UN OUTIL D'APPARIEMENT SUR IDENTIFIANTS INDIRECTS

L'EXEMPLE DU SYSTÈME D'INFORMATION SUR L'INSERTION DES JEUNES

*Loïc Midy**

Le service statistique du ministère de l'Éducation Nationale (la Depp) réalise depuis longtemps deux enquêtes annuelles d'insertion dans la vie active, portant sur les sortants de l'apprentissage et de la voie professionnelle scolaire. Mais elles ne permettent pas de publier des statistiques au niveau des établissements, comme requis depuis 2018 par la loi pour la Liberté de choisir son avenir professionnel. Afin de répondre à ce besoin, la Depp et la Dares ont construit un nouveau dispositif, appelé InserJeunes, qui apparie des sources administratives, principalement sur identifiants indirects. Cette problématique était centrale pour la réussite du dispositif, qu'il s'agisse des choix méthodologiques, du paramétrage des algorithmes ou des développements informatiques. Le processus choisi comporte, classiquement, cinq étapes : normalisation des données, indexation, calcul de similarités, classification supervisée et évaluation de la qualité. Le choix des méthodes adaptées est présenté à travers un cas réel de production : si elles ont été implémentées à travers un outil d'appariement développé spécifiquement pour InserJeunes, elles restent transposables dans des environnements similaires.

The French Statistical Office of the National Education Ministry (the* DEPP*) carry out two surveys on the labour market integration of the students who just finished their study as apprentice or in vocational school path. But they don't enable to publish statistics at the establishment level as required by the 2018 Act for the Liberty to choose one professional future. So the* DEPP *and the* DARES *(Statistical Office of the Labour Ministry) have designed a new information system,* InserJeunes*, based on the record linkage of administrative data sources. Record linkage is central in this device, from the methodological, the algorithmic and the IT development standpoints. In* InserJeunes*, the record linkage process has five steps: data normalisation, indexing, similarities calculation, supervised classification and quality evaluation. The methods are presented through a real production example from the* InserJeunes *information system. They were implemented through a record linkage tool developed by the* InserJeunes *team, which can be reused for other record linkage processes.

* Directeur du projet Mesure de l'insertion des jeunes, Depp,
loic.midy@education.gouv.fr

MIEUX CONNAÎTRE L'EFFICACITÉ DES ÉTABLISSEMENTS EN MATIÈRE D'INSERTION DES JEUNES

L'orientation des élèves se construit tout au long de la scolarité avec des étapes clés en fin de troisième, seconde et terminale. Ainsi, l'orientation en voie professionnelle peut commencer dès la fin de troisième avec un choix entre apprentissage ou voie professionnelle scolaire. L'insertion dans l'emploi étant la première finalité de la formation professionnelle, connaître les taux d'insertion des formations initiales permet d'éclairer les choix des jeunes et de leur famille.

La direction de l'Évaluation, de la prospective et de la performance (Depp) réalise, depuis le début des années 1990, deux enquêtes d'insertion annuelles[1] permettant de suivre l'entrée dans la vie active des sortants d'apprentissage et de voie professionnelle scolaire. Ces opérations apportent des informations précieuses, mais ne permettent pas de publier des statistiques au niveau établissement, compte tenu des taux de réponse observés[2].

Or, la loi du 5 septembre 2018 pour la Liberté de choisir son avenir professionnel[3] prévoit la publication de statistiques par établissement sur le parcours scolaire et l'insertion dans l'emploi des jeunes en formation professionnelle. Afin de répondre à ce besoin, la Depp et la direction de l'Animation de la recherche, des études et des statistiques (Dares) ont construit un nouveau dispositif[4] : InserJeunes est basé sur l'appariement de sources administratives exhaustives, relatives à la scolarité des élèves et apprentis, à la réussite aux examens, aux contrats d'apprentissage et aux contrats salariés de la déclaration sociale nominative (DSN[5]). Les premiers résultats ont été diffusés début février 2021.

