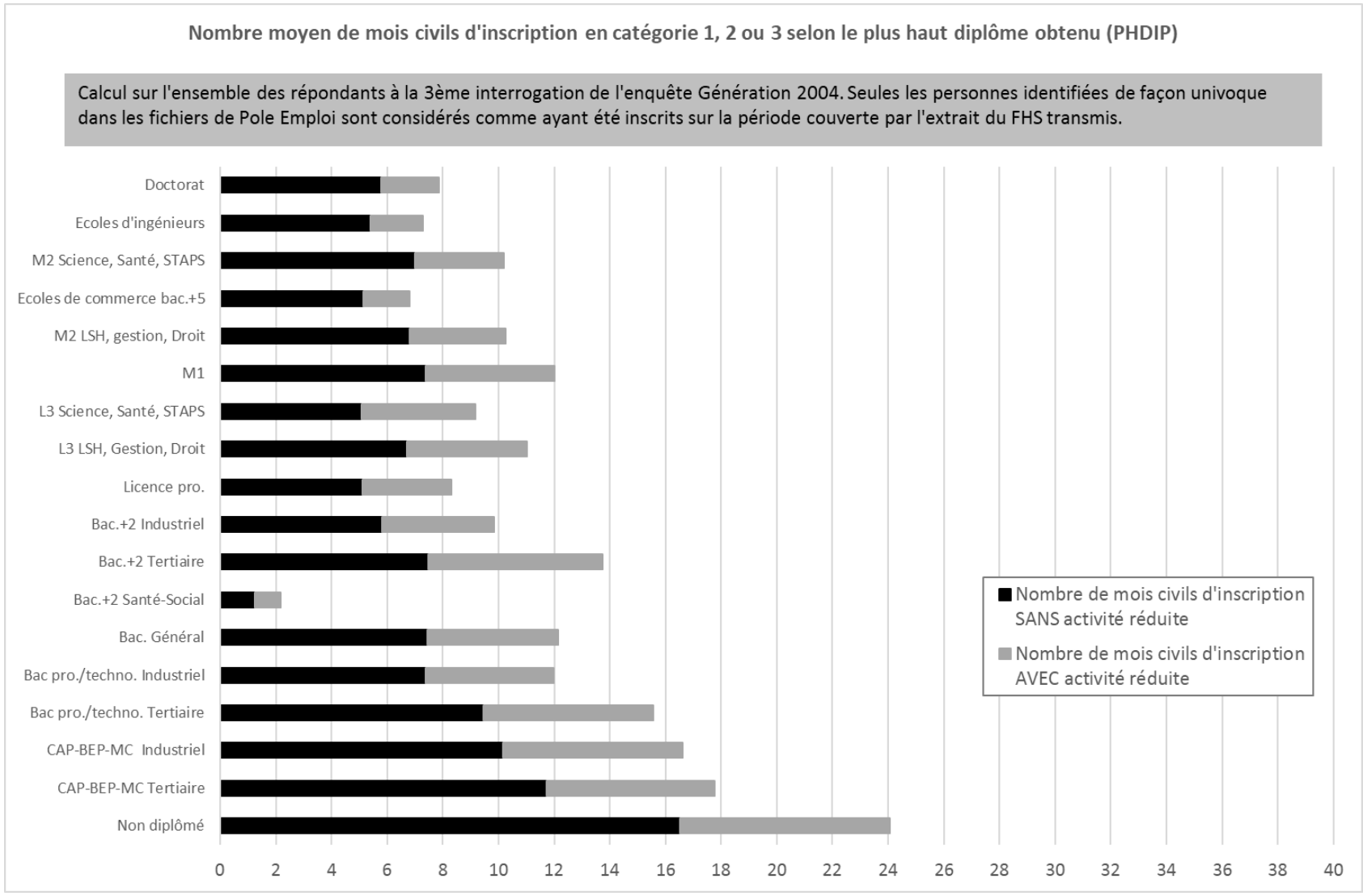
Le lecteur pourra mettre ces résultats en regard avec ceux publiés par le Céreq à partir de l'enquête Génération 2004, notamment dans *Quand l'Ecole est finie... Premiers pas dans la vie active de la Génération 2004* (Cereq, janvier 2008, 86 pages) et dans Mazari Z. & Recotillet I., « Génération 2004 : des débuts de trajectoire durablement marqués par la crise ? » (*Bref*, n° 311, Juillet 2013, Céreq, 4 pages).

Graphique 1 •



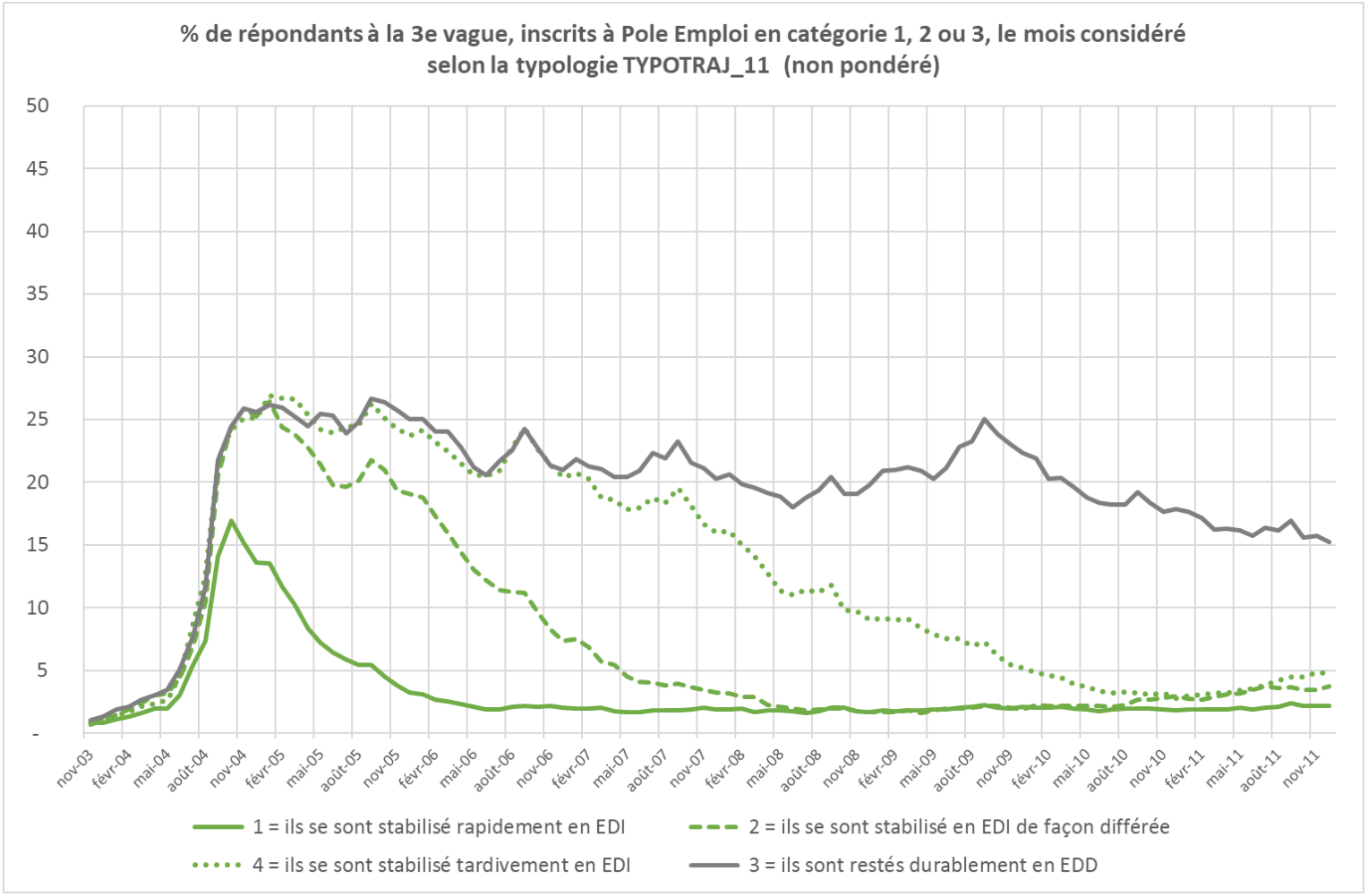
La collection *Working Paper* publie des textes pour engager le débat avec d'autres chercheurs.
La publication n'engage que l'auteur.

Graphique 2 •



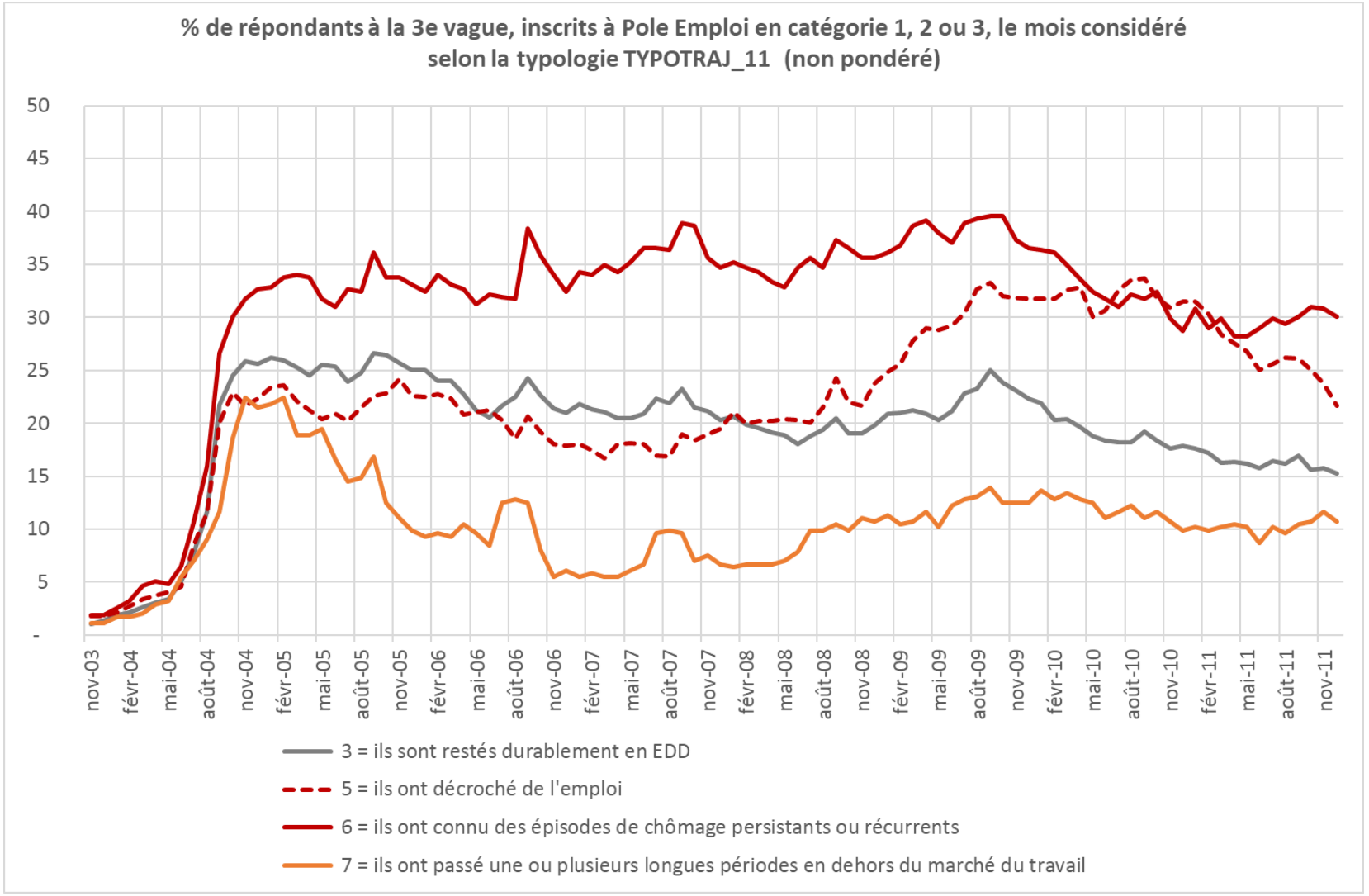
La collection *Working Paper* publie des textes pour engager le débat avec d'autres chercheurs.
La publication n'engage que l'auteur.

Graphique 3 •



La collection *Working Paper* publie des textes pour engager le débat avec d'autres chercheurs.
La publication n'engage que l'auteur.

Graphique 4 •



La collection *Working Paper* publie des textes pour engager le débat avec d'autres chercheurs.
La publication n'engage que l'auteur.

Conclusion

Les travaux exploratoires sur le rapprochement expérimental entre le fichier historique des demandeurs d'emploi et l'enquête Génération permettent de tirer les enseignements suivants :

1° Un nombre important de jeunes répondants à l'enquête Génération a été identifié dans les fichiers de Pôle emploi. La proportion des échos positifs est toujours élévée, quelles que soient les caractéristiques étudiées, même si elle semble aussi corrélée avec les difficultés d'insertion sur le marché du travail. Cette proportion élevée rappelle le rôle important de Pole emploi sur le marché du travail, qui vaut dès les premières années de vie active.

2° Le rapprochement de la table DE avec l'enquête Génération permet de calculer des durées moyennes d'inscription selon le plus haut diplôme, selon le niveau de sortie ou selon le type de trajectoire d'insertion, qui aboutissent à des constats cohérents avec ceux produits par la seule enquête Génération.

3° En conséquence, même si de nombreux jeunes au chômage ne s'inscrivent pas à Pôle emploi, il serait utile de relancer un appariement pour une prochaine enquête Génération en visant les apports mentionnés au 1° et au 4° de la partie 1.4. En particulier cet appariement permettrait d'enrichir l'enquête, d'examiner les apports possibles pour les traitements de la non-réponse et de l'attrition.

4° La pertinence de maintenir dans l'enquête Génération des questions sur l'inscription à Pôle emploi devrait être aussi réexaminée.

5° En lien avec le point précédent et pour les raisons évoquées au 1° et 2° de la partie 1.3, il conviendrait de limiter dans un premier temps cet appariement à un nombre restreint de tables du fichier historique, essentiellement les tables DE et E0 du FHS, à l'exclusion des autres. Il semble en effet plus raisonnable de limiter les ambitions pour s'assurer d'une exploitation effective des données demandées, avant d'envisager un élargissement éventuel ultérieur.

6° Si une nouvelle opération de rapprochement devait être envisagée, il importerait que le Céreq discute et documente de façon précise dès le départ avec Pôle emploi les procédures qui seront mises en œuvre pour identifier les personnes dans les fichiers de Pôle emploi et pour récupérer les informations de ces individus dans le FHS.

7° Toujours si une nouvelle opération de rapprochement devait être envisagée, il conviendrait d'examiner la possibilité d'un recours indirect au NIR, disponible dans les systèmes d'information de Pôle emploi. Son usage impliquerait de collecter la date de naissance complète, le nom à la naissance et le lieu de naissance des personnes dans l'enquête Génération. Il impliquerait de recourir aussi au service d'identification à la BRPP de l'Insee.

8° Il conviendrait d'examiner la question des personnes qui sembles présentes en doublon, avec plusieurs identifiants dans l'enquête Génération.

La collection *Working Paper* publie des textes pour engager le débat avec d'autres chercheurs.
La publication n'engage que l'auteur.

Annexe 1 – Fiche programme n°402

Rapprochement exploratoire de l'enquête Génération avec les données de Pôle emploi

Fiche n° 402 - Unité de rattachement : DEEVA Retour Mettre à jour

Version 2 publiée le 19 septembre 2020 - Fiche validée

COUPPIE Thomas (DEEVA), ILARDI Valérie (DEEVA), JUGNOT Stéphane (DIR)

CHOMAGE ENQUETE METHODOLOGIE D'ENQUETE STATISTIQUE D'EMPLOI ENQUETE GENERATION 2010

OMT 2014-2017 Aucun de ces 7 domaines en particulier
OMT 2019-2022 Entre brouillage des frontières et enjeux de mobilité, comment se construisent les parcours des individus ?

Pourquoi ? (objectifs et hypothèses)

Le travail exploratoire proposé vise à reprendre une piste initiée en 2012 qui n'a pû être menée à son terme en raison de la priorité donnée à d'autres sujets pour le refonte du dispositif Génération. Il s'agira de comparer les séquences de chômage enregistrées par l'enquête génération et celles de Pôle Emploi mais surtout d'examiner les apports possibles de cette source administrative pour le dispositif Génération. De nombreux travaux montrent que la non inscription à Pôle Emploi des personnes se déclarant chômeurs est plus fréquente chez les jeunes qu'à d'autres tranches d'âges. Les données administratives ne pourront donc pas se substituer aux données d'enquête. Néanmoins, elles peuvent constituer un enrichissement sur les aspects de l'accompagnement des jeunes qui y ont recours. Les résultats de ce travail exploratoire doivent permettre une prise en compte de ses conclusions pour l'enquête Génération 2021.

Comment ? (méthode et déroulé)

Le travail d'appariement stricto sensu a déjà été réalisé. Les questions sur les variables d'identification et l'algorithme à utiliser pour décider de rapprocher ou non les observations des deux sources pour un individu présumé identique ne seront donc pas examinés. Dans un premier temps, il s'agira d'examiner la cohérence des séquences d'inscription dans les fichiers de Pôle Emploi (sans activité réduite ou avec activité réduite) et des parcours déclarés dans l'enquête Génération, en fonction des niveaux des diplômes. Dans un second temps, une analyse des informations fournies par Pôle Emploi par niveaux de diplôme permettra d'illustrer les apports possibles de l'appariement.

Annexe 2 - Liste des tables composant le fichier historique administratif en 2011

Source : Pôle emploi, version 3.7 du 10 février 2011 du *Dictionnaire des données du Fichier historique administratif des demandeurs d'emploi* (73 pages).

DE : Table des demandeurs / demandes d'emploi. C'est la table centrale. L'observation est la demande d'emploi. Elle comprend les caractéristiques de la demande (date et motif d'inscription, date et motif d'annulation, emploi recherché, nature du contrat souhaité, …). Elle comprend aussi l'identifiant du demandeur d'emploi et ses caractéristiques (sexe, âge, niveau de formation, bénéficiaire d'une reconnaissance de handicap, du RMI/RSA,…).

Un demandeur d'emploi peut avoir plusieurs demandes d'emploi en cours un même jour[33].

L'identifiant du demandeur d'emploi (IDENT) est non signifiant et incrémenté au sein de chaque zone Assedic. Si le demandeur d'emploi change de zone Assedic, il change d'identifiant.