> *« InserJeunes est basé sur l'appariement de sources administratives exhaustives, relatives à la scolarité des élèves et apprentis, à la réussite aux examens, aux contrats d'apprentissage et aux contrats salariés de la déclaration sociale nominative. »*

Côté Éducation nationale, les bases sur les élèves peuvent s'apparier sur un identifiant spécifique, l'identifiant national élève (INE)[6]. Mais il n'existe pas d'identifiant commun pour apparier les bases « scolarité » des élèves et apprentis, avec les contrats salariés de la DSN. Ces appariements ne sont donc possibles qu'indirectement, et réalisés à partir des cinq variables que sont les noms, prénoms, date et lieu de naissance et le sexe. La mise en place d'un outil d'appariement sur identifiants indirects performant et de qualité est donc un enjeu central d'InserJeunes. Cette problématique bien connue des statisticiens fait l'objet d'une vaste littérature (voir par exemple (Kilss et Alvey, 1985)).

Cet article présente la démarche d'ensemble retenue, les sources principales utilisées, le cadre juridique à respecter ainsi que les choix effectués entre les différentes méthodes et outils informatiques d'appariement sur identifiants indirects.

1. Les enquêtes *Insertion dans la vie active* (IVA) pour les sortants de la formation professionnelle des lycées, et *Insertion professionnelle des apprentis* (IPA).
2. De l'ordre de 60 %.
3. Voir les références juridiques en fin d'article.
4. InserJeunes a bénéficié d'un financement du fonds pour la transformation de l'action publique.
5. Pour plus d'information sur la DSN, voir (Humbert-Bottin, 2018).
6. L'INE, mis en place en 2017, est un identifiant unique de chaque élève.

LES PRINCIPES DU DISPOSITIF INSERJEUNES

Le processus principal s'articule autour de plusieurs phases ***(figure 1)***. Dans un premier temps, pour une année scolaire donnée, le champ des élèves en année terminale de formation est calculé en mobilisant trois bases de données administratives « scolarité »[7], chacune couvrant une partie du champ d'InserJeunes : l'apprentissage, la voie professionnelle scolaire dans un établissement du ministère de l'Éducation nationale et celle dans un établissement du ministère de l'Agriculture. Ces bases contiennent les variables indirectement identifiantes, ainsi que l'INE, et des informations sur l'établissement et la formation suivie.

Dans une deuxième phase, on établit le champ des élèves sortants, c'est-à-dire ceux qui ne sont plus en formation. Pour ce faire, on recherche, principalement sur l'INE, si ces élèves sont encore présents l'année scolaire suivante dans l'ensemble des bases de données élèves disponibles c'est-à-dire les trois bases déjà mobilisées dans la phase précédente ainsi que trois bases supplémentaires[8] afin d'être le plus exhaustif possible[9]. Tout élève retrouvé est noté comme étant toujours en étude, les autres sont appelés les sortants de formation[10]. Cela permet d'établir le **taux de poursuites d'études**.