Les informations retenues sont les plus récentes saisies à la date de production du fichier. Pour certaines variables, l'historique des différentes valeurs prises sur la demande d'emploi est conservée dans des tables dédiées, la valeur la plus récente étant toujours dans la table DE :

- **CATREGR** : Table d'historisation du type d'emploi recherché (nature du contrat et quotité de travail)
- **ROME :** Table d'historisation du métier recherché dans la nomenclature ROME
- **ZUS :** Table d'historisation de l'appartenance à une ZUS (depuis janvier 2007) – le n° zus serait mal renseigné
- **OBLIGEM** : Table d'historisation de reconnaisance d'un handicap
- **RMI :** Table d'historisation des droits au RMI, puis RSA
- **STATUT :** Table d'historisation du statut du demandeur d'emploi (PAP ou PAE, avec numéro d'ordre)
- **STRSUIVP :** Table d'historisation de la structure principale de suivie (depuis octobre 2008, rarement renseigné)
- **STRSUIVD :** Table d'historisation de la structure déléguée de suivie (depuis octobre 2008. Code ALE, co-traitant, opérateur privé…)

E0 : Table des activités réduites. Chaque déclaration d'activité réduite au cours d'un mois donne lieu à une observation dans cette table, avec le nombre d'heures déclarées, le mois d'exercice de l'activité réduite et le mois d'enregistrement de la déclaration (pour une même demande d'emploi, plusieurs activités réduites peuvent être déclarées successivement comme se rapportant au même mois).

D2 : Table des demandeurs d'emploi indemnisable. Cette table donne les informations sur les droits à indemnisation des demandeurs d'emploi (régime, filière, nature de l'allocation, date d'ouverture des droits, date de fin théorique des droits). Le fait d'être indemnisable n'implique pas une indemnisation effective, en particulier en cas d'exercice d'une activité réduite. Les informations pour une période donnée peuvent évoluer à chaque actualisation du fichier. Elles ne sont considérées comme « consolidées » qu'avec un recul d'un an.

DRSA : Table des droits au RSA (pour les nouveaux bénéficiaires à partir de juin 2009).

P4 : Table des parcours (depuis octobre 2006) : type de parcours, date d'entrée et de sortie du parcours et s'il y a lieu, motif de changement de parcours

[33] Par exemple, si la date d'annulation d'une demande est modifiée d'un mois sur l'autre à la suite à une actualisation tardive, la demande annulée demeure dans le fichier et une autre demande restant en cours apparait, avec les mêmes caractéristiques, don't la même date de début.

La collection *Working Paper* publie des textes pour engager le débat avec d'autres chercheurs.
La publication n'engage que l'auteur.

E1CONTACT : Table de contacts (dates des contacts sans entretien physique, pour des transmissions d'informations)

E1ENT : Table des entretiens (hors entretiens PAE)

E1ENTPP : Table des entretiens de suivi mensuel (dont PAE) : offre de service (si pas d'offre de service, non mis ici depuis 2006)

E1PAP : Table des entretiens PAP[34] (dates des entretiens et phases du PAP, mises en œuvre avant la mise en place du PPAE)

E3CONS : Table des actions conseillées (code identifiant le type d'action, date de préconisation, le cas échéant, nature du besoin de formation dans le référentiel Formacode)

E3 : Table des actions réalisées ou non. Quand une action est conseillée, elle est intégrée dans E3CONS. Une fois l'inscription réalisée ou l'action réalisée, abandonnées ou supprimée, l'action disparait de la table E3CONS et bascule dans la table E3, avec des informations complentaires, dont la date d'inscription, la date de début de la prestation et le statut de l'action (inscription, réalisée, abandonnée…).

M0 : Table des mises en relation (date de la mise en relation, nature du contrat proposé, numéro de l'offre d'emploi, résultat de la mise en relation)

P2 : Table des formations des demandeurs d'emploi indemnisés (nature du besoin de formation dans le référentiel Formacode, objectif de la formation, niveau de la formation, dates de début et de fin, nombre d'heures, existence d'une période en entreprise, type de diplôme validé, SIRET de l'entreprise réalisant la prestation)

TRANSFERT : Cette table permet de suivre les changements d'identifiants liés aux changements de zone Assedic (IDENT, date de transfert à autre Assedic, ALE d'origine, ALE de destination, ancien IDENT dans l'ALE d'origine)

[34] Le projet d'action personnalisé est un parcours d'accompagnement du demandeur d'emploi mis en place en 2001 (pour en savoir plus, voir Debauche E., Jugnot S. (2005), « Le projet d'action personnalisé du demandeur d'emploi : un accompagnement individualisé de masse », *Premières synthèses*, Dares, n°30.2, juillet. Le PAP a été remplace par le PPAE, projet personnalisé d'accès à l'emploi, par la loi n°2008-758 du 1er août 2008 relative aux droits et aux devoirs des demandeurs d'emploi, qui introduit la notion d'offre d'emploi « raisonnable ».

Annexe 3 – Extraits du « rapport intermédiaire » relatif à l'« axe 1 : dimensions techniques et méthodologiques » du « groupe de travail sur l'avenir du dispositif Génération »[35]

[...]

Fonctionnement du groupe de travail « Axe 1 »

Le groupe s'est réuni 4 fois au cours du premier semestre auquel s'est ajouté une cinquième réunion de tout le département pour réfléchir au protocole de l'expérimentation de 2016.

Les contributions aux réflexions sur cet axe ont été réalisées par les personnes suivantes : Waida Aboudou (WA), Christophe Barret (CB), Thomas Couppié (TC), Christophe Dzikowski (CD), Céline Goffette (CG), Graziella Marouillat (GM), Zora Mazari (ZM), Boris Ménard (BM), Virginie Mora (VM), Pascale Rouaud (PR), Florence Ryk (FR). Les personnes suivantes ont également participé aux échanges lors des réunions du sous-groupe de travail : Jean-Claude Sigot, Pierre Hallier, Dominique Epiphane et Valentine Henrard. Les contributions sont les suivantes :

Chantiers		Noms
A. Collecte par internet		
A1. Bibliographie Multimode		CD
A2. Autres expériences nationales et internationales		CG
Bilan de collecte	A3. Bilan Cawi Géné2010_3a	CD, CB
	A4. Bilan enquête compétence	BM
A5. Méthodologie	A5a. comparaison Cati/CAwi	VM, CD
	A5b. mesure des effets de satisficing/primacy	WA, EG
	A5c. Analyse des résidents étrangers	WA, EG
A6. Nouveau protocole expérimental 2015		CB
A7. Ergonomie questionnaire WEB		ZM, FR
B. Appariement à des sources administratives		
B1. Appariement FHS		
B2. Appariement DADS		CB, CD
B3. Appariement SISE/APOGEE		BM
B4. Expertise de l'information sur les entreprises		ZM, CB
B5. Mise en place de Sicore embarqué		ZM
C. Organisation		
C1. Externalisation ou Internalisation ?		ZM, BM, FR,
C2. Bilan des coûts Génération		PR, GM
C3. Le multimode est-il rentable ?		CB

[35] Le seul document récupéré est une « version provisoire » du 7 septembre 2015. Aucune version définitive n'a pu être retrouvée, ni auprès du DEEVA, ni sur le réseau.

La collection *Working Paper* publie des textes pour engager le débat avec d'autres chercheurs.
La publication n'engage que l'auteur.

[...]

A. Appariements à des sources administratives

Les appariements à des sources administratives représentent quant à elles une opportunité pour réduire la durée du questionnement (et donc le coût), pour enrichir les données issues de l'enquête (et donc ouvrir de nouvelles perspectives d'études), pour fiabiliser certaines informations (par exemple sur les données entreprises). Trois chantiers, à différents degré d'avancement, sont envisagés : le fichier FHS de Pôle emploi, le fichier DADS, Les fichiers issus d'Apogée ou les fichiers Sise.

a) Appariements FHS

Les données de Génération 2004 (ensemble des répondants de la première vague) ont été appariées (à l'été 2014) avec les données du fichier historique de Pôle emploi sur 7 années. L'exploitation de cet appariement comporte deux volets : une expertise méthodologique et une phase d'études/recherches.

L'expertise méthodologique se décompose en deux étapes :

- La première étape consiste à gérer notamment les échos multiples et à évaluer finement la qualité de l'appariement.
- La seconde étape consiste à comparer les données issues des deux sources et voir à quel point celles-ci sont compatibles. Par exemple, un intérimaire pendant un an avec alternance de mission et de chômage est susceptible d'être perçu très différemment dans les deux sources.

Sur les réinterrogations, les données FHS (qui suit le jeune sur 7 ans) nous permettrait également de mener une analyse sur l'attrition (les chômeurs sont-ils sur ou sous-représentés parmi les non-répondants des vagues 2 et 3).

Pour l'avenir du dispositif, les intérêts de cet appariement sont potentiellement multiples :

- réduire partiellement la durée de questionnement
- consolider certaines informations issues de l'enquête
- obtenir des informations complémentaires à celles recueillis dans l'enquête. Par exemple, les données FHS apportent des éléments sur les formations, les propositions d'emploi, les mises en relation, …
- améliorer le traitement de la non-réponse

Il faut rappeler qu'il n'y a pas d'inscription obligatoire à pôle emploi pour les jeunes. Il serait alors intéressant d'envisager des appariements avec les données des missions locales, de l'Apec, …

Le travail sur ce chantier n'a pas pu être mené dans le cadre de ce groupe de travail. Il devra faire l'objet d'un chantier Deeva dans les mois à venir.

Malgré tout deux éléments de cadrage intéressants peuvent être fournis.
Il y a 33.655 individus recherchés dans les fichiers Pôle-Emploi. On obtient un écho pour 21.344 individus, soit 63,4% de l'échantillon.

En construisant une indicatrice de chômage dans Génération 2004, qui comptabilise le nombre de mois de chômage déclaré (que l'individu soit répondant seulement à 3 ans, seulement à 5 ans ou à 7 ans, on prend sur la période la plus longue), on obtient le résultat suivant. Parmi ceux qui n'ont pas déclaré de chômage (moismax=0) 55% ne sont pas dans les bases Pôle-Emploi et 45% y sont (ayant pu rentrer dans les fichiers Pôle-Emploi en dehors de la fenêtre d'observation). Pour ceux qui ont au moins 5 mois de chômage, on monte à 82%. Reste maintenant à affiner cette indicatrice car un peu de chômage en tout début de vie active ne donne pas forcément des droits à indemnisation, motivation forte pour aller voir Pôle-Emploi.

b) Appariement DADS […]

La collection *Working Paper* publie des textes pour engager le débat avec d'autres chercheurs.
La publication n'engage que l'auteur.

Annexe 4 – Lettre du Cereq à Pôle emploi (février 2012)

Lettre de Frédéric Wacheux, directeur du Céreq, à Stéphane Ducatez, Sous-Directeur de l'Évaluation et des Prévisions - Direction Générale de Pôle emploi, du 21 février 2012 (réf. : DEEVA/2012/07)

Objet : Enrichissement de l'enquête Génération par appariement à une extraction du Fichier Statistique Historique.

Monsieur,

J'ai l'honneur de vous soumettre une demande d'autorisation d'utilisation des informations issues des fichiers historiques de Pôle emploi pour enrichir le fichier individuel de l'enquête Génération du CEREQ réalisée en 2007 auprès des jeunes sortis de formation initiale en 2003-2004.

Les enquêtes d'insertion du CEREQ sont constitutives du dispositif « Génération » et sont réalisées en conformité de la loi n°51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques. Leur objectif premier est d'étudier les conditions d'accès à l'emploi et les premiers pas dans la vie active des jeunes issus de tous les niveaux de formation du système éducatif.

Dans le cadre de cette demande, Il s'agit plus précisément de la cohorte constituant la Génération des sortants du système éducatif en 2004, enquêtée une première fois en 2007 puis à nouveau en 2009 et 2011. Chacune de ces interrogations a fait l'objet d'une déclaration auprès de la CNIL : récépissé n°1210067 du 19 janvier 2007, récépissé n°1342783 du 11 février 2009 et récépissé n°1519175 du 8 juillet 2011.

Le rapprochement que nous souhaiterions pouvoir réaliser avec le FHS vise un double objectif. Le premier objectif est relatif à une expérimentation dans la perspective de l'enquête 2013 auprès de la « génération 2010 » permettant de réduire à terme le questionnaire posé aux jeunes interrogés et de satisfaire ainsi aux exigences du Cnis en matière de redondance d'information collectée. Le second objectif vise à enrichir l'analyse du lien entre la sortie du chômage pour les jeunes et l'accès à l'emploi. Cela permettrait en particulier de voir si le dispositif « Génération » est adapté pour le repérage et le suivi des trajectoires des jeunes connaissant une forte précarité, avec une alternance de périodes d'emplois courts et de chômage (ici limité au chômage « administratif »). Il permettrait aussi d'étudier l'effet sur les parcours des jeunes du recours au système public de l'emploi avec une information enrichie.

Des contacts pris avec la sous direction des enquêtes et des prévisions, il ressort qu'un tel appariement serait fructueux dans près de 85 % des cas.

Les objectifs et les conditions de mise en œuvre sont présentés de façon plus précise en annexe.

Je vous remercie de l'intérêt que vous porterez à la demande du Céreq et vous prie de recevoir, Monsieur, l'expression de mes très sincères salutations.

Le Directeur du Céreq

Annexe – présentation du traitement

A. Présentation synthétique du projet

Le traitement consiste, en collaboration avec Pôle emploi, à **enrichir les résultats des enquêtes du CEREQ auprès des jeunes sortis de formation initiale en 2003-2004 par des informations issues du fichier historique statistique de Pôle emploi.**

Les enquêtes du CEREQ concernées par ce projet relèvent du dispositif « Génération » et sont donc réalisées dans le cadre de **la loi n°51-711 du 7 juin 1951** sur l'obligation, la coordination et le secret en matière de statistiques. Le dispositif « Génération » a pour objectif d'étudier les conditions d'accès à l'emploi et les premiers pas dans la vie active des jeunes, en fonction de la formation initiale suivie.

Le rapprochement concernerait plus précisément la cohorte constituant la « Génération 2004 » des jeunes sortis de formation initiale au cours ou à l'issue de l'année scolaire 2003-2004, soit un échantillon d'environ 35 000 répondants, représentatifs des 750 000 jeunes sortis de formation initiale cette année là.

Ces jeunes, diplômés ou non, ont été enquêtés une première fois en 2007 puis à nouveau en 2009 et 2011. Chacune de ces interrogations a fait l'objet d'une déclaration auprès de la CNIL : récépissé n°1210067 du 19 janvier 2007, récépissé n°1342783 du 11 février 2009 et récépissé n°1519175 du 8 juillet 2011.

Le rapprochement envisagé a un **double objectif** :

- Une expérimentation dans la perspective de l'enquête 2013 auprès de la « Génération 2010 ». Dans le **cadre de la RGPP**, il s'agirait de tester une mobilisation de sources administratives permettant de diminuer les coûts de collecte du dispositif «Génération » tout en gagnant en qualité.
- la réalisation d'études : les informations appariées, très riches en matière de profil scolaire des jeunes et d'expériences d'emploi permettraient **d'étudier de façon très ciblée** l'effet différentiel du recours au système public de l'emploi sur les processus de stabilisation dans l'emploi.

Le rapprochement envisagé **s'effectuerait sur la base du nom, du prénom, du mois et de l'année de naissance, du sexe.** Les informations relatives au lieu de résidence figurant dans les deux sources sont également utilisables au niveau département.

B. Présentation du dispositif « Génération » : un dispositif unique pour étudier les premiers pas dans la vie active, selon la formation initiale suivie

A la fin des années quatre-vingt-dix, le Céreq a mis en place un dispositif d'enquêtes original qui permet d'étudier l'accès à l'emploi des jeunes à l'issue de leur formation initiale.

Tous les trois ans, une nouvelle enquête est réalisée auprès de jeunes qui ont en commun d'être sortis du système éducatif la même année quel que soit le niveau ou le domaine de formation atteint, d'où la notion de « génération ».

L'enquête permet de reconstituer les parcours des jeunes au cours de leurs trois premières années de vie active et d'analyser ces parcours au regard notamment du parcours scolaire et des diplômes obtenus. Certaines cohortes sont interrogées plusieurs fois pour suivre les débuts de carrière.

En s'appuyant sur un calendrier qui décrit mois par mois la situation des jeunes et sur des informations plus précises concernant le premier emploi et l'emploi occupé au bout de trois années passées sur le marché du travail, ce dispositif permet non seulement d'analyser les trajectoires d'entrée dans la vie active mais aussi de distinguer, d'une génération à l'autre, les aspects structurels et conjoncturels de l'insertion.

Les cohortes de jeunes sortis de formation initiale en 1992, 1998, 2001, 2004 et 2007 ont ainsi été interrogées. La prochaine cohorte enquêtée concernera, en 2013, les jeunes sortis au cours ou à l'issue de l'année scolaire 2009-2010.

Le présent projet concerne la cohorte constituant la génération de jeunes sortis de formation initiale au cours ou à l'issue de l'année scolaire 2003-2004. 65 000 jeunes avaient été enquêtés lors de la première interrogation, une partie avec un questionnaire complet, une autre avec un questionnaire réduit. Le présent projet concernera uniquement les jeunes concernés par le questionnaire complet,

soit **un échantillon d'environ 35 000 répondants à la première interrogation, en 2007. Ils sont représentatifs des 750 000 jeunes sortis de formation initiale cette année là.** Ils ont ensuite été réinterrogés en 2009 et 2011, permettant ainsi de couvrir les premiers pas dans la vie active au cours des sept années suivant la sortie du système éducatif. Le projet concernera tous les jeunes ayant répondu à la première interrogation et concernés par les interrogations de 2009 et 2011, qu'ils aient répondu ou non à ces deux dernières interrogations.

Chacune de ces interrogations a fait l'objet d'une déclaration auprès de la CNIL :
- Récépissé n°1210067 du 19 janvier 2007, pour l'enquête de 2007 ;
- Récépissé n°1342783 du 11 février 2009, pour l'enquête de 2009 ;
- Récépissé n°1519175 du 8 juillet 2011, pour l'enquête de 2011.

C. Finalités

C.1 - Finalités générales des enquêtes du dispositif « Génération »

Conformément aux dossiers déposés auprès de la CNIL, les enquêtes du dispositif « Génération » ont pour finalité :

- D'une part, la production d'indicateurs statistiques : Il s'agit de constituer et diffuser des repères statistiques sur l'insertion sur le marché du travail des jeunes à l'issue de leur formation initiale, en fonction de leur niveau de diplôme et de leur type de formation, en tenant compte également de leurs caractéristiques individuelles (origine sociale, lieu de résidence, origine nationale, situation maritale, sexe…). Plusieurs dimensions de la qualité de l'insertion sont prises en compte : situation d'activité, nature du premier emploi (type de contrat, niveau de rémunération, quotité de temps de travail, profession, secteur d'activité, taille de l'entreprise), nature de l'emploi trois ans après la sortie, récurrence du chômage sur la période, vitesse d'accès à l'emploi…

Des premiers résultats sont publiés dans un délai relativement court après l'interrogation dans une perspective d'information des acteurs, d'évaluation du système éducatif et d'aide à la décision. Ces indicateurs sont présentés par niveau de diplôme à l'issue des diverses filières de formations. Les nomenclatures de diplômes et de formation peuvent être plus ou moins détaillées, selon l'échantillon disponible. Sur l'ensemble du champ, la NSF (Nomenclature des Spécialités de Formation) est mobilisée. Pour les sortants d'université, la nomenclature SISE est également utilisée.

Des indicateurs éclairent aussi l'insertion dans des professions et dans des secteurs. Pour l'approche par profession détaillée, on aura recours, par exemple, à la nomenclature FAP des Familles d'Activité Professionnelle et à la PCS.

Ces indicateurs font l'objet de publications, notamment les collections du Céreq, sur le site internet du Céreq et dans celles des partenaires d'enquêtes.

- D'autre part, contribuer à la compréhension des processus d'insertion et des différenciations des parcours en début de carrière : Il s'agit ici notamment d'analyser les formes que prennent les parcours sur le marché du travail, avec une stabilisation plus ou moins rapide en emploi, des décrochages éventuels, des réorientations, etc. Au-delà du rôle que peut jouer la formation initiale, l'enquête permet de tester les effets des expériences de travail en cours d'études, l'importance de certains réseaux sociaux pour l'accès aux premiers emplois. Plusieurs dimensions des parcours professionnels sont examinées comme la correspondance formation-emploi, la dimension territoriale, la mobilité externe… L'enquête permet aussi d'étudier les différenciations de parcours en lien avec le sexe, les origines sociales ou nationales, le fait d'habiter un quartier de politique de la ville.

Ces travaux d'études et de recherche supposent la mise à disposition de fichiers statistiques individuels, anonymes et respectant le secret statistique au sens de la loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques.

C.2 - <u>Finalités spécifiques du projet d'enrichissement avec des informations des fichiers historiques de Pôle emploi</u>

Le projet à un double objectif :

1) Expérimenter ce rapprochement dans la perspective d'un projet plus large envisagé dans le cadre de l'enquête 2013 auprès de la « génération 2010 ».

La question posée ici est de voir si, à terme, les informations administratives peuvent permettre de réduire la taille du questionnaire et d'améliorer la qualité de l'information en évitant les effets de mémoire. Une **réduction des coûts de collecte serait alors possible**. Le projet présenté dans le présent dossier permettrait d'amorcer la réflexion, sans attendre la réalisation de l'enquête 2013.

2) Permettre de réaliser des études sur le recours au service public de l'emploi et évaluer ses effets sur les parcours dans les premières années de la vie active.

L'enrichissement permettrait en effet de disposer d'informations précises, complémentaires aux questionnaires des trois interrogations de la « Génération 2004 » : les périodes d'inscription à Pôle emploi et la nature de l'accompagnement proposé par Pôle emploi. Le dispositif «génération » pourra alors être mobilisé pour évaluer sur une cohorte de jeunes débutants **l'incidence des dispositifs du Service Public de l'Emploi**, notamment dans leur capacité à faciliter la transition entre le système éducatif et le marché du travail.

D) Les catégories de données traitées

D.1) Informations issues des enquêtes « Génération »

Les questionnaires de l'enquête « Génération » abordent le parcours scolaire et le parcours professionnel au cours des premières années suivant la sortie de formation initiale, notamment :

- *Le parcours scolaire* : retard ou non à l'entrée en 6e, nature de la classe de 3e et de la classe suivie après la troisième, orientations successives, diplômes obtenus, stages effectués, spécialités fines de formation, migrations en cours d'études, …

- *Les caractéristiques individuelles et la situation familiale* : pays de naissance des parents de l'enquêté, pays de naissance de l'enquêté (France, Etranger), situation d'activité et situation professionnelle des parents à la date de fin d'études ; décohabitation et vie en couple ou non de l'enquêté, résidences successives,…

- *Le calendrier professionnel* : calendrier des séquences d'emploi et de non-emplois sur la base d'un calendrier mensuel déclaratif, portant sur les trois années allant de la sortie de formation initiale jusqu'à la date d'enquête. Il permet de distinguer six situations exclusives l'une de l'autre : deux situations d'emploi (en intérim ou sous contrat direct) ; une situation de recherche d'emploi ; une situation de reprise d'études ; une situation de formation et une situation regroupant les autres cas assimilée à de l'inactivité.

- *Les emplois occupés :* adresse de l'établissement, secteur d'activité, taille, nature et PCS de l'emploi occupé, nature juridique du contrat de travail, temps de travail, salaire perçu, mode d'accès à l'emploi, origine de l'embauche, origine de la fin du contrat,…

D.2) Informations issues de Pôle emploi

Les informations récupérées auprès de Pôle emploi porteraient sur :

- *Les périodes d'inscription à pôle emploi* : catégorie de demandeur d'emploi, ROME de l'emploi recherché, nature de l'accompagnement, formations prescrites et suivies, financeurs de ces formations, mises en relations avec les entreprises et résultats de ces mises en relation, bénéfice du RMI RSA, revenus d'allocations, périodes d'indemnisations, montants de l'indemnité et dispositif financeur.

D.3) Durée de conservation

- Le nom et le prénom (informations nécessaires au rapprochement de fichier) : ces informations seront détruites dans le cadre de ce traitement trois mois après le rapprochement et, au plus tard le 31 décembre 2012.

- Les autres informations, constitutives du fichier d'études, seront conservées sans limitation de durée.

E) Mise en œuvre de l'enrichissement

L'appariement pourrait s'effectuer de la manière suivante :

Étape 1 - création par le Cereq (équipe gestion d'enquêtes du Département des Entrées et Evolutions dans la Vie Active, DEEVA), d'une table d'identification contenant les informations suivantes : identifiant non signifiant différent de l'identifiant non signifiant utilisé dans les fichiers de résultats des enquêtes de la « Génération 2004» ; nom ; prénoms ; date de naissance ; sexe ; département de résidence (avec conservation d'une copie sécurisée de la table au sein de l'équipe gestion d'enquête).

Étape 2 – Transmission à Pôle emploi de la table d'identification selon les modalités prescrites par Pôle emploi et validées par la CNIL (échanges sécurisés de données cryptées sur serveurs FTP, échange physique d'un CDROM ou d'une clé USB,…).

Étape 3 – Appariement, traitement des doublons et récupération par Pôle emploi des informations destinées à enrichir les fichiers de résultats des enquêtes de la « Génération 2004 » pour les individus de la table d'identification.

Étape 4 – Transmission au Céreq (équipe gestion d'enquête du DEEVA) de la table d'informations issues de Pôle emploi indexée sur l'identifiant non signifiant d'origine, selon des modalités validées par la CNIL.

Étape 5 – Création par l'équipe gestion d'enquêtes du DEEVA d'une table « Pôle emploi » substituant à l'identifiant non signifiant utilisé pour les échanges avec Pôle emploi l'identifiant non signifiant utilisé dans les fichiers de résultats des enquêtes de la « Génération 2004 ».

Étape 6 – Destruction par Pôle emploi de la table d'identification et de la table transmise en retour au Céreq, au plus tard le 31 décembre 2012. Destruction de la copie de la table d'identification par le Céreq 31 décembre 2012.

F) Sécurités et secret

Les bâtiments du Céreq sont doublement protégés par un système de gardiennage et de verrouillage des accès (ascenseurs et escaliers), en dehors des heures de travail (nuit, fin de semaine, jours fériés).

Les postes informatiques individuels sont protégés par un mot de passe individuel.

Les agents du Céreq sont soumis au secret professionnel.

Au sein du Céreq, les fichiers *détail complet* des enquêtes « Génération » et, lors des phases de réalisation des enquêtes qui le nécessitent, les fichiers nominatifs, sont stockés sur des serveurs localisés dans des locaux protégés et dont l'accès est strictement limité aux personnes responsables du service informatique. Ils sont accessibles par réseau aux seuls ayant droit.

L'équipe de gestion d'enquête est sensibilisée au cadre législatif de la loi « Informatique et Libertés ». La liste des personnes composant l'équipe de gestion d'enquêtes sera actualisée au moins une fois par an auprès du service informatique responsable de la gestion des droits d'accès.

G) Catégories de destinataires

Fichiers non anonymes :

La table d'identification est conservée et échangée entre l'équipe de gestions d'enquêtes du DEEVA, au Céreq, et le service informatique en charge de l'appariement à Pôle emploi, (cf. supra chapitre E). A l'issue, des traitements, et au plus tard le 31 décembre 2012, cette table est détruite.

Fichiers enrichis :

Les fichiers des enquêtes « Génération 2004 » enrichis de la table des informations issues de Pôle emploi sont destinés aux chargés d'études du Céreq (y compris ceux en poste dans ses centres associés). Une convention peut prévoir une mise à disposition d'autres partenaires du service public (Pôle emploi, DARES) ou des laboratoires de recherches)

Annexe 5 – Complément d'informations transmises par le Céreq à la CNIL (décembre 2012)

CÉREQ

Enrichissement de l'enquête Génération par appariement à une extraction du Fichier Historique de Pôle emploi

Complément d'information à la demande d'autorisation n°1636140 en ligne du 4 décembre 2012

04/12/2012

La collection *Working Paper* publie des textes pour engager le débat avec d'autres chercheurs.
La publication n'engage que l'auteur.

A. .....Présentation synthétique du projet 54

B. Présentation du dispositif « Génération » : un dispositif unique pour étudier les premiers pas dans la vie active, selon la formation initiale suivie 55

C. .....Finalités 55

1) Finalités générales des enquêtes du dispositif Génération..... 55

2) Finalités spécifiques du projet d'enrichissement avec des informations des fichiers historiques de Pôle emploi 56

D. .....Les catégories de données traitées ..... Erreur ! Signet non défini.

1) Informations issues des enquêtes « Génération » ..... **Erreur ! Signet non défini.**

2) Informations issues de Pôle emploi..... **Erreur ! Signet non défini.**

3) Durée de conservation 57

E. ..... Mise en œuvre de l'enrichissement 58

F..... Sécurité et secret 60

G. ..... Catégories de destinataires 60

A. Présentation synthétique du projet

Le traitement consiste, en collaboration avec Pôle emploi, à **enrichir les résultats des enquêtes du CEREQ** (enquêtes auprès des jeunes sortis de formation initiale) **par des informations issues du fichier historique statistique et du fichier historique administratif de Pôle emploi.**

Les enquêtes du CEREQ concernées par ce projet relèvent du dispositif « Génération » et sont donc réalisées dans le cadre de **la loi n°51-711 du 7 juin 1951** sur l'obligation, la coordination et le secret en matière de statistiques. Le dispositif « Génération » a pour objectif d'étudier les conditions d'accès à l'emploi et les premiers pas dans la vie active des jeunes, en fonction de la formation initiale suivie.

Le premier traitement concernera plus précisément la cohorte constituant la Génération 2004 des jeunes sortis de formation initiale au cours ou à l'issue de l'année scolaire 2003-2004, soit un échantillon d'environ 35 000 répondants, représentatifs des 750 000 jeunes sortis de formation initiale cette année-là.

Ces jeunes, diplômés ou non, ont été enquêtés une première fois en 2007 puis à nouveau en 2009 et 2011. Chacune de ces interrogations a fait l'objet d'une déclaration auprès de la CNIL : récépissé n°1210067 du 19 janvier 2007, récépissé n°1342783 du 11 février 2009 et récépissé n°1519175 du 8 juillet 2011.

Cette demande d'autorisation concerne chaque nouvelle enquête Génération à partir de la prochaine interrogation qui aura lieu au printemps 2013.

Calendrier des enquêtes à venir :

	Sortie du système éducatif	Première interrogation	Deuxième interrogation	Troisième interrogation
Enquête Génération 2004	2004	2007	2009	2011
Enquête Génération 2007	2007	2010		
Enquête Génération 2010	2010	2013	2015	2017
Enquête Génération 2013	2013	2016		
Enquête Génération 2016	2016	2019	2021	2023

Enquête non concernée par l'enrichissement.

Le rapprochement a un **double objectif** :

- Une expérimentation dans la perspective des futures enquêtes Génération. Dans le cadre de la RGPP, il s'agirait de tester une mobilisation de sources administratives permettant de diminuer les coûts de collecte du dispositif «Génération » tout en gagnant en qualité.

- La réalisation d'études : les informations appariées, très riches en matière de profil scolaire des jeunes et d'expériences d'emploi permettraient d'étudier de façon très ciblée l'effet différentiel du recours au système public de l'emploi sur les processus de stabilisation dans l'emploi.

Le rapprochement envisagé **s'effectuera sur la base du nom, du prénom, du mois et de l'année de naissance.**

B. Présentation du dispositif « Génération » : un dispositif unique pour étudier les premiers pas dans la vie active, selon la formation initiale suivie

A la fin des années quatre-vingt-dix, le Céreq a mis en place un dispositif d'enquêtes original qui permet d'étudier l'accès à l'emploi des jeunes à l'issue de leur formation initiale.

Tous les trois ans, une nouvelle enquête est réalisée auprès de jeunes qui ont en commun d'être sortis du système éducatif la même année quel que soit le niveau ou le domaine de formation atteint.

L'enquête permet de reconstituer les parcours des jeunes au cours de leurs trois premières années de vie active et d'analyser ces parcours au regard notamment du parcours scolaire et des diplômes obtenus. Certaines cohortes sont interrogées plusieurs fois pour suivre les débuts de carrière.

En s'appuyant sur un calendrier qui décrit mois par mois la situation des jeunes et sur des informations plus précises concernant l'ensemble des périodes d'emploi occupées sur le marché du travail, ce dispositif permet non seulement d'analyser les trajectoires d'entrée dans la vie active mais aussi de distinguer, d'une enquête Génération à l'autre, les aspects structurels et conjoncturels de l'insertion.

Les cohortes de jeunes sortis de formation initiale en 1992, 1998, 2001, 2004 et 2007 ont ainsi été interrogées. La prochaine cohorte enquêtée concernera, en 2013, les jeunes sortis au cours ou à l'issue de l'année scolaire 2009-2010.

Le premier traitement faisant l'objet d'une demande d'accès à des données à caractère personnel auprès du Cnis concerne la cohorte constituant la génération de jeunes sortis de formation initiale au cours ou à l'issue de l'année scolaire 2003-2004. 65 000 jeunes avaient été enquêtés lors de la première interrogation, une partie avec un questionnaire complet (35 000), une autre avec un questionnaire réduit (30 000). Le présent projet concernera uniquement les jeunes concernés par le questionnaire complet, soit **un échantillon d'environ 35 000 répondants à la première interrogation, en 2007.** Ils sont représentatifs des 750 000 jeunes sortis de formation initiale cette année-là. Ils ont ensuite été réinterrogés en 2009 et 2011, permettant ainsi de couvrir les sept premières années suivant la sortie du système éducatif. Le projet concernera tous les jeunes ayant répondu à la première interrogation et concernés par les interrogations de 2009 et 2011, qu'ils aient répondu ou non à ces deux dernières interrogations.

C. Finalités

1) Finalités générales des enquêtes du dispositif Génération

Conformément aux déclarations déposées ou inscrites au registre (cil nommé en mai 2012) auprès de la CNIL, les enquêtes du dispositif Génération ont pour finalité :

- D'une part, la production d'indicateurs statistiques : Il s'agit de constituer et diffuser des repères statistiques sur l'insertion sur le marché du travail des jeunes à l'issue de leur formation initiale, en fonction de leur niveau de diplôme et de leur type de formation, en tenant compte également de leurs caractéristiques individuelles (origine sociale, lieu de résidence, origine nationale, situation maritale, sexe…). Plusieurs dimensions de la qualité de l'insertion sont prises en compte : situation d'activité, nature du premier emploi (type de contrat, niveau de rémunération, quotité de temps de travail,

profession, secteur d'activité, taille de l'entreprise), nature de l'emploi trois ans après la sortie, récurrence du chômage sur la période, vitesse d'accès à l'emploi…

Des premiers résultats sont publiés dans un délai relativement court après l'interrogation dans une perspective d'information des acteurs, d'évaluation du système éducatif et d'aide à la décision. Ces indicateurs sont présentés par niveau de diplôme à l'issue des diverses filières de formations. Les nomenclatures de diplômes et de formation peuvent être plus ou moins détaillées, selon l'échantillon disponible. Sur l'ensemble du champ, la NSF (Nomenclature des Spécialités de Formation) est mobilisée. Pour les sortants d'université, la nomenclature SISE est également utilisée.

Des indicateurs éclairent aussi l'insertion dans des professions et dans des secteurs. Pour l'approche par profession détaillée, on aura recours, par exemple, à la nomenclature FAP des Familles d'Activité Professionnelle et à la PCS.

Ces indicateurs font l'objet de publications, notamment les collections du Céreq, sur le site internet du Céreq et dans celles des partenaires d'enquêtes.

- D'autre part, contribuer à la compréhension des processus d'insertion et des différenciations des parcours en début de carrière : Il s'agit ici notamment d'analyser les formes que prennent les parcours sur le marché du travail, avec une stabilisation plus ou moins rapide en emploi, des décrochages éventuels, des réorientations, etc. Au-delà du rôle que peut jouer la formation initiale, l'enquête permet de tester les effets des expériences de travail en cours d'études, l'importance de certains réseaux sociaux pour l'accès aux premiers emplois. Plusieurs dimensions des parcours professionnels sont examinées comme la correspondance formation-emploi, la dimension territoriale, la mobilité externe… L'enquête permet aussi d'étudier les différenciations de parcours en lien avec le sexe, les origines sociales ou nationales, le fait d'habiter un quartier de politique de la ville.

Ces travaux d'études et de recherche supposent la mise à disposition de fichiers statistiques individuels, anonymes et respectant le secret statistique au sens de la loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques.

2) Finalités spécifiques du projet d'enrichissement avec des informations des fichiers historiques de Pôle emploi

L'enrichissement permettrait en effet de disposer d'informations précises, complémentaires aux questionnaires des interrogations des enquêtes Génération : les périodes d'inscription à Pôle emploi et la nature de l'accompagnement proposé par Pôle emploi. Le dispositif Génération pourra alors être mobilisé pour évaluer sur une cohorte de jeunes débutants l'incidence des dispositifs du Service Public de l'Emploi, notamment dans leur capacité à faciliter la transition entre le système éducatif et le marché du travail.