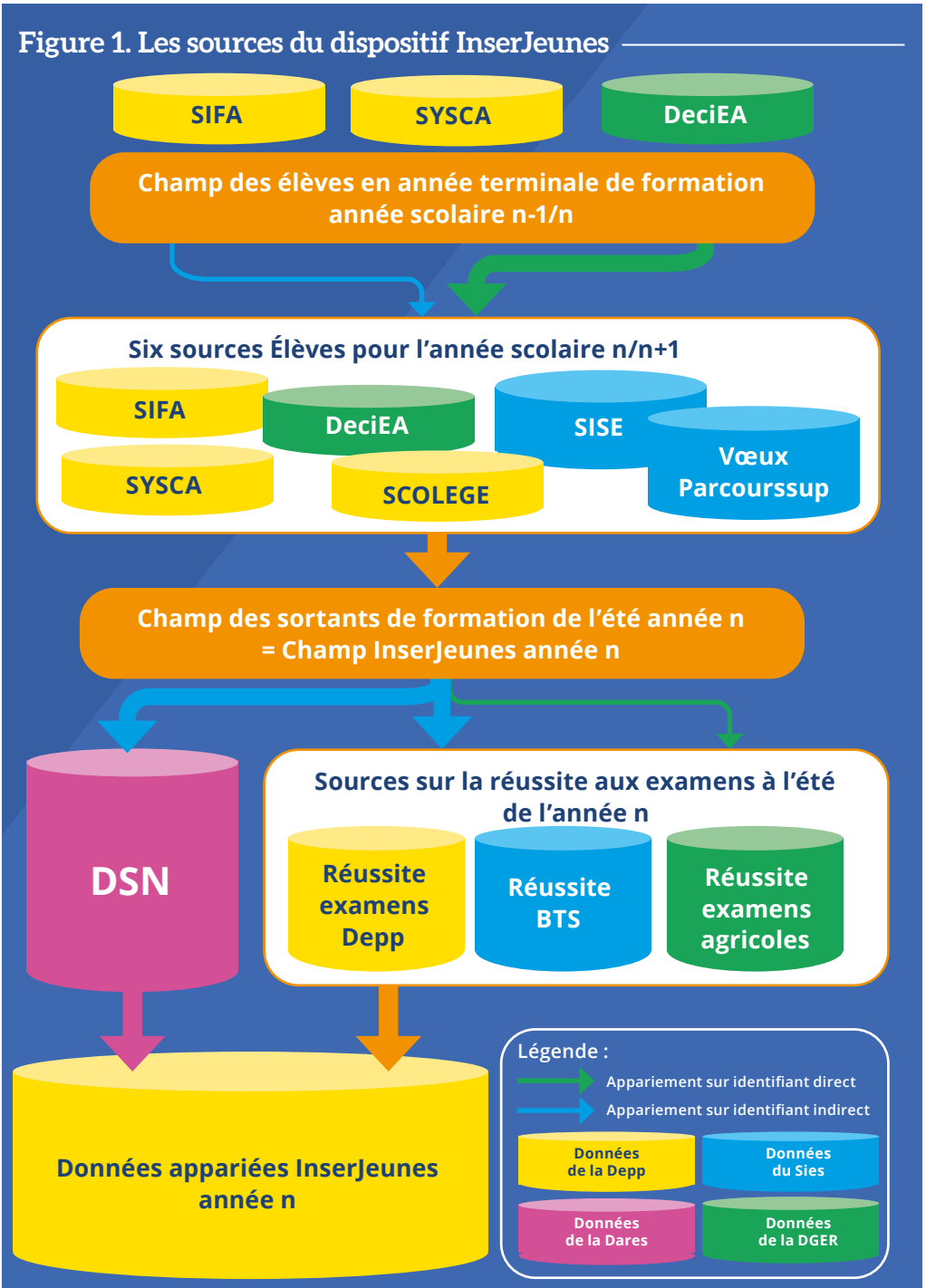
Dans la troisième phase, les bases élèves/apprentis sont enrichies avec leur réussite aux examens (selon les cas cet appariement est réalisé sur l'INE ou sur identifiants indirects), ce qui permet de calculer le **taux d'interruption en cours de formation**.

Enfin, lors de la quatrième phase, les bases élèves/apprentis sortants sont appariées sur identifiants indirects avec la DSN[11], ce qui permet de mesurer un **taux d'emploi salarié en France des sortants** puis la **valeur ajoutée de l'établissement sur ce taux d'emploi**[12]. La DSN contient des informations détaillées sur les contrats salariés (type de contrat, salaire, quotité de travail, catégorie socioprofessionnelle, etc.) ainsi que sur l'établissement employeur (secteur, commune d'implantation, etc.) : de ce fait, InserJeunes pourra permettre également de réaliser des études statistiques, par exemple, sur l'adéquation formation/emploi.