D. Les catégories de données traitées

1) Informations issues des enquêtes « Génération »

Les questionnaires de l'enquête Génération abordent le parcours scolaire et le parcours professionnel au cours des premières années suivant la sortie de formation initiale, notamment :

- *Le parcours scolaire* : retard ou non à l'entrée en 6e, nature de la classe de 3e et de la classe suivie après la troisième, orientations successives, diplômes obtenus, stages effectués, spécialités fines de formation, migrations en cours d'études, …

- *Les caractéristiques individuelles et la situation familiale* : pays de naissance des parents de l'enquêté, pays de naissance de l'enquêté (France, Etranger), situation d'activité et situation professionnelle des parents à la date de fin d'études ; décohabitation et vie en couple ou non de l'enquêté, résidences successives,…

- *Le calendrier professionnel* : calendrier des séquences d'emploi et de non-emploi sur la base d'un calendrier mensuel déclaratif, portant sur les trois années allant de la sortie de formation initiale jusqu'à la date d'enquête. Il permet de distinguer six situations exclusives l'une de l'autre : deux situations d'emploi (en intérim ou sous contrat direct) ; une situation de recherche d'emploi ; une situation de reprise d'études ; une situation de formation et une situation regroupant les autres cas assimilés à de l'inactivité.

- *Les emplois occupés :* adresse de l'établissement, secteur d'activité, taille, nature et PCS de l'emploi occupé, nature juridique du contrat de travail, temps de travail, salaire perçu, mode d'accès à l'emploi, origine de l'embauche, origine de la fin du contrat,…

- *Les périodes de non-emploi (chômage, inactivité, formation, reprise d'études)* : inscription à Pôle emploi, démarches de recherche d'emploi, formation la plus longue suivie, objectif de la reprise d'études à temps plein, diplôme préparé en reprise d'études,…

2) Informations issues de Pôle emploi

Les informations obtenues auprès de Pôle emploi porteraient sur les périodes d'inscription à Pôle emploi : catégorie de demandeur d'emploi, ROME de l'emploi recherché, nature de l'accompagnement, formations prescrites et suivies, financeurs de ces formations, mises en relations avec les entreprises et résultats de ces mises en relation, bénéfice du RMI RSA, revenus d'allocations, périodes d'indemnisations, montants de l'indemnité et dispositif financeur.

Elles seront extraites de trois tables SAS du FHA et des tables du FHS (à l'exception de quelques variables indirectement nominatives).

Extraction du FHA :

- Table E3 : table de toutes les actions (conseillées, réalisées ou non)
- Table M0 : table des mises en relations (positives ou non)
- Table P2 : table des formations (sauf variable SIRET)

Extraction du FHS :

- Table DE : Table des demandeurs – demandes (sauf variable DEPCOM)
- -table D2 : table des indemnisations
- Table E0 : Table des activités réduites (rattachées à des demandes en catégorie 123678 uniquement)
- Table PAP : Table des entretiens PAP (depuis juillet 2001 et uniquement pour des entretiens rattachés à des demandes 123678)
- Table PARCOURS : Tables des parcours des demandeurs d'emploi
- Table STATDE : Table de statistiques individuelles sur le demandeur (sauf variable REGION)
- Table RSA : table des droits RSA (sauf variable REGION)
- Tables des variables historicisées (sauf variables REGION et NUMZUS)

3) Durée de conservation

- Le nom, prénom, mois et année de naissance (informations nécessaires au rapprochement de fichier) : ces informations seront détruites dans le cadre de ce traitement au plus tard un mois après le rapprochement.

- Les autres informations anonymisées à des fins statistiques, constitutives du fichier d'études, seront conservées dans la limite de conservation requise par le comité du secret (2 ans pour les données ménages).

E. Mise en œuvre de l'enrichissement

L'appariement s'effectuera de la manière suivante :

Étape 1 - création par le Cereq (équipe gestion d'enquêtes du Département des Entrées et Evolutions dans la Vie Active, DEEVA), d'une table d'identification contenant les informations suivantes : identifiant non signifiant différent de l'identifiant non signifiant utilisé dans les fichiers de résultats des enquêtes de la Génération 2004; nom ; prénoms ; date de naissance.

Étape 2 – Transmission à Pôle emploi de la table d'identification selon les modalités prescrites par Pôle emploi et validées par la CNIL (échanges sécurisés de données cryptées sur serveurs FTP, échange physique d'un CDROM ou d'une clé USB,…).

Étape 3 – Appariement, traitement des doublons et récupération par Pôle emploi des informations destinées à enrichir les fichiers de résultats des enquêtes de la Génération 2004 pour les individus de la table d'identification.

Étape 4 – Transmission au Céreq (équipe gestion d'enquête du DEEVA) de la table d'informations issue de Pôle emploi indexée sur l'identifiant non signifiant d'origine, selon des modalités validées par la CNIL.

Étape 5 – Création par l'équipe gestion d'enquêtes du DEEVA d'une table « Pôle emploi » substituant à l'identifiant non signifiant utilisé pour les échanges avec Pôle emploi l'identifiant non signifiant utilisé dans les fichiers de résultats des enquêtes de la « Génération 2004 ».

Étape 6 – Destruction par Pôle emploi de la table d'identification et de la table transmise en retour au Céreq, au plus tard un mois après le rapprochement. Destruction de la copie de la table d'identification par le Céreq au plus tard un mois après le rapprochement.

Le schéma 1 présente les étapes 1 à 6.
Le schéma 2 présente le circuit des données identifiées, rapprochables à des données à caractère personnel.

Schéma 1 :

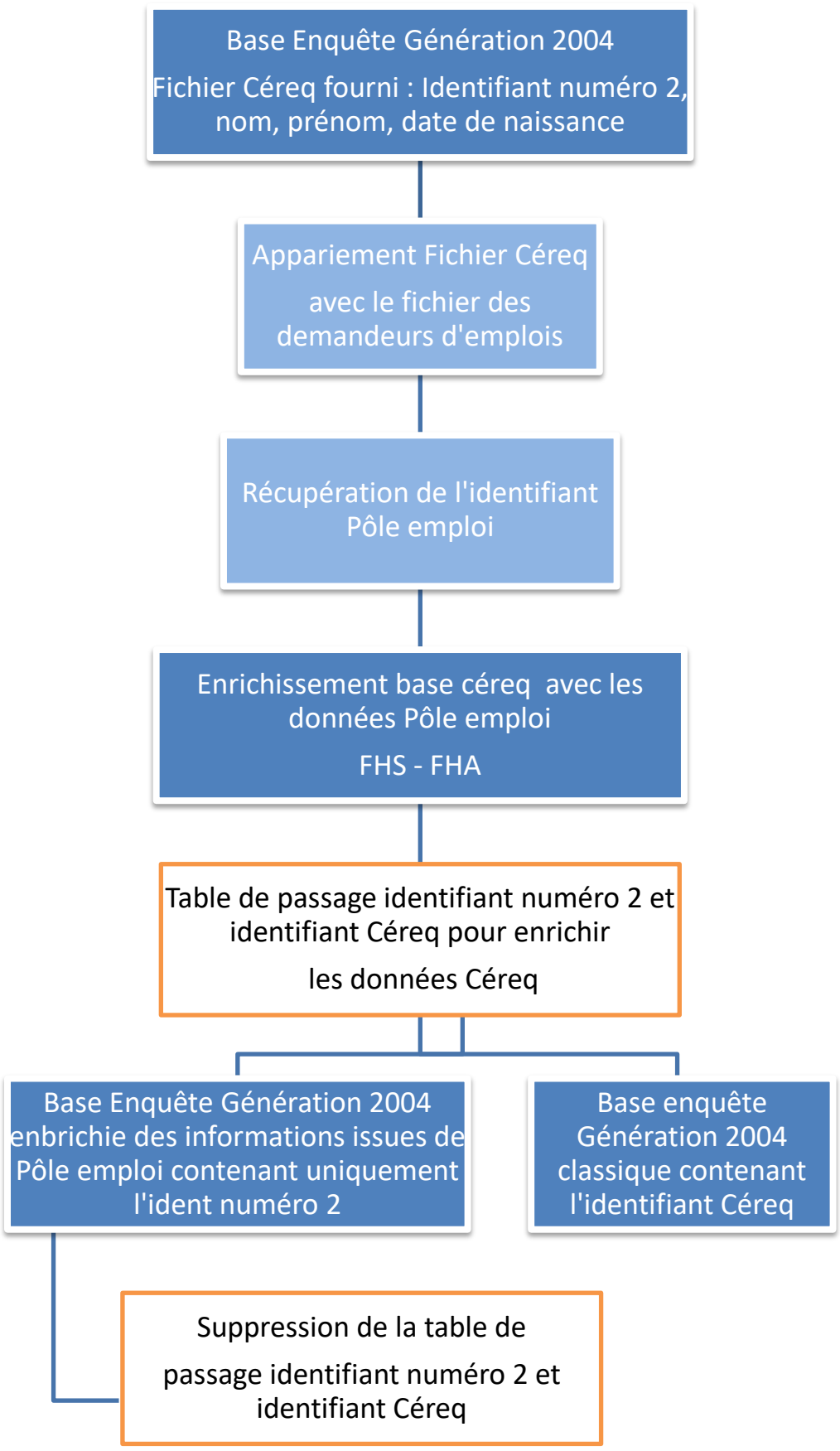
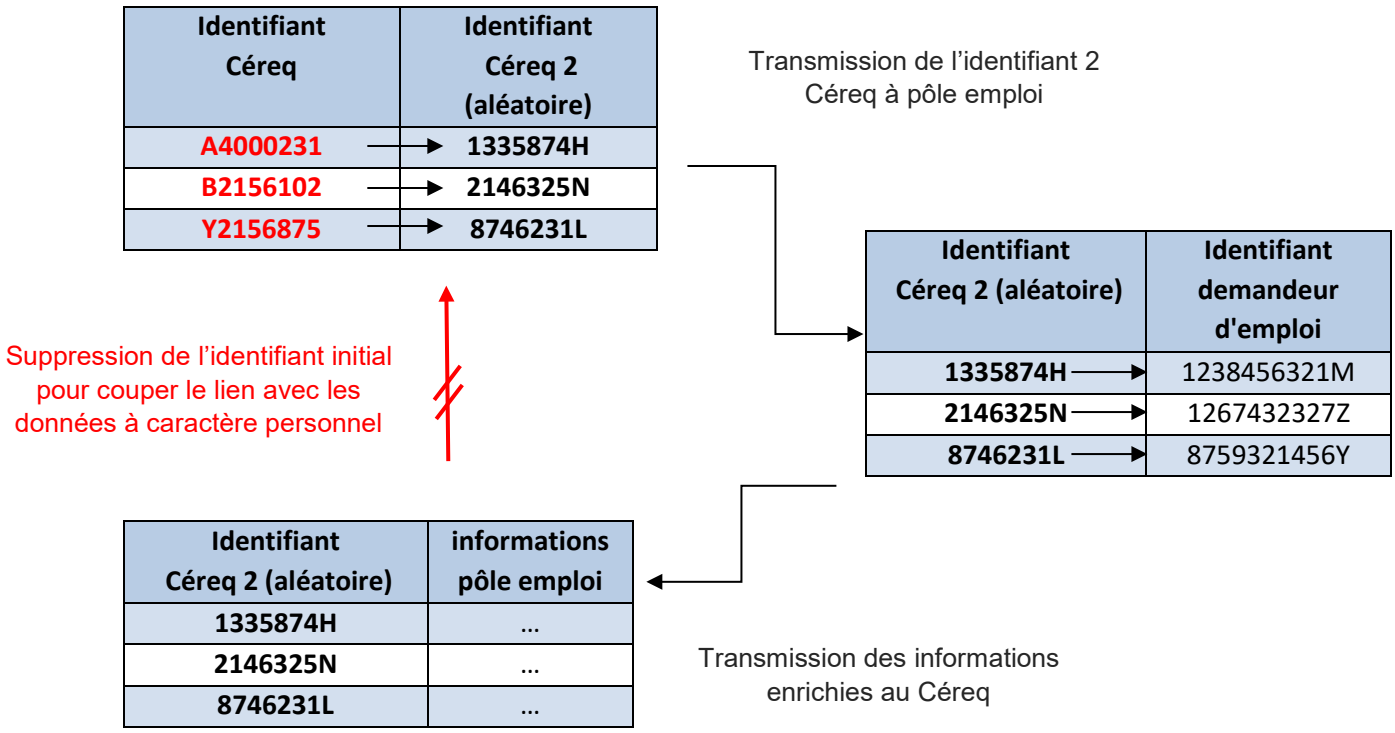

Base Enquête Génération 2004
Fichier Céreq fourni : Identifiant numéro 2, nom, prénom, date de naissance
Appariement Fichier Céreq avec le fichier des demandeurs d'emplois
Récupération de l'identifiant Pôle emploi
Enrichissement base céreq avec les données Pôle emploi
FHS - FHA
Table de passage identifiant numéro 2 et identifiant Céreq pour enrichir les données Céreq
Base Enquête Génération 2004 enbrichie des informations issues de Pôle emploi contenant uniquement l'ident numéro 2
Base enquête Génération 2004 classique contenant l'identifiant Céreq
Suppression de la table de passage identifiant numéro 2 et identifiant Céreq

Schéma 2 :



F. Sécurité et secret

Les bâtiments du Céreq sont doublement protégés par un système de gardiennage et de verrouillage des accès (ascenseurs et escaliers), en dehors des heures de travail (nuit, fin de semaine, jours fériés).

Les postes informatiques individuels sont protégés par un mot de passe individuel.

Les agents du Céreq sont soumis au secret professionnel.

Au sein du Céreq, les fichiers *détail complet* des enquêtes Génération et, lors des phases de réalisation des enquêtes qui le nécessitent, les fichiers nominatifs, sont stockés sur des serveurs localisés dans des locaux protégés et dont l'accès est strictement limité aux personnes responsables du service informatique. Ils sont accessibles par réseau aux seuls ayant droit.

L'équipe de gestion d'enquête est sensibilisée au cadre législatif de la loi « Informatique et Libertés ». Un des membres de cette équipe est le correspondant informatique et libertés du Céreq.

G. Catégories de destinataires

Fichiers non anonymes :

La table d'identification est conservée et échangée entre l'équipe de gestions d'enquêtes du DEEVA, au Céreq, et le service informatique en charge de l'appariement à Pôle emploi. A l'issue des traitements, et au plus tard un mois après le rapprochement, cette table est détruite.

Fichiers enrichis :

Les fichiers des enquêtes Génération 2004 enrichis de la table des informations issues de Pôle emploi sont destinés aux chargés d'études du Céreq (y compris ceux en poste dans ses centres associés). Une convention peut prévoir une mise à disposition d'autres partenaires du service public ou des laboratoires de recherches.

Annexe 6 – Contenu du fichier intitulé « Matching Cereq » daté du 11 avril 2014, créé le 14 janvier 2014 à Pôle emploi

Étapes de l'appariement

Descriptif du fichier en entrée : identifiant Cereq, nom, prénom, année de naissance, mois de naissance.

Étape de traitement préliminaire :
Suppression des espaces et des tirets dans les noms, prénoms.
Année et mois de naissance mis en numériques

Fichier individu :
A partir de l'axe individu de SID
Filtre sur les années de naissance des individus pour réduire la taille du fichier en sortie et pouvoir faire le matching (limite aux années 69 à 89)
Récupération des variables suivantes : date de mise à jour, identifiant de l'individu, identifiant BNI de l'individu, nom de naissance, prénom de naissance, année de naissance, mois de naissance.
Suppression des espaces et des tirets dans les noms, prénoms.

Matching :
Étape 1 : jointure de table à partir du nom, de l'année et du mois de naissance.
Étape 2 : conserver les individus dont le prénom Cereq ressemble au prénom Pôle emploi (=*).
Étape 3 : pour éliminer les doublons nous avons choisi d'utiliser la notion d'identifiant unique rattachée au numéro BNI ainsi que la date de mise à jour.

Si un individu est identifié au moins 2 fois avec le même nom, la même année de naissance, le même mois de naissance et avec un prénom ressemblant, alors :

Cas 1 : identifiant BNI renseigné sur chaque ligne et identique.
On supposera alors qu'il s'agit d'un seul et même individu qui a eu plusieurs identifiants Sigma. On conservera une ligne pour cet individu avec le dernier identifiant connu (date de mise à jour max) et on identifiera les anciens identifiants Sigma classés du plus récent au plus ancien.

Cas 2 : identifiant BNI n'est pas renseigné sur chaque ligne ou bien il est différent.
On considèrera alors qu'il peut s'agir de plusieurs individus distincts et on conservera autant de lignes qu'il y a d'identifiants BNI distincts. Pour chaque identifiant BNI on retrouvera la même logique que dans le cas 1 ; on conserve l'historique des identifiants Sigma se rapportant à cet identifiant BNI.

Nous créons 20 colonnes identifiants du plus récent au plus ancien par individu (ident1—ident20) au cas où cet identifiant serait utile pour capter des informations dans le SI.

Fichier en entrée :
35 075 lignes
35 046 lignes distinctes (hors identifiant Cereq)
21 376 individus avec au moins une correspondance (matching 61%)

Nombre de correspondances	Nombre d'individus
1	16789
2	3122
3	964
4	297
5	116
6	47
7	25
8	10
9	3
10	1
13	1
16	1

Parmi les 4587 individus avec plusieurs correspondances, l'élimination des doublons permet d'avoir :

Nombre de correspondances	Nombre d'individus
1	3359
2	1045
3	121
4	38
5	16
6	3
7	3
8	1
10	1

3 fichiers sont donc constitués :

Fichier avec les correspondances uniques
→ Id_unique (16789 individus, 16789 lignes)
Fichier avec correspondances multiples et un seul identifiant BNI
→ Id_mult (3359 individus, 3359 lignes)
Fichier avec correspondances multiples et plusieurs identifiants BNI
→ Id_mult_BNI (1228 individus, 2742 lignes)

Annexe 7 – Durée moyenne d'inscription en catégorie 1, 2 ou 3 selon diverses caractéristiques

Cette annexe présente les résultats des exploitations évoquées dans la cinquième partie du corps du document, qui portent sur les durées moyennes d'inscription en catégorie 1, 2 ou 3, selon diverses caractéristiques issues de l'enquête Génération 2004: la typologie des trajectoires sur sept ans (TYPOTRAJ_11), le niveau de sortie (NIVSOR9), le plus haut diplôme obtenu (PHDIP) et le nombre de sequences de non-emploi déclarées (NSCHO_TOT).

L'analyse est restreinte au champ des répondants à la troisième interrogation de l'enquête Génération 2004.

Seules les personnes retrouvées de manière univoque dans les fichiers de Pôle emploi sont considérées comme ayant été inscrites au moins une fois sur la fenêtre d'observation couverte par l'extrait du FHS transmis au Céreq. Pour les autres, non identifiées ou identifiées de façon non univoque, la durée d'inscription est mise à zero.

Seules les demandes d'emploi de catégorie 1, 2 ou 3 sont prises en compte pour calculer la durée totale d'inscription. Celle-ci est calculé en "mois civils". Si une personne s'inscrit au cours du mois de janvier et sort des listes au cours du mois suivant, sa durée d'inscription est alors de deux mois (janvier et février). La durée d'inscription est donc surestimée. Ce choix a été fait parce qu'en cas d'exercice d'une activité réduite, la seule information disponible est le nombre de mois concernés par l'exercice d'une activité réduite. La durée d'inscription sans activité réduite est calculée par soustraction du nombre de mois d'exercice d'une activité réduite à la durée totale d'inscription.

Les résultats n'utilisent pas les pondérations disponibles dans l'enquête Génération 2004.

Les durées moyennes sont d'abord calculées sur l'ensemble des répondants à la troisième interrogation de l'enquête Génération 2004, avec une durée nulle pour les répondants absents des fichiers de Pôle emploi ou retrouvés de façon non univoque (tableaux à gauche dans les pages suivantes). Elles sont ensuite calculées pour les seules personnes inscrites au moins une fois en catégorie 1, 2 ou 3 (tableaux à droite dans les pages suivantes). Les premiers tableaux permettent d'évaluer le degré d'exposition au "chômage administratif" de l'ensemble du groupe de personnes étudiées. Les seconds tableaux donnent la durée du "chômage administratif" réellement subie par les personnes concernées du groupe.

Distribution de la durée d'inscription en catégorie 1, 2 ou 3 selon la typologie des trajectoires à 7 ans

		Calcul sur tous				*Calcul sur les seules personnes inscrites*	
		Durée totale en catégorie 1, 2 ou 3					
TYPOTRAJ_11	**Effectif**	**Moy.**	**P25**	**Mediane**	**P75**	**effectif**	**soit**
1 = ils se sont stabilisé rapidement en EDI	4 753	4,1	0,0	0.0	4,0	1720	36%
2 = ils se sont stabilisé en EDI de façon différée	2 440	8,7	0,0	3.0	12,0	1433	59%
3 = ils sont restés durablement en EDD	1 310	25,0	0,0	12.0	43,0	865	66%
4 = ils se sont stabilisé tardivement en EDI	2 254	14,7	0,0	9.0	23,0	1502	67%
5 = ils ont décroché de l'emploi	832	29,8	0,0	25.0	51,0	567	68%
6 = ils ont connu des épisodes de chômage persistants ou récurrents	432	41,6	0,0	39.5	69,0	315	73%
7 = ils ont passé une ou plusieurs longues périodes en dehors du marché du travail	344	13,9	0,0	6.0	22,0	211	61%

	Durée sans activité réduite				Durée sans activité réduite			
	Moy.	**P25**	**Mediane**	**P75**	**Moy.**	**P25**	**Mediane**	**P75**
1 = ils se sont stabilisé rapidement en EDI	2,7	0,0	0,0	3.0	7,4	2	5	9
2 = ils se sont stabilisé en EDI de façon différée	5,0	0,0	2,0	7.0	8,6	3	6	11
3 = ils sont restés durablement en EDD	12,7	0,0	7,0	21.0	19,3	7	15	29
4 = ils se sont stabilisé tardivement en EDI	9,0	0,0	5,0	14.0	13,5	5	10	19
5 = ils ont décroché de l'emploi	20,4	0,0	17,0	34.0	30,0	16	28	41
6 = ils ont connu des épisodes de chômage persistants ou récurrents	33,5	0,0	31,0	54.0	45,9	25	41	62
7 = ils ont passé une ou plusieurs longues périodes en dehors du marché du travail	9,7	0,0	4,0	13.5	15,8	6	11	23

	Durée avec une activité réduite				durée avec une activité réduite			
	Moy.	**P25**	**Mediane**	**P75**	**Moy.**	**P25**	**Mediane**	**P75**
1 = ils se sont stabilisé rapidement en EDI	1,4	0,0	0,0	0,0	3,8	0	1	4
2 = ils se sont stabilisé en EDI de façon différée	3,7	0,0	0,0	4,0	6,2	0	3	9
3 = ils sont restés durablement en EDD	12,2	0,0	2,0	18,0	18,5	2	11	28
4 = ils se sont stabilisé tardivement en EDI	5,7	0,0	1,0	7,0	8,6	1	4	12
5 = ils ont décroché de l'emploi	9,4	0,0	2,0	14,0	13,8	2	9	19
6 = ils ont connu des épisodes de chômage persistants ou récurrents	8,1	0,0	2,0	12,0	11,1	1	6	16
7 = ils ont passé une ou plusieurs longues périodes en dehors du marché du travail	4,2	0,0	0,0	3,0	6,9	0	2	9

Distribution de la durée d'inscription en catégorie 1, 2 ou 3 selon le niveau de sortie

		Calcul sur tous				*Calcul sur les seules personnes inscrites*	
		Durée totale en catégorie 1, 2 ou 3					
NIVSOR9	**Effectif**	**Moy.**	**P25**	**Mediane**	**P75**	**effectif**	**soit**
Non diplômé	1 056	24,7	0,0	13,0	42,0	714	68%
CAP-BEP-MC	1 887	15,7	0,0	3,0	23,0	1067	57%
Bac	2 938	14,1	0,0	4,0	20,0	1706	58%
Deug	265	11,9	0,0	0,0	18,0	125	47%
BTS-DUT-Santé social	2 549	6,8	0,0	0,0	7,0	1046	41%
Licence, L3	1 186	9,6	0,0	0,0	11,0	535	45%
Maitrise, M1,…	683	12,5	0,0	4,0	16,0	407	60%
DEA, DESS, M2	1 609	9,4	0,0	3,0	13,0	943	59%
Doctorat	192	8,0	0,0	0,0	9,5	70	36%

	Durée sans activité réduite				**Durée sans activité réduite**			
	Moy.	**P25**	**Mediane**	**P75**	**Moy.**	**P25**	**Mediane**	**P75**
Non diplômé	16,6	0,0	8,0	27,0	24,5	8,0	18,0	36,0
CAP-BEP-MC	10,1	0,0	2,0	14,0	17,8	4,0	11,0	27,0
Bac	8,6	0,0	2,0	11,0	14,8	4,0	9,0	20,0
Deug	7,1	0,0	0,0	9,0	15,1	5,0	10,0	20,0
BTS-DUT-Santé social	3,7	0,0	0,0	4,0	9,1	2,0	5,0	11,0
Licence, L3	5,7	0,0	0,0	6,0	12,7	3,0	7,0	17,0
Maitrise, M1,…	7,8	0,0	2,0	10,0	13,1	4,0	8,0	17,0
DEA, DESS, M2	6,5	0,0	2,0	9,0	11,0	4,0	7,0	14,0
Doctorat	5,8	0,0	0,0	8,0	16,0	6,0	13,0	25,0

	Durée avec une activité réduite				**Durée avec une activité réduite**			
	Moy.	**P25**	**Mediane**	**P75**	**Moy.**	**P25**	**Mediane**	**P75**
Non diplômé	8,1	0,0	1,0	11,0	12,0	1,0	5,0	17,0
CAP-BEP-MC	5,6	0,0	0,0	5,0	9,9	1,0	4,0	14,0
Bac	5,5	0,0	0,0	5,0	9,4	1,0	4,0	13,0
Deug	4,7	0,0	0,0	4,0	10,1	1,0	4,0	14,0
BTS-DUT-Santé social	3,0	0,0	0,0	1,0	7,4	0,0	3,0	9,0
Licence, L3	3,9	0,0	0,0	3,0	8,6	0,0	3,0	10,0
Maitrise, M1,…	4,7	0,0	0,0	4,0	7,9	0,0	3,0	9,0
DEA, DESS, M2	2,9	0,0	0,0	2,0	5,0	0,0	1,0	6,0
Doctorat	2,1	0,0	0,0	0,0	5,9	0,0	1,0	6,0

Distribution de la durée d'inscription en catégorie 1, 2 ou 3 selon le plus haut diplôme obtenu

		Calcul sur tous				*Calcul sur les seules personnes inscrites*	
		Durée totale en catégorie 1, 2 ou 3					
PHDIP	**Effectif**	**Moy.**	**P25**	**Mediane**	**P75**	**effectif**	**soit**
Non diplômé	961	24,1	0,0	11.0	42	635	66%
CAP-BEP-MC Tertiaire	820	17,8	0,0	5.5	25	496	60%
CAP-BEP-MC Industriel	961	16,6	0,0	4.0	24	555	58%
Bac pro./techno. Tertiaire	1 438	15,6	0,0	5.0	23	877	61%
Bac pro./techno. Industriel	951	12,0	0,0	2.0	15	497	52%
Bac. Général	634	12,2	0,0	3.0	17	351	55%
Bac.+2 Santé-Social	1 292	2,2	0,0	0.0	0	285	22%
Bac.+2 Tertiaire	951	13,8	0,0	5.0	18	585	62%
Bac.+2 Industriel	610	9,8	0,0	3.0	12	348	57%
Licence pro.	345	8,3	0,0	3.0	10	193	56%
L3 LSH, Gestion, Droit	520	11,0	0,0	0.0	13	228	44%
L3 Science, Santé, STAPS	210	9,2	0,0	0.0	11	79	38%
M1	742	12,0	0,0	2.0	16	413	56%
M2 LSH, gestion, Droit	696	10,3	0,0	4.0	14	402	58%
Ecoles de commerce bac.+5	120	6,8	0,0	2.0	11	65	54%
M2 Science, Santé, STAPS	433	10,2	0,0	4.0	14	263	61%
Ecoles d'ingénieurs	475	7,3	0,0	2.0	8	267	56%
Doctorat	206	7,9	0,0	0.0	10	74	36%

	Durée sans activité réduite				**Durée sans activité réduite**			
	Moy.	**P25**	**Mediane**	**P75**	**Moy.**	**P25**	**Mediane**	**P75**
Non diplômé	16,5	0,0	7,0	28.0	25,0	8,0	19,0	37.0
CAP-BEP-MC Tertiaire	11,7	0,0	3,0	17.0	19,4	5,0	12,0	28.5
CAP-BEP-MC Industriel	10,2	0,0	3,0	15.0	17,6	5,0	12,0	27.0
Bac pro./techno. Tertiaire	9,5	0,0	3,0	13.0	15,5	4,0	10,0	21.0
Bac pro./techno. Industriel	7,4	0,0	1,0	9.0	14,1	3,0	8,0	19.0
Bac. Général	7,4	0,0	2,0	10.0	13,4	4,0	9,0	19.0
Bac.+2 Santé-Social	1,2	0,0	0,0	0.0	5,6	2,0	3,0	7.0
Bac.+2 Tertiaire	7,5	0,0	3,0	10.0	12,2	4,0	7,0	15.0
Bac.+2 Industriel	5,8	0,0	2,0	7.0	10,2	3,0	6,0	11.0
Licence pro.	5,1	0,0	2,0	6.0	9,1	3,0	6,0	11.0
L3 LSH, Gestion, Droit	6,7	0,0	0,0	7.0	15,3	4,0	10,0	21.5
L3 Science, Santé, STAPS	5,1	0,0	0,0	4.0	13,5	3,0	11,0	21.0
M1	7,4	0,0	2,0	10.0	13,2	4,0	9,0	18.0
M2 LSH, gestion, Droit	6,8	0,0	3,0	10.0	11,8	4,0	8,0	15.0
Ecoles de commerce bac.+5	5,1	0,0	2,0	7.5	9,5	4,0	6,0	12.0
M2 Science, Santé, STAPS	7,0	0,0	3,0	10.0	11,5	4,0	8,0	15.0
Ecoles d'ingénieurs	5,4	0,0	2,0	7.0	9,6	3,0	6,0	12.0
Doctorat	5,8	0,0	0,0	7.0	16,1	6,0	13,0	25.0

	Durée avec une activité réduite				**Durée avec une activité réduite**			
	Moy.	**P25**	**Mediane**	**P75**	**Moy.**	**P25**	**Mediane**	**P75**
Non diplômé	7,5	0,0	1,0	9.0	11,4	1,0	5.0	15
CAP-BEP-MC Tertiaire	6,0	0,0	0,0	7.5	10,0	1,0	4.5	15
CAP-BEP-MC Industriel	6,5	0,0	0,0	6.0	11,2	1,0	4.0	15
Bac pro./techno. Tertiaire	6,1	0,0	0,0	7.0	10,0	1,0	4.0	14
Bac pro./techno. Industriel	4,6	0,0	0,0	4.0	8,8	0,0	3.0	11
Bac. Général	4,7	0,0	0,0	4.0	8,5	1,0	3.0	11
Bac.+2 Santé-Social	0,9	0,0	0,0	0.0	4,3	0,0	1.0	4
Bac.+2 Tertiaire	6,3	0,0	0,0	7.0	10,2	1,0	4.0	13
Bac.+2 Industriel	4,0	0,0	0,0	4.0	7,1	1,0	3.0	9
Licence pro.	3,2	0,0	0,0	3.0	5,7	0,0	2.0	6
L3 LSH, Gestion, Droit	4,3	0,0	0,0	3.0	9,9	1,0	5.0	13
L3 Science, Santé, STAPS	4,1	0,0	0,0	3.0	10,9	1,0	5.0	17
M1	4,7	0,0	0,0	4.0	8,4	0,0	3.0	9
M2 LSH, gestion, Droit	3,5	0,0	0,0	3.0	6,0	0,0	2.0	7
Ecoles de commerce bac.+5	1,7	0,0	0,0	1.0	3,1	0,0	1.0	2
M2 Science, Santé, STAPS	3,2	0,0	0,0	3.0	5,3	0,0	1.0	6
Ecoles d'ingénieurs	1,9	0,0	0,0	1.0	3,4	0,0	1.0	3
Doctorat	2,1	0,0	0,0	0.0	5,8	0,0	1.0	6

La collection *Working Paper* publie des textes pour engager le débat avec d'autres chercheurs.
La publication n'engage que l'auteur.

Distribution de la durée d'inscription en catégorie 1, 2 ou 3 selon le nombre de sequences de non-emploi déclarées dans le calendrier de l'enquête Génération

	Calcul sur tous					*Calcul sur les seules personnes inscrites*	
		Durée totale en catégorie 1, 2 ou 3					
nscho_tot	Effectif	Moy.	P25	Mediane	P75	effectif	soit
0	5 516	3,7	0,0	0.0	2,0	1613	29%
1	295	15,8	0,0	8.0	27,0	186	63%
2	394	16,0	0,0	9.0	24,0	268	68%
3	2 720	11,5	0,0	6.0	15,0	1954	72%
4	366	24,8	0,0	16.0	39,0	260	71%
5	419	24,4	2,0	17.0	38,0	321	77%
6	950	19,0	2,0	12.0	24,0	726	76%
7	279	32,5	0,0	28.0	51,0	209	75%
8	290	28,1	0,0	22.5	45,0	210	72%
9	377	26,4	8,0	20.0	38,0	300	80%
10	154	35,8	8,0	37.0	58,0	119	77%
11	153	34,5	9,0	33.0	52,0	119	78%
12	141	34,1	11,0	26.0	52,0	113	80%
13 et plus	311	35,1	0,0	31.0	56,0	215	69%

	Durée sans activité réduite				Durée sans activité réduite			
	Moy.	P25	Mediane	P75	Moyenne	P25	Medianne	P75
0	2,0	0,0	0.0	1,0	7,0	2,0	4,0	8,0
1	9,5	0,0	4.0	15,0	15,1	5,0	11,0	21,0
2	9,4	0,0	5.5	14,0	13,8	5,0	10,0	20,0
3	7,3	0,0	4.0	9,0	10,1	3,0	6,0	12,0
4	15,0	0,0	9.5	24,0	21,1	8,0	17,0	29,0
5	16,8	1,0	10.0	24,0	22,0	8,0	17,0	30,0
6	12,5	1,0	8.0	16,0	16,3	6,0	11,0	20,0
7	21,8	0,0	17.0	33,0	29,1	14,0	23,0	40,0
8	19,1	0,0	14.5	30,0	26,4	12,0	23,0	35,0
9	17,4	4,0	12.0	24,0	21,9	9,0	15,0	30,0
10	23,6	5,0	22.0	36,0	30,6	17,0	30,0	39,0
11	23,1	6,0	21.0	33,0	29,7	16,0	28,0	38,0
12	19,4	7,0	15.0	31,0	24,2	12,0	20,0	34,0
13 et plus	21,8	0,0	19.0	35,0	31,5	18,0	28,0	44,0

	Durée avec une activité réduite				Durée avec une activité réduite			
	Moy.	P25	Mediane	P75	Moy.	P25	Mediane	P75
0	1,7	0,0	0,0	0,0	5,9	0,0	1.0	7.0
1	6,3	0,0	1,0	8,0	10,0	1,0	4.0	13.0
2	6,7	0,0	1,0	8,0	9,8	1,0	4.0	13.5
3	4,2	0,0	0,0	4,0	5,9	0,0	1.0	6.0
4	9,8	0,0	2,0	14,0	13,8	1,0	7.0	20.0
5	7,6	0,0	2,0	9,0	9,9	1,0	5.0	13.0
6	6,6	0,0	2,0	8,0	8,6	1,0	3.5	11.0
7	10,7	0,0	4,0	14,0	14,2	2,0	8.0	19.0
8	9,0	0,0	4,0	12,0	12,4	3,0	7.5	18.0
9	9,0	0,0	5,0	13,0	11,3	2,0	7.5	15.5
10	12,2	0,0	8,0	18,0	15,7	5,0	13.0	22.0
11	11,4	0,0	6,0	18,0	14,7	4,0	10.0	22.0
12	14,7	1,0	10,0	20,0	18,3	5,0	13.0	23.0
13 et plus	13,3	0,0	5,0	20,0	19,2	4,0	15.0	24.0

La collection *Working Paper* publie des textes pour engager le débat avec d'autres chercheurs.
La publication n'engage que l'auteur.

Annexe 8 – Programme SAS

```
/*************************************************/
/* EXPLORATION DONNEES GENERATION 2004 A  7 ANS  */
/*************************************************/

/* répertoire de Génération2004 à 7 ans - table g047ansindividus8sspi  */
libname gene "U:\DEEVA\03-ENQUETES GENERATION\02-bases de données\GENE2004\9- Bases
3eme interro";
*Libname gene "\\cereq\u\personnel\JUGNOT\FHS\livraison";

/* répertoire des tables livrées par Pôle emploi */
Libname fhs2 "U:\DEEVA\04-REFONTE
DISPOSITIF\2.AVENIR_DU_DISPOSITIF\2_Appariements_experimentaux\Gene-FHS\1-Bases de
donnees";
Libname fhs "\\cereq\u\personnel\JUGNOT\FHS\livraison";

/* repertoire de sorties pour les travaux */
Libname sor "U:\DEEVA\04-REFONTE
DISPOSITIF\2.AVENIR_DU_DISPOSITIF\2_Appariements_experimentaux\Gene-FHS\3-
Travaux2021";
*Libname sor "\\cereq\u\personnel\JUGNOT\FHS";

/*************************************************/
/*************************************************/
/* EXPLORATION DONNEES DE LA LIVRAISON FHS       */
/*************************************************
/*************************************************/

Title "Liste des tables dans la librairie FHS2";
PROC CONTENTS DATA=FHS2._ALL_ OUT=LISTE_TABLE NOPRINT;
run;

Title "Description de tables du FHS (pour repérer les identifiants disponibles)" ;
Proc contents data=fhs2.fus_statde;
Proc contents data=fhs2.fus_pap;
Proc contents data=fhs2.fus_de;
Proc contents data=fhs2.fus_E0;
proc print data=fhs2.fus_DE (obs=10);
run;

Title "Comparaison variable IDENT et IDTIDV";
Data ComparId; set fhs2.fus_E0 (keep=IDENT IDTIDV);
comparaison=(ident=idtidv);
run;
Proc freq data=ComparID;table comparaison;
run;

Title "Description de tables du FHA (pour repérer les identifiants disponibles)" ;
Proc contents data=fhs.fus_E3;
Proc contents data=fhs.fus_M0;
run;

Title "Description des tables de rapprochement des identifiants faites lors des
travaux d'appariement initiaux" ;
Proc contents data=fhs2.id_unique; proc print data=fhs2.id_unique (obs=10);
Proc contents data=fhs2.id_mult; proc print data=fhs2.id_mult (obs=10);
Proc contents data=fhs2.id_mult_bni;proc print data=fhs2.id_mult_bni (obs=10);
Proc contents data=fhs2.compil;proc print data=fhs2.compil (obs=10);
Proc freq data=fhs2.compil;tables fic;
run;

Title "Description d'autres tables potentiellement utiles produites lors des
travaux initiaux" ;
Proc contents data=fhs2.table_passage_entiere;proc print
data=fhs2.table_passage_entiere (obs=10);
Proc contents data=fhs2.table_identification;proc print
data=fhs2.table_identification (obs=10);
```

```
run;

/*************************************************/
/**** Verif sur table CORRESP_IDENT pour voir si */
/**** 1° variable IDENT = variable IDTIDV -> Oui (et non vide) */
/**** 2° IDX toujours renseigné -> non dans 10% des observations  */
/**** 3° le meme IDX est la plusieurs fois */
/**** -> 24884 IDX unique (et non vide),  261 IDX en double, 1 IDX en triple */
/**** 4° le même IDENT est la plusieurs fois -> non, tous différents*/
/*************************************************/
Data Verif;set fhs2.corresp_ident;
If Ident="" then verif_ident="Vide  ";
   Else if Ident=idtidv then verif_ident="IDTIDV";
   Else verif_ident="DIFFER";

If idtidv="" then verif_idtidv="Vide";
   Else verif_idtidv="OK  ";

If IDX="" then verif_idx="Vide";
   Else verif_idx="OK  ";

Title "1° La variable IDENT est-elle égale à la variable IDTIDV dans CORRESP_IDENT
?";
Proc freq data=verif; tables verif_ident*verif_idtidv/nocol norow nopercent;

Title "2° La variable IDX est-elle parfois non renseignée ?";
Proc freq data=verif; tables verif_idx/nocol norow nopercent;

Proc sort data=verif;by idx;run;
Data verif; set verif; retain n_idx idx1;
If idx1=idx then n_idx=n_idx+1;
Else do; n_idx=1; idx1=idx;end;
run;
Title "3° La variable IDX est-elle parfois en double ou plus ?";
Proc freq data=verif; tables n_idx/nocol norow nopercent;

Proc sort data=verif;by ident;run;
Data verif; set verif; retain n_ident ident1;
If ident1=ident then n_ident=n_ident+1;
Else do; n_ident=1; ident1=ident;end;
run;
Title "3° La variable IDENT est-elle parfois en double ou plus ?";
Proc freq data=verif; tables n_ident/nocol norow nopercent;

proc sort data=verif; by idx;
Data fhs2.doublons ;* creation de table en sortie des IDX associés à plusieurs
IDENT/IDTIDV;
set verif ;by idx; if n_idx=2;
Keep idx ;
run;
Proc sort data=fhs2.corresp_ident;by idx;
Data fhs2.doublons ;
merge fhs2.corresp_ident (keep=IDX IDTIDV)
      fhs2.doublons (in=a);
by idx; if a;
run;
Data Verif; set _null_;
run;

/*************************************************/
/****** Examen de l'unicité de IDENT_BNI dans la table ID_UNIQUE ******/
/*************************************************/

Data id_unique ;set fhs2.id_unique;ident2 = substr(identifiant2,1,7);run;
proc sort data=id_unique;by ident_BNI;run;
Data id_unique ; set id_unique ; by ident_BNI; retain n_BNI;
if first.ident_BNI then n_BNI=1; else n_BNI=n_BNI+1;
run;
Title "Repérage des IDENT_BNI multiples dans ID_UNIQUE";
Proc freq data=id_unique; tables n_BNI;run;
Data verif_doublon; set id_unique ;
```

```
If n_BNI ge 2;
If ident_BNI="YYYY"  then AnomalieBNI="YYYY  ";
else if ident_BNI="0000000000" then Anomalie="000000";
else if ident_BNI="" then AnomalieBNI="Vide  ";
Else Anomalie="Doublon";
selec=1-(AnomalieBNI in ("YYYY  ","Vide  "));
run;
Proc freq data=verif_doublon; tables AnomalieBNI;run;
Title "Liste des IDENT_BNI en doublon dans ID_UNIQUE";
Proc print data=verif_doublon (keep=Ident_BNI selec where=(selec=1));
run;

/************************************************/
/****** Examen de l'unicité de IDENT2 dans la table ID_UNIQUE ******/
/************************************************/
Data id_unique ;set fhs2.id_unique;ident2 = substr(identifiant2,1,7);run;
proc sort data=id_unique;by ident2;run;
Data id_unique ; set id_unique ; by ident2; retain n_ident2;
if first.ident2 then n_ident2=1; else n_ident2=n_ident2+1;
run;
Title "Repérage des IDENT2 multiples dans ID_UNIQUE";
Proc freq data=id_unique; tables n_ident2;run;
Data verif_doublon; set id_unique ;
If n_ident2 ge 2;
run;
Title "Liste des IDENT2 en doublon dans ID_UNIQUE";
Proc print data=verif_doublon (keep=Ident2) ;

/************************************************/
/****** Examen de l'unicité de IDTIDV dans la table ID_UNIQUE ******/
/************************************************/
Data id_unique ;set fhs2.id_unique;ident2 = substr(identifiant2,1,7);run;
proc sort data=id_unique;by idtidv;run;
Data id_unique ; set id_unique ; by idtidv; retain n_idtidv;
if first.idtidv then n_idtidv=1; else n_idtidv=n_idtidv+1;
run;
Title "Repérage des IDTIDV multiples dans ID_UNIQUE";
Proc freq data=id_unique; tables n_idtidv;run;
Data verif_doublon; set id_unique ;
If n_idtidv ge 2;
run;
Title "Liste des IDTIDV en doublon dans ID_UNIQUE";
Proc print data=verif_doublon (keep=Idtidv) ; run ;

/************************************************/
/**** Création de la table ID_UNIQUE_Sj avec ajout de variables sur les anomalies
d'identifiants **/
/************************************************/

Data fhs2.id_unique_Sj ;set fhs2.id_unique;
ident2 = substr(identifiant2,1,7);

if Ident2 in([liste des identifiants repérés en doublon]) then
AnomalieBNI="Doublon";
       * recup des IDENT2 en doublon sur des IDENT_BNI non renseignés;
else If ident_BNI="YYYY"  then AnomalieBNI="YYYY   ";
else if ident_BNI="" then AnomalieBNI="Vide   ";
else if ident_BNI="0000000000" then AnomalieBNI="0000   ";
else if ident_BNI in ([liste des identifiants repérés en doublon]) then
AnomalieBNI="Doublon";
Else AnomalieBNI="OK     ";

If ident2 in ([liste des identifiants repérés en doublon]) then
AnomalieIdent2="Doublon_AB";
Else AnomalieIdent2="OK       ";

If Idtidv in ([liste des identifiants repérés en doublon]) then
AnomalieIdtidv="Doublon";
Else AnomalieIdtidv="OK     ";
```

```
If AnomalieBNI="OK      " and AnomalieIdent2="OK        " and AnomalieIdtidv="OK
" then Anomalie="OK      ";
else if AnomalieBNI in ("YYYY    ","Vide    ","0000    ") then Anomalie = "SansBNI";
Else Anomalie="A VOIR ";
run;

Title "Examen des Anomalies d'identifiants dans la table ID_Unique";
Proc freq data=fhs2.id_unique_Sj;
tables AnomalieBNI*AnomalieIdent2 Anomalie AnomalieBNI AnomalieIdent2
AnomalieIdtidv/nopercent norow nocol;
run;

Proc sort data=fhs2.id_unique_Sj ; by AnomalieBNI ident_BNI;run;
Proc print data=fhs2.id_unique_Sj (where=(Anomalie="A VOIR "));
run;


/*************************************************/
/****** Examen de l'unicité de IDENT_BNI dans la table ID_MULT ******/
/*************************************************/

Data id_mult ;set fhs2.id_mult;ident2 = substr(identifiant2,1,7);run;
proc sort data=id_mult;by ident_BNI;run;
Data id_mult ; set id_mult ; by ident_BNI; retain n_BNI;
if first.ident_BNI then n_BNI=1; else n_BNI=n_BNI+1;
run;
Proc freq data=id_mult; tables n_BNI;run;
Data verif_doublon; set id_mult ;
If n_BNI ge 2;
If ident_BNI="YYYY"  then AnomalieBNI="YYYY";
else if ident_BNI="" then AnomalieBNI="Vide";
If AnomalieBNI in ("YYYY","Vide") then delete;
run;
Proc print data=verif_doublon (keep=Ident_BNI);*recup manuel des ident_BNI en
doublon ;
run;


/*************************************************/
/****** Examen de l'unicité de IDENT2 dans la table ID_MULT ******/
/*************************************************/
Data id_mult ;set fhs2.id_mult;ident2 = substr(identifiant2,1,7);run;
proc sort data=id_mult;by ident2;run;
Data id_mult ; set id_mult ; by ident2; retain n_ident2;
if first.ident2 then n_ident2=1; else n_ident2=n_ident2+1;
run;
Proc freq data=id_mult; tables n_ident2;run;
Data verif_doublon; set id_mult ;
If n_ident2 ge 2;
run;
Proc print data=verif_doublon (keep=Ident2);*recup manuel des ident2 en doublon ;

/*************************************************/
/**** Création de la table ID_MULT_Sj avec ajout de variables sur les anomalies
d'identifiants **/
/*************************************************/


Data fhs2.id_mult_Sj ;set fhs2.id_mult;
ident2 = substr(identifiant2,1,7);

if ident_BNI in ([liste des identifiants repérés en doublon]) then
AnomalieBNI="Doublon";
Else AnomalieBNI="OK      ";

If ident2 in ([liste des identifiants repérés en doublon])  then
AnomalieIdent2="Doublon";
Else AnomalieIdent2="OK      ";

If AnomalieBNI="OK      " and AnomalieIdent2="OK      " then Anomalie="OK      ";
Else Anomalie="A VOIR ";
```

```
run;

Title "Examen des Anomalies d'identifiants dans la table ID_Mult";
Proc freq data=fhs2.id_mult_Sj;
tables AnomalieBNI*AnomalieIdent2 Anomalie/nopercent norow nocol;
run;

Proc sort data=fhs2.id_mult_Sj ; by AnomalieBNI ident_BNI;run;
Proc print data=fhs2.id_mult_Sj (where=(Anomalie="A VOIR "));
run;

/*************************************************/
/**** Calcul du nombre d'Identifiants sigma différents (IDTIDV) **/
/**** par personne identifiée dans ID_MULT
/*************************************************/
Data NB_IDTIDV ;set fhs2.id_mult;
Nb_idtidv=(IDTIDV ne "")+(IDTIDV1 ne "")+(IDTIDV2 ne "")+(IDTIDV3 ne "")+(IDTIDV4
ne "")+(IDTIDV5 ne "")+
          (IDTIDV6 ne "")+(IDTIDV7 ne "")+(IDTIDV8 ne "")+(IDTIDV9 ne "")+(IDTIDV10
ne "")+(IDTIDV11 ne "")+
          (IDTIDV12 ne "")+(IDTIDV13 ne "")+(IDTIDV14 ne "")+(IDTIDV15 ne
"")+(IDTIDV16 ne "")+(IDTIDV17 ne "")+
          (IDTIDV18 ne "")+(IDTIDV19 ne "")+(IDTIDV20 ne "");
Title "Nombre d'identifiants IDTIDV par personnes dans la table ID_MULT";
Proc freq data=NB_IDTIDV;tables Nb_idtidv;run;
Title "Observation anormalement présente avec un seul IDTIDV";
Proc print data=NB_IDTIDV (where=(Nb_idtidv=1));
run;
Data NB_IDTIDV; set _null_;
run;


/*************************************************/
/****** Examen du nbe d'IDENT_BNI dans la table ID_MULT_BNI ******/
/*************************************************/

Data id_mult_bni ;set fhs2.id_mult_bni;ident2 = substr(identifiant2,1,7);run;
proc sort data=id_mult_bni;by ident_BNI;run;
Data id_mult_bni ; set id_mult_bni ; by ident_BNI; retain n_BNI;
if first.ident_BNI then n_BNI=1; else n_BNI=n_BNI+1;
run;
Title "Nbre d'IDENT_BNI dans ID_MULT_BNI";
Proc freq data=id_mult_bni; tables n_BNI;run;
Title "IDENT_BNI présents 2 fois ou plus dans ID_MULT_BNI";
Proc print data=id_mult_bni (where=(n_BNI ge 2));
run;

/*************************************************/
/**** Création de la table ID_MULT_BNI_Sj avec ajout de variables sur les anomalies
d'identifiants **/
/*************************************************/
Data fhs2.id_mult_bni_Sj; set id_mult_bni;
if ident_BNI="" then AnomalieBNI="Vide   ";
else if ident_BNI="0000000000" then AnomalieBNI="0000   ";
else if ident_BNI in ([liste des identifiants repérés en doublon]) then
AnomalieBNI="Doublon";
else AnomalieBNI="OK     ";
run;
Title "Anomalie sur IDENT_BNI dans ID_MULT_BNI";
Proc freq data=fhs2.id_mult_bni_Sj; tables AnomalieBNI;
run;
Title "Les BNI en doublon dans ID_MULT_BNI";
Proc print data=fhs2.id_mult_bni_Sj (where=(AnomalieBNI="Doublon"));
run;

/*************************************************/
/****** Examen de l'unicité de IDENT2 dans la table ID_MULT_BNI ******/
/*************************************************/
Data id_mult_bni_Sj ;set fhs2.id_mult_bni;
ident2 = substr(identifiant2,1,7);run;
proc sort data=id_mult_bni_Sj;by ident2;run;
```

```
Data id_mult_bni_Sj; set id_mult_bni_Sj ; by ident2; retain n_ident2;
if first.ident2 then n_ident2=1; else n_ident2=n_ident2+1;
if ident_BNI="" then AnomalieBNI="Vide   ";
else if ident_BNI="0000000000" then AnomalieBNI="0000   ";
else if ident_BNI in ([liste des identifiants repérés en doublon]) then
AnomalieBNI="Doublon";
else AnomalieBNI="OK     ";
run;
Data id_mult_bni; set id_mult_bni_Sj ; by ident2;
if last.ident2 ;
run;
Title "Repérage du nombre d'occurence des IDENT2 dans ID_MULT_BNI";
Proc freq data=id_mult_bni; tables n_ident2*AnomalieBNI/nopercent norow nocol;
run;


/*************************************************/
/**** Calcul du nombre d'Identifiants sigma différents (IDTIDV) **/
/**** par personne identifiée dans ID_MULT_BNI
/*************************************************/

Data NB_IDTIDV ;set fhs2.id_mult_BNI;
Nb_idtidv=(IDTIDV ne "")+(IDTIDV1 ne "")+(IDTIDV2 ne "")+(IDTIDV3 ne "")+(IDTIDV4
ne "")+(IDTIDV5 ne "")+
          (IDTIDV6 ne "")+(IDTIDV7 ne "")+(IDTIDV8 ne "")+(IDTIDV9 ne "")+(IDTIDV10
ne "")+(IDTIDV11 ne "")+
          (IDTIDV12 ne "")+(IDTIDV13 ne "")+(IDTIDV14 ne "")+(IDTIDV15 ne
"")+(IDTIDV16 ne "")+(IDTIDV17 ne "")+
          (IDTIDV18 ne "")+(IDTIDV19 ne "")+(IDTIDV20 ne "");
Title "Nombre d'identifiants IDTIDV par personnes dans la table ID_MULT_BNI";
Proc freq data=NB_IDTIDV;tables Nb_idtidv;run;
Title "Observation anormalement présente avec un seul IDTIDV";
Proc print data=NB_IDTIDV (where=(Nb_idtidv=1));
run;

/*************************************************/
/**** Les doublons repérés supra pour les IDENT_BNI identiques associés
/**** à deux IDENT2 différents correspondent-ils à des identités proches dans la
table
/**** transmise par le CEREQ (on ecarte ici les cas ou c'est la date de naissance
qui est différente)
/*************************************************/
Data identif1 ; set fhs2.id_unique_Sj (keep=ident2 AnomalieBNI);
If AnomalieBNI = "Doublon";
source="ID_UNIQUE  ";
Run;
Data identif2 ; set fhs2.id_mult_Sj (keep=ident2 AnomalieBNI);
If AnomalieBNI="Doublon";
source="ID_MULT    ";
run;
Data identif3 ; set fhs2.id_mult_BNI_Sj (keep=ident2 AnomalieBNI);
If AnomalieBNI="Doublon";
source="ID_MULT_BNI";
run;
Data Identif ; set identif1 identif2 identif3;
run;
Proc sort data=identif;by ident2;
Proc sort data=fhs2.table_passage_entiere;by ident2;
run;
Data identif; merge fhs2.table_passage_entiere identif;by ident2;
If AnomalieBNI="Doublon";
run;
Proc sort data=identif; by nom pren; run;
Title " liste des identifiants IDENT2 repéré comme doublon sur une même personne";
Proc print data=identif;
run;


/*************************************************/
/****** Examen de l'unicité de IDENT2 dans la table COMPIL ******/
/*************************************************/
```

```
Data compil ;set fhs2.compil;ident2 = substr(identifiant2,1,7);run;
proc sort data=compil;by ident2;run;
Data compil ; set compil ; by ident2; retain n_ident2;
if first.ident2 then n_ident2=1; else n_ident2=n_ident2+1;
run;
Title "Repérage des IDENT2 multiples dans COMPIL";
Proc freq data=compil; tables n_ident2;run;
Proc freq data=compil; tables n_ident2*fic/nopercent nocol norow;run;

Data verif_doublon; set compil ;
If n_ident2 ge 2;
run;
Title "Liste des IDENT2 en doublon dans ID_UNIQUE";
Proc print data=verif_doublon (keep=Ident2 fic) ;*recup manuel des ident2 en
doublon ;
run;

Data doublon;set id_unique;*extraction des observations en vrai doublon ;
If ident2 in ([liste des identifiants repérés en doublon]));
Proc print data=doublon; run;
Data doublon; set _null_;run;

Data verif_doublon2; set compil;
If n_ident2 ge 2;
If fic in ("Id_unique","Id_mult");
run;

/*liste des observations avec les IDENT2 repérés dans VERIF_DOUBLON2*/
Data verif_doublon3; set compil;
If IDENT2 in ([liste des identifiants repérés en doublon]);
Drop IDTIDV2--IDTIDV20 DATMAJ PRENOM_;
run;


/*********************************************/
/*********************************************/
/* Synthèse                                  */
/*********************************************/
/*********************************************/

** table des échos uniques - Données Pôle emploi ;
** utilisation de la table suffixée "SJ" pour récupérer la variable ANOMALIE créé
supra ;
Data id_unique; set fhs2.id_unique_SJ (keep= prenom an_nais mois_nais identifiant2
                                            ident2 ident_bni anomalie) ;
Echo="id_unique  ";
Suffixe=substr(identifiant2,8,1);
If suffixe="A" then do; an_nais_A=an_nais; mois_nais_A=mois_nais;
ident_bni_A=ident_bni;Echo_A=Echo; end;
If suffixe="B" then do; an_nais_B=an_nais; mois_nais_B=mois_nais;
ident_bni_B=ident_bni;Echo_B=Echo;end;
run;
Proc sort data=id_unique; by ident2;
run;

** table des échos multiples pour Sigma (IDTIDV) mais même BNI - données Pôle
emploi;
** utilisation de la table suffixée "SJ" pour récupérer la variable ANOMALIE créé
supra ;
Data id_mult; set fhs2.id_mult_SJ (keep=prenom an_nais mois_nais identifiant2
ident2 ident_bni anomalie) ;
Echo="id_mult    ";
Suffixe=substr(identifiant2,8,1);
If suffixe="A" then do; an_nais_A=an_nais; mois_nais_A=mois_nais;
ident_bni_A=ident_bni;Echo_A=Echo; end;
If suffixe="B" then do; an_nais_B=an_nais; mois_nais_B=mois_nais;
ident_bni_B=ident_bni;Echo_B=Echo;end;
run;
Proc sort data=id_mult; by ident2;
run;
```

```
** table des échos multiples sur BNI - données Pôle emploi;
***** Un identifiant2 est répété autant de fois que d'échos BNI;
***** Mais comme même identiant2-nom-prenoms-mois et année de naissance ;
***** on ne garde qu'une ligne pour chaque identifiant2;
Data id_mult_bni; set fhs2.id_mult_bni_Sj (keep=prenom an_nais mois_nais
identifiant2 ident2 ident_bni) ;
Echo="id_mult_bni";
Anomalie="COMPLEX";
Suffixe=substr(identifiant2,8,1);
If suffixe="A" then do; an_nais_A=an_nais; mois_nais_A=mois_nais;
ident_bni_A=ident_bni;Echo_A=Echo; end;
If suffixe="B" then do; an_nais_B=an_nais; mois_nais_B=mois_nais;
ident_bni_B=ident_bni;Echo_B=Echo;end;
run;
Proc sort data=id_mult_bni; by ident2;
run;

Proc sort data=id_mult_bni; by identifiant2;
Data id_mult_bni; set id_mult_bni; by identifiant2; if first.identifiant2;
run;

** table ECHO_PÔLE_EMPLOI agrégeant les trois tables créées ci-dessus ;
** avec une ligne par valeur de IDENTIFIANT2 dans chacune des tables
d'identification de Pôle emploi;
Data fhs2.Echo_Pôle_Emploi ; set id_unique id_mult id_mult_bni;
run;
Proc sort data= fhs2.Echo_Pôle_Emploi; by ident2;
run;
Title "Répartition des identifiant2 par type d'écho (sur les seules données Pôle
emploi)";
Proc freq data=fhs2.Echo_Pôle_Emploi; tables Echo*Anomalie/nopercent norow nocol;
run;

** verif si des IDENT2 sont communs à ID_UNIQUE, ID_MULT, ID_MULT_BNI (deux à
deux);
Proc sort data=id_unique;by ident2;
Proc sort data=id_mult;by ident2;
Data verif; merge id_unique (keep=ident2 in=a) id_mult (keep=ident2 Echo);
By ident2; if a ; run;
Title "IDENT2 communs à ID_UNIQUE et ID_MULT";
Proc freq data=verif ;tables Echo; Proc print data=verif (where=(Echo="id_mult
"));
run;
/****/
Proc sort data=id_unique;by ident2;
Proc sort data=id_mult_bni;by ident2;
Data verif; merge id_unique (keep=ident2 in=a) id_mult_bni (keep=ident2 Echo);
By ident2; if a ; run;
Title "IDENT2 communs à ID_UNIQUE et ID_MULT_BNI";
Proc freq data=verif ;tables Echo; Proc print data=verif
(where=(Echo="id_mult_bni"));
run;
/****/
Proc sort data=id_mult;by ident2;
Proc sort data=id_mult_bni;by ident2;
Data verif; merge id_mult (keep=ident2 in=a) id_mult_bni (keep=ident2 Echo);
By ident2; if a ; run;
Title "IDENT2 communs à ID_MULT et ID_MULT_BNI";
Proc freq data=verif ;tables Echo; Proc print data=verif
(where=(Echo="id_mult_bni"));
run;
/*** listage des IDENT2 présents dans plusieurs des trois tables
d'identification***/
Data fhs2.Echo_Pôle_Emploi ; set fhs2.Echo_Pôle_Emploi;
If ident2 in ([liste des identifiants repérés en doublon]) then Plusieurs_tables=1;
Else plusieurs_tables=0;
run;
Title "Synthèse des cas d'IDENT2 présents dans plusieurs des trois tables
d'identification";
```

```
Proc sort data=fhs2.Echo_Pôle_Emploi ; by ident2;run;
Proc print data=fhs2.Echo_Pôle_Emploi (where=(Plusieurs_tables=1));
run;
Proc freq data=fhs2.Echo_Pôle_Emploi; tables Echo*Plusieurs_tables/nopercent nocol
norow;
run;


** Table COMPARAISON agrégeant :
****** - la table ECHO_PÔLE_EMPLOI
****** - avec la table d'identification extraite de GENERATION (on préfixe les
variables par "G_");
Data table_identification ;
set fhs2.table_identification (rename=(annee=G_annee mois=G_mois nom=G_nom
pren=G_pren));
run;
Proc  sort data=table_identification; by identifiant2;
run;

*** On s'assure d'une seule ligne par IDENTIFIANT2 dans la table ECHO_PÔLE_EMPLOI;
*** (normalement c'est le cas);
Proc sort data=fhs2.Echo_Pôle_Emploi;by identifiant2;run;
Data Echo_Pôle_Emploi; set fhs2.Echo_Pôle_Emploi; by identifiant2;
Ident_Echo_PE=ident2;
If first.identifiant2; run;

Data fhs2.Comparaison; merge table_identification Echo_Pôle_Emploi;
by identifiant2;
ident2 = substr(identifiant2,1,7);
If Echo="" then Echo="Absent PE  ";
If Anomalie="" then Anomalie="Absent ";

**** construction de la variable de SELECTION;
If Echo="id_mult_bni" then selection="identification complexe";

        * cas des IDENT2 presents dans plusieurs tables d'identification;
Else If ident2 in ([liste des identifiants repérés en doublon])      then
do ;selection="identification complexe";Anomalie="A VOIR ";end;


Else if Echo in ("id_unique  ","id_mult    ") and Anomalie="A VOIR "
                    then selection="identification complexe";
Else if Echo="id_unique  " then Selection="Identification simple  ";
Else if Echo="id_mult    " then Selection="Identification simple  ";
Else Selection="Non identifié par PE   ";

If suffixe="A" then Selection_A=Selection;
If suffixe="B" then Selection_B=Selection;
run;

Title "Répartition des identifiant2 par type d'écho (sur les données appariées)";
Proc freq data=fhs2.Comparaison; tables Echo Echo*Anomalie/nopercent norow nocol;
run;
Proc freq data=fhs2.Comparaison; tables Selection*(Echo Anomalie)/nopercent norow
nocol;
run;

/*********************************************/
/** Création de la table SELECTION utilisée infra pour l'appariement
/*********************************************/
Data Comparaison_A; set fhs2.Comparaison;
if suffixe="A";
suffixe_A="A";
Drop prenom an_nais mois_nais Echo Selection Ident_bni G_nom G_mois G_annee
Anomalie Suffixe an_nais_B mois_nais_B Echo_B Selection_B Suffixe_B Ident_BNI_B;
run;

Data Comparaison_B; set fhs2.Comparaison;
if suffixe="B";
suffixe_B="B";
```

```
Drop prenom an_nais mois_nais Echo Selection Ident_bni G_nom G_mois G_annee
Anomalie Suffixe an_nais_A mois_nais_A Echo_A Selection_A Suffixe_A Ident_BNI_A;
run;

Proc sort data=Comparaison_A;by ident2;run;
Proc sort data=Comparaison_B;by ident2;run;

Proc sort data=fhs2.table_passage_entiere ; by ident2; run;

Data fhs2.Selection; merge fhs2.table_passage_entiere (keep=ident2 ident q2 q3)
Comparaison_A;
                     by ident2;
run;
Data fhs2.Selection; merge fhs2.Selection Comparaison_B;
                     by ident2;
run;

Data fhs2.Selection; set fhs2.Selection;
If Suffixe_A="A" and Suffixe_B="B" then Suffixe="AB";
else if Suffixe_A="A" then Suffixe="A ";
else if Suffixe_B="B" then Suffixe="B ";

if Suffixe="A " then do; Selection=Selection_A ; Echo=Echo_A;end;
else if Suffixe="B " then do; Selection=Selection_B ; Echo=Echo_B;end;
else if Suffixe="AB" then do; selection="identification complexe"; Echo="Multi
table";end;
else do; Selection="Non identifié par PE   "; Echo="Absent PE  ";end;

If Echo="id_unique  " and Selection="Identification simple  "
     then Selection ="Identification simple 1";

run;
Title "Répartition des observations de Géné selon la variable Selection";
Proc freq data=fhs2.Selection;
tables Selection_A Selection_B Suffixe Selection selection*suffixe Echo/nopercent
norow nocol;
run;


/************************************************/
/************************************************/
/************************************************/
/************************************************/
/* ESSAI APPARIEMENT          */
/************************************************/
/************************************************/
/************************************************/
/************************************************/


/************************************************/
/************************************************/
/* Exploration de Génération 2004 à 7 ans       */
/* et construction de tables de synthèses sur le non-emploi*/
/* qui seront utilisés plus bas */
/************************************************/
/************************************************/

Data Verif ; set gene.g047ansnonemp5sspi;

If ra1_07=1 or ra1_09=1 or ra1_11 =1 then ra1_agrege=1;
else if ra1_07=3 or ra1_09=3 or ra1_11 =3 then ra1_agrege=3;
else if ra1_07=2 or ra1_09=2 or ra1_11 =2 then ra1_agrege=2;
Else ra1_agrege=0;

If ra2_0407="01" or ra2_0709="01" or ra2_0911 ="01" then ra2_agrege="01";
Else if ra2_0407 in ("02","03","04","05","06","07","08","09")
     or ra2_0709 in ("02","03","04","05","06","07","08","09")
        or ra2_0911 in ("02","03","04","05","06","07","08","09") then
ra2_agrege="02";
Else ra2_agrege="00";
```

```
Title "Exploratoire - Table des séquences de non-emploi - type des séquences ?";
Proc freq data=Verif; tables typeseq;
run;

Title "Exploratoire - Table des séquences de non-emploi - combien se disent
inscrits aux Assedic (=1)?";
Proc freq data=Verif; tables ra1_07 ra1_09 ra1_11 ra1_agrege
typeseq*ra1_agrege/nopercent nocol norow;
run;

Title "Exploratoire - Table des séquences de non-emploi - qui dit avoir été à
l'ANPE/PE en démarche(=01)?";
Proc freq data=Verif; tables ra2_agrege typeseq*ra2_agrege/nopercent nocol norow;
run;


**** TABLE SeqNonEmp : repérage des IDENT present dans la table des séquences de
non-emploi;
Proc sort data=verif;by ident;run;
Data SeqNonEmp;set verif; by ident ; if first.ident; SeqNonEmp=1;
Keep ident SeqNonEmp;
run;

**** TABLE Inscrits : repérage des IDENT déclarant etre inscrit sur au moins
séquences de non-emploi;
Data Inscrits;set verif; if ra1_agrege=1;Inscrits=1;
Keep Ident Inscrits;
Proc sort data=Inscrits;by ident;run;
Data Inscrits; set Inscrits; by ident; if first.ident;
run;

**** TABLE Demarche : repérage des IDENT déclarant avoir été au moins une fois à
l'ANPE/PE lors de la séquences de non-emploi;
Data Demarche;set verif; if ra2_agrege="01"; Demarche=1;
Keep Ident Demarche;
Proc sort data=Demarche;by ident;run;
Data Demarche; set Demarche; by ident; if first.ident;
run;

**** TABLE ChoC : repérage des IDENT déclarant au moins une sequence de chomage
court;
Data ChoC;set verif; if typeseq="chc"; ChoC=1;
Keep Ident ChoC;
Proc sort data=ChoC;by ident;run;
Data ChoC; set ChoC; by ident; if first.ident;
run;

**** TABLE ChoL : repérage des IDENT déclarant au moins une sequence de chomage
court;
Data ChoL;set verif; if typeseq="chl"; ChoL=1;
Keep Ident ChoL;
Proc sort data=ChoL;by ident;run;
Data ChoL; set ChoL; by ident; if first.ident;
run;

**** TABLE NonEmp : table synthétique sur les IDENT présents dans la table des
sequence de non-emploi;
Data NonEmp; Merge SeqNonEmp Inscrits ; by ident ; run;
Data NonEmp; Merge NonEmp Demarche ; by ident ; run;
Data NonEmp; Merge NonEmp ChoC ; by ident ; run;
Data fhs2.NonEmp; Merge NonEmp ChoL ; by ident ;
SeqNonEmp=(SeqNonEmp=1);
Inscrits=(Inscrits=1);
Demarche=(Demarche=1);
ChoC=(Choc=1);
ChoL=(ChoL=1);
ChoCL=(Choc=1 or ChoL=1);
run;
Title "Exploratoire - Table des individus ayant des séquences de non-emploi";
Proc freq data=fhs2.NonEmp; tables SeqNonEmp/nopercent nocol norow;
```

```
run;
Proc freq data=fhs2.NonEmp;
tables Choc*ChoL (ChoCL ChoC ChoL)*inscrits Demarche*inscrits/nopercent nocol
norow;
run;


/*****************************************************/
/*****************************************************/
***** Agrégation de la TABLE avec la synthèse de l'identification Pôle EMPLOI via
la table SELEC2;
Proc sort data=fhs2.NonEmp;by ident; run;
Proc sort data=fhs2.selection;by ident; run;

Data fhs2.GeneFHS_noemp;
merge fhs2.NonEmp
      fhs2.selection (keep=ident selection);
by ident;
SeqNonEmp=(SeqNonEmp=1);
run;
Proc freq data=fhs2.GeneFHS_noemp; tables SeqNonEmp*Selection/nopercent;
run;
Title" restriction au filtre=1";
Proc freq data=fhs2.GeneFHS_noemp ; tables SeqNonEmp*Selection/nopercent;
Proc freq data=fhs2.GeneFHS_noemp ; tables (ChoCL ChoC ChoL
inscrits)*Selection/nopercent;
run;


/*****************************************************/
/*****************************************************/
/*****************************************************/
/* Association avec la table individu de Génération */
/* Restreint aux cas simples (de table ID_UNIQUE sans anomalies
/* En repartant de la table SELEC2 créée juste au dessus
/*****************************************************/
/*****************************************************/
/*****************************************************/

Proc sort data=fhs2.selection;by ident;run;
Proc sort data=gene.g047ansindividus8sspi;by ident;run;

Data fhs2.GeneFHS;merge gene.g047ansindividus8sspi (in=a) fhs2.selection;by ident;
if a ;

If Selection ="Identification simple 1" then Selection="Identification simple  ";

nscho_tot=nscho_07+nscho_09+nscho_11;
If nscho_tot ge 13 then nscho_tot = 13; * pour lisibilité du tableau, on regroupe
dans "13" les 13 et plus;
If typotraj_07=. then typotraj_07=0; *non répondant à 7 ans;
If typotraj_11=. then typotraj_11=0; *non répondant à 7 ans;
run;

Title "Exploratoire basique : identification selon le sexe ";
Proc freq data=fhs2.GeneFHS; tables q1*selection /nopercent nocol norow; run;

Title "Exploratoire basique : identification selon type de trajectoire à 3 ans ";
Proc freq data=fhs2.GeneFHS; tables typotraj_07*selection /nopercent nocol norow;
run;

Title "Exploratoire basique : identification selon type de trajectoire à 5 ans";
Proc freq data=fhs2.GeneFHS; tables typotraj_11*selection/nopercent nocol norow;
run;

Title "Exploratoire basique : identification selon le nbre de séquence de non-
emploi ('13'='13 et plus')";
Proc freq data=fhs2.GeneFHS; tables nscho_tot* selection/ nopercent norow nocol;
run;

Title "Exploratoire basique : identification selon le niveau de sortie(NIVSOR9)";
```

```
Proc freq data=fhs2.GeneFHS; tables nivsor9* selection/ nopercent nocol norow; run;

Title "Exploratoire basique : identification selon le niveau de sortie(NIVSOR9)-
PONDERE";
Proc freq data=fhs2.GeneFHS; tables nivsor9* selection/ nopercent nocol; weight
pondef; run;

Title "Exploratoire basique : identification selon le plus haut diplôme (PHDIP)";
Proc freq data=fhs2.GeneFHS; tables phdip* selection/ nopercent nocol norow; run;

Title "Exploratoire basique : identification selon le niveau de sortie(PHDIP)-
PONDERE";
Proc freq data=fhs2.GeneFHS; tables phdip* selection/ nopercent nocol; weight
pondef; run;

Proc freq data=fhs2.GeneFHS; tables typotraj_07 nivsor9 phdip/ nopercent nocol;
run;



/*********************************************************/
/*********************************************************/
/*********************************************************/
/*********************************************************/
/***  ESSAI D'EXPLOITATION EN ESSAYANT DE RETROUVER
/***  LES BONNES INFO SUR LA BASE D'UNE CLEF IDTIDV-MOIS NAIS-ANNEE NAIS
/*********************************************************/
/*********************************************************/
/*********************************************************/
/*********************************************************/

/* TEMPORAIRE1 : Table de passage entre l'IDENT de génération et IDENT2*/
/* pour les seules individus identifiés de façon simple */
Data Temporaire1;set fhs2.selection;
If Selection in ("Identification simple 1","Identification simple  ");
If suffixe ="A" then do; an_nais=an_nais_A;mois_nais=mois_nais_A;end;
If suffixe ="B" then do; an_nais=an_nais_B;mois_nais=mois_nais_B;end;
Keep Selection Echo Ident Ident2 an_nais mois_nais;
run;


/* TEMPORAIRE2 : Table de passage entre l'IDENT2 et IDTIDV */
/* construite avec les seules tables ID_UNIQUE et ID_MULT*/
/* puisqu'on se limitera ensuite aux identifications simples */
Data Temp2_unique ;set fhs2.id_unique;
ident2 = substr(identifiant2,1,7);
Keep ident2 idtidv;
run;

Data Temp2_mult0 ;set fhs2.id_mult;
ident2 = substr(identifiant2,1,7);
Keep ident2 idtidv;
run;

%Macro Temp2_mult;
    %do i = 1 %to 8;
       Data Temp2_mult&i. ;set fhs2.id_mult;
       ident2 = substr(identifiant2,1,7);
       If Idtidv&i. ne "";
       idtidv=Idtidv&i.;
       Keep ident2 idtidv;
       run;
       %end ;
%mend;
%Temp2_mult;
run;

Data Temporaire2; set Temp2_unique Temp2_mult0 Temp2_mult1 Temp2_mult2 Temp2_mult3
                      Temp2_mult4 Temp2_mult5 Temp2_mult6 Temp2_mult7 Temp2_mult8;
run;
```

```
/* TEMPORAIRE3 : Table de passage entre IDTIDV et l'IDENT de génération */
/* pour les seules individus identifiés de façon simple */
/* (donc normalement, une ligne par IDTIDV mais non vérifié)*/
/* Et creation de la clef IDTIDV_Mois_nais-An_Nais */


Proc sort data=Temporaire1; by ident2;run;
Proc sort data=Temporaire2; by ident2;run;
Data Temporaire3 ; merge Temporaire2 temporaire1 (in=a);
by ident2 ; if a;
length ID $11. Date $4.;
ID=substr(IDTIDV,1,11);
Date=put(100*an_nais+mois_nais,$4.);
Clef=ID!!"-"!!Date;
run;
Title "Verification de la construction de la clef";
proc print data=temporaire3 (obs=10);run;

/* TEMP_DE : table extraite de DE (selection de variables) avec ajout de la clef*/
Data Temp_de; set fhs2.fus_de;
length ID $11. Date $4.;
ID=substr(IDTIDV,1,11);
Mois=month(datnais);an=year(datnais)-1900;
Date=put(100*an+mois,$4.);
Clef=ID!!"-"!!Date;
Keep IDTIDV NDEM SEXE DATNAIS DEPCOM NIVFOR DATINS MOTINS DATANN MOTANN NBAR78
NBAR79 CATREGR
     ID Mois an Date Clef;
run;

/* TEMP_DE2 : table précédente avec injection de l'identifiant Génération sur la
base de la clef*/
/* Creation de la variable DUR_INS sur la durée d'inscription en mois. Le mois
d'inscription */
/* compte pour 1, de même que le mois de sortie, pour rapprocher avec les mois
d'activité réduite*/
/* Création de la variable NB0AR de mois sans activité réduite */
Proc sort data=temporaire3;by Clef;run;
Proc sort data=Temp_de;by Clef;run;

Data fhs2.TEMP_DE2 ; merge Temp_De Temporaire3 (keep=clef IDENT in=a);
By clef; if a;
Mois_ins=month(DATINS);An_ins=year(datins);
Mois_ann=month(DATANN);An_ann=year(datann);
If datins ne . and datann=. then do ;Mois_ann="03";An_ann="2014";end;
DUR_INS=12*(An_ann-An_ins)+Mois_ann-Mois_ins+1;
NB0AR=DUR_INS-NBAR78-NBAR79;
NBAR=NBAR78+NBAR79;
run;

Title "Catégorie des demandes";
Proc freq data=fhs2.TEMP_DE2; tables catregr; run;


/* TEMP_SYNTHESE_DE : table sommant les variables de mois par identifiants Géné*/
/*********************************************************/
/* ATTENTION : on se restreint aux catégories 1, 2 ou 3 */
/* Car l'exploratoire est ensuite fait sur les durées   */
/* d'inscription en catégorie 1, 2 OU 3                 */
/*********************************************************/

Proc sort data=fhs2.Temp_DE2; by ident;
Proc summary data=fhs2.Temp_DE2 (where=(catregr in ("1","2","3")));
class ident ;
var NB0AR NBAR78 NBAR79 DUR_INS NBAR;
Output out=Temp_Synthese_De Sum=;
run;
Data Temp_Synthese_De; set Temp_Synthese_De; if _type_=1;
run;
```

```
/* ESSAI_GENE_DE : Appariement de la table de synthèse des info de DE */
/* avec la table Génération enrichie crée plus haut*/

proc sort data=Temp_Synthese_De;by ident;run;
Proc sort data=fhs2.GeneFHS;by ident;run;

Data fhs2.Essai_Gene_DE;merge fhs2.GeneFHS (in=a) Temp_Synthese_De ;
by ident; if a;
Nb_DE123=_Freq_;
If _freq_ ge 1 then present=1;
else do; present=0; NB0AR=0; NBAR78=0; NBAR79=0; DUR_INS=0; NBAR=0;end;
run;

Title "Nbre de personnes ayant eu au moins une DE de catégorie 1, 2 3 récupérée";
Proc freq data=fhs2.Essai_Gene_DE; tables (typotraj_11 nivsor9 Phdip)*present/
nocol nopercent norow;
run;

/**** TABLEAUX DE SORTIE- LOT 1 ***/
/**** Durées sans Activité réduite (champ = tous, y compris ceux jamais inscrits)
***/

Title "Exploratoire sur les durées sans AR : selon type de trajectoire à 5 ans";
Title2 "Champ = tous, meme les non inscrits";
Proc univariate data=fhs2.Essai_Gene_DE;
Class typotraj_11; Var NB0AR ;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
Proc print data=Univariate_Typo; run;

Title "Exploratoire sur les durées sans AR : selon le niveau de sortie";
Proc univariate data=fhs2.Essai_Gene_DE;
Class Nivsor9; Var NB0AR ;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
Proc print data=Univariate_Typo; run;

Title "Exploratoire sur les durées sans AR : selon le plus haut diplome";
Proc univariate data=fhs2.Essai_Gene_DE;
Class PHDIP; Var NB0AR ;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
Proc print data=Univariate_Typo;run;


Title "Exploratoire sur les durées sans AR : selon le nbre de sequences de non-
emploi";
Proc univariate data=fhs2.Essai_Gene_DE;
Class nscho_tot; Var NB0AR ;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
Proc print data=Univariate_Typo;run;

/**** TABLEAUX DE SORTIE - LOT 2 ***/
/**** Durées sans Activité réduite (champ = uniquement ceux inscrits) ***/

Title "Exploratoire sur les durées sans AR : selon type de trajectoire à 5 ans";
Title2 "Champ = Uniquement les inscrits";
Proc univariate data=fhs2.Essai_Gene_DE (where=(present=1));
Class typotraj_11; Var NB0AR ;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
Proc print data=Univariate_Typo;run;

Title "Exploratoire sur les durées sans AR : selon le niveau de sortie";
Proc univariate data=fhs2.Essai_Gene_DE (where=(present=1));
Class Nivsor9; Var NB0AR ;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
```

```
Proc print data=Univariate_Typo;run;

Title "Exploratoire sur les durées sans AR : selon le plus haut diplome";
Proc univariate data=fhs2.Essai_Gene_DE (where=(present=1));
Class PHDIP; Var NB0AR ;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
Proc print data=Univariate_Typo;run;

Title "Exploratoire sur les durées sans AR : selon le nbre de sequences de non-
emploi";
Proc univariate data=fhs2.Essai_Gene_DE (where=(present=1));
Class nscho_tot; Var NB0AR ;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
Proc print data=Univariate_Typo;run;



/**** TABLEAUX DE SORTIE  - LOT 3 ***/
/**** Durées de l'activité réduite (champ = tous, y compris ceux jamais inscrits)
***/

Title "Exploratoire sur les durées de l'AR : selon type de trajectoire à 5 ans";
Title2 "Champ = tous, meme les non inscrits";
Proc univariate data=fhs2.Essai_Gene_DE;
Class typotraj_11; Var NBAR ;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
Proc print data=Univariate_Typo; run;

Title "Exploratoire sur les durées de l'AR  : selon le niveau de sortie";
Proc univariate data=fhs2.Essai_Gene_DE;
Class Nivsor9; Var NBAR ;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
Proc print data=Univariate_Typo; run;

Title "Exploratoire sur les durées de l'AR  : selon le plus haut diplome";
Proc univariate data=fhs2.Essai_Gene_DE;
Class PHDIP; Var NBAR ;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
Proc print data=Univariate_Typo;run;

Title "Exploratoire sur les durées de l'AR  : selon le nbre de sequences de non-
emploi";
Proc univariate data=fhs2.Essai_Gene_DE;
Class nscho_tot; Var NBAR ;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
Proc print data=Univariate_Typo;run;

/**** TABLEAUX DE SORTIE - LOT 4 ***/
/**** Durées de l'activité réduite (champ = uniquement ceux inscrits) ***/

Title "Exploratoire sur les durées de l'AR : selon type de trajectoire à 5 ans";
Title2 "Champ = Uniquement les inscrits";
Proc univariate data=fhs2.Essai_Gene_DE (where=(present=1));
Class typotraj_11; Var NBAR ;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
Proc print data=Univariate_Typo; run;

Title "Exploratoire sur les durées de l'AR  : selon le niveau de sortie";
Proc univariate data=fhs2.Essai_Gene_DE (where=(present=1));
Class Nivsor9; Var NBAR ;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
Proc print data=Univariate_Typo; run;
```

```
Title "Exploratoire sur les durées de l'AR  : selon le plus haut diplome";
Proc univariate data=fhs2.Essai_Gene_DE(where=(present=1));
Class PHDIP; Var NBAR ;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
Proc print data=Univariate_Typo;run;

Title "Exploratoire sur les durées de l'AR  : selon le nbre de sequences de non-
emploi";
Proc univariate data=fhs2.Essai_Gene_DE (where=(present=1));
Class nscho_tot; Var NBAR ;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
Proc print data=Univariate_Typo;run;


/**** TABLEAUX DE SORTIE  - LOT 5 ***/
/**** Durées totale d'inscription en 1, 2 ou 3 (champ = tous, y compris ceux jamais
inscrits) ***/

Title "Exploratoire sur les durées d'inscription en 1, 2 ou 3 : selon type de
trajectoire à 5 ans";
Title2 "Champ = tous, meme les non inscrits";
Proc univariate data=fhs2.Essai_Gene_DE;
Class typotraj_11; Var DUR_INS ;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
Proc print data=Univariate_Typo; run;

Title "Exploratoire sur les durées d'inscription en 1, 2 ou 3 : selon le niveau de
sortie";
Proc univariate data=fhs2.Essai_Gene_DE;
Class Nivsor9; Var DUR_INS;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
Proc print data=Univariate_Typo; run;

Title "Exploratoire sur les d'inscription en 1, 2 ou 3 : selon le plus haut
diplome";
Proc univariate data=fhs2.Essai_Gene_DE;
Class PHDIP; Var DUR_INS ;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
Proc print data=Univariate_Typo;run;

Title "Exploratoire sur les d'inscription en 1, 2 ou 3 : selon le nbre de sequences
de non-emploi";
Proc univariate data=fhs2.Essai_Gene_DE;
Class nscho_tot; Var DUR_INS ;
Output out=Univariate_Typo Mean=Moyenne P25=P25 Median=Medianne P75=P75 n=effectif;
run;
Proc print data=Univariate_Typo;run;

/*****************************************************/
/*****************************************************/
/*****************************************************/
/* Construction d'un calendrier avec FHS :
/*** Travail sur la table des demandes d'emploi
/*** de catégorie 1, 2 ou 3 sans distinction de l'activité réduite
/*** pour créer des variables MOISFH1 (Nov03) à MOISFH98 (Dec11)
/*** qui vaut 1 si le mois i est en demande d'emploi, 0 sinon
/*******************/

Data CalFHS1; set fhs2.Temp_DE2 (where=(catregr in ("1","2","3")));
If datins="" and datann="" then delete;
andeb=year(datins);
anfin=year(datann); if anfin=. then anfin=9999;

/* identification des demandes achevées avant novembre 2003=MOIS1 */
DE_precoce=(datann < MDY(11,1,2003));
If datann="" then DE_precoce=9;*demande encore en cours lors de l'extraction;
```

```
/* identification des demandes commençant après décembre 2011 =MOIS98 */
DE_tardif=(datins > MDY(12,31,2011));
run;

Title " repérage des demandes d'emploi hors période du calendrier Génération";
Proc freq data=calFHS1; tables DE_precoce DE_tardif andeb*DE_tardif
anfin*DE_precoce DE_tardif*DE_precoce/nopercent norow nocol;
run;

Data CalFHS1;set CalFHS1;
* suppression des demandes d'emploi hors période d'observation;
If DE_tardif = 1 then delete;
If DE_precoce = 1 then delete;

* construction des variables de mois d'inscription pour que le mois 1=nov 2003 et
le mois 98=dec 2011;
indexdeb=(year(datins)-2003)*12+month(datins)-10;
indexfin=(year(datann)-2003)*12+month(datann)-10;
If indexfin="" then indexfin=999;
%Macro Indexation1;
      %do i = 1 %to 98 ;
      If indexdeb le &i. and indexfin ge &i. then moisPE&i.=1;
      Else moisPE&i.=0;
      %end ;
%mend;
%Indexation1;
run;
* passage d'une table de demandes à une table des demandeurs;
* les variables Mois_PEx compte le nbre de demandes en cours pour l'individu IDENT
le mois x=1 à 98;
Data CalFHS2; set CalFHS1;
Retain TempmoisPE1-TempmoisPE98 IDTIDV1 n;
If IDENT=IDTIDV1 then do;
   n=n+1;
   %Macro Indexation2; %do i = 1 %to 98;
    moisPE&i.=moisPE&i.+TempmoisPE&i.;
    TempmoisPE&i.=moisPE&i.;
    %end ;
   %mend;
   %Indexation2;
   end;
Else do;
   n=1;
   %Macro Indexation3; %do i = 1 %to 98;
    TempmoisPE&i.=moisPE&i.;
    %end ;
   %mend;
   %Indexation3;
   IDTIDV1=IDENT;
   end;
run;

Proc sort data=CalFHS2;by ident n;run;
Data fhs2.CalFHS2; set CalFHS2; by IDENT; if last.IDENT;
Drop TempmoisPE1--TempmoisPE98;
run;

Title " Distribution du nombre de demandes d'emploi comptabilisés par jeunes
inscrits ";
Proc freq data= fhs2.CalFHS2; tables n/nopercent;
Proc univariate data= fhs2.CalFHS2; var n;
run;
Title "Nombre de demandes d'emplois comptabilisées par jeunes inscrits pour chaque
mois i=1 (nov03) à 98 (dec11)";
Proc freq data= fhs2.CalFHS2; tables moisPE1--moisPE98;

/****** rapprochement de CALFHS avec Génération */
/****** pour construire des calendriers mensuels d'inscrits par type selon
TYPOTRAJ_11*/
```

```
proc sort data=fhs2.CalFHS2;by ident;run;
Proc sort data=fhs2.GeneFHS;by ident;run;

Data fhs2.Essai_Gene_calendrier_DE;merge fhs2.GeneFHS (in=a keep=ident typotraj_11)
fhs2.CalFHS2 ;
by ident; if a;
%Macro Indexation4;
       %do i = 1 %to 98 ;
       If moisPE&i.=. then moisPE&i.=0;
       %end ;
%mend;
%Indexation4;
run;

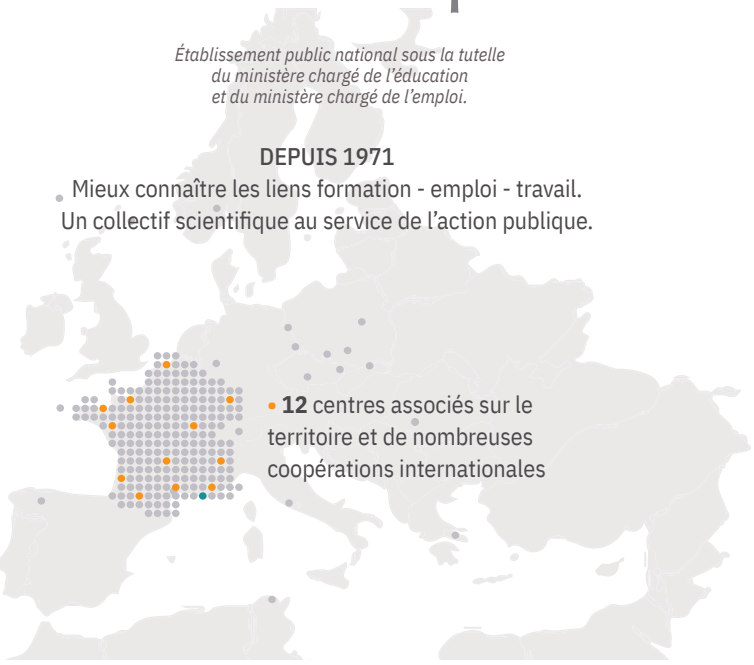
Proc summary data=fhs2.Essai_Gene_calendrier_DE;
var moisPE1--moisPE98;
class typotraj_11;
Output out=Calendrier sum=;
run;
Proc print data=calendrier;
run;
```

Céreq

Établissement public national sous la tutelle du ministère chargé de l'éducation et du ministère chargé de l'emploi.

DEPUIS 1971

Mieux connaître les liens formation - emploi - travail.
Un collectif scientifique au service de l'action publique.



• **12** centres associés sur le territoire et de nombreuses coopérations internationales

+ d'infos
et tous les travaux

À explorer
www.cereq.fr

+ de 600 publications
Accessibles librement