Dans InserJeunes, le taux d'appariement ne donne aucune indication du niveau de qualité du processus. Par exemple, lorsqu'un sortant n'est pas apparié avec la DSN, il n'est pas possible de savoir si c'est parce qu'il n'est pas en emploi salarié ou en raison d'une erreur dans le processus d'appariement. Mais le dispositif comporte un appariement sur identifiants indirects appelé « appariement qualité », annuel, pour lequel le taux théorique est de 100 % : il s'agit du rapprochement du fichier recensant les apprentis au 31 décembre ayant un contrat d'apprentissage actif[13] avec la DSN. Ainsi, le taux d'appariement réel obtenu constitue un indicateur de la qualité du processus d'appariement.

7. SIFA (Système d'information de la formation des apprentis) pour les apprentis, SYSCA (Système d'information statistique consolidé académique) pour les élèves de voie professionnelle scolaire du ministère de l'Éducation nationale et DeciEA pour les élèves de voie professionnelle scolaire du ministère de l'Agriculture.
8. SIFA, SYSCA, DeciEA plus les élèves du secteur privé hors contrat avec la source SCOLEGE et le supérieur *via* les enquêtes SISE (Système d'information sur le suivi des étudiants) et les vœux validés *via* Parcoursup dans un institut de formation en soins infirmiers.
9. En particulier en prenant en compte les poursuites dans l'enseignement supérieur.
10. En réalité quelques sortants de formation peuvent en fait être encore en étude : par exemple on ne repère pas les poursuites d'études à l'étranger.
11. Pour être tout à fait précis, on utilise la source DMMO (Déclaration de mouvement de main d'œuvre) basée sur la DSN (Déclaration sociale nominative).
12. La notion de valeur ajoutée est un concept largement développé dans un précédent article, voir (Evain, 2020).
13. En dehors de la fonction publique, car les emplois publics ne sont pas encore intégrés en DSN.

Figure 1. Les sources du dispositif InserJeunes



SIFA : système d'information de la formation des apprentis
SYSCA : système d'information statistique consolidé académique
DeciEA pour les élèves de voie professionnelle scolaire du ministère de l'Agriculture
SCOLEGE : scolarité léger, application web développée par la Depp
SISE : système d'information sur le suivi des étudiants

> *Le processus statistique InserJeunes comporte au total dix appariements sur identifiants indirects pour chaque année scolaire.*

Le processus statistique InserJeunes comporte au total dix appariements sur identifiants indirects pour chaque année scolaire. La problématique des appariements sur identifiants indirects est donc centrale. Cela implique de mettre au point un processus d'appariement général, puis d'en faire une implémentation informatique générique et rapide.

AU CŒUR DU SYSTÈME D'INFORMATION : UN PROCESSUS D'APPARIEMENT EN CINQ ÉTAPES

Dans InserJeunes, chaque appariement sur identifiants indirects est réalisé entre deux tables individuelles[14] sans double compte. Le processus d'appariement retenu pour InserJeunes ***(figure 2)*** comporte cinq étapes successives, comme c'est aussi le cas dans la présentation de Peter Christen ***(figure 3)*** (Christen, 2012).

Les données sont tout d'abord normalisées. Puis vient l'étape d'indexation, qui consiste à établir une liste de taille raisonnable de paires « potentiellement intéressantes ». Une paire correspond au croisement d'une ligne de la première table avec une ligne de la seconde table. Chaque paire comporte donc un/des noms, un/des prénoms, une date de naissance, un lieu de naissance et une variable sexe provenant de chacune des deux tables qu'on apparie. En troisième lieu, une similarité est calculée pour chacun des cinq couples d'identifiants indirects de chaque paire (par exemple couple de noms, couple de dates de naissance). Quatrièmement, chaque paire est classifiée : les paires supposées relever du même individu (i.e. lorsque les cinq similarités calculées à l'étape précédente sont suffisamment élevées) sont acceptées et les autres sont rejetées. Enfin, la qualité du processus d'appariement est évaluée.

NORMALISER DES DONNÉES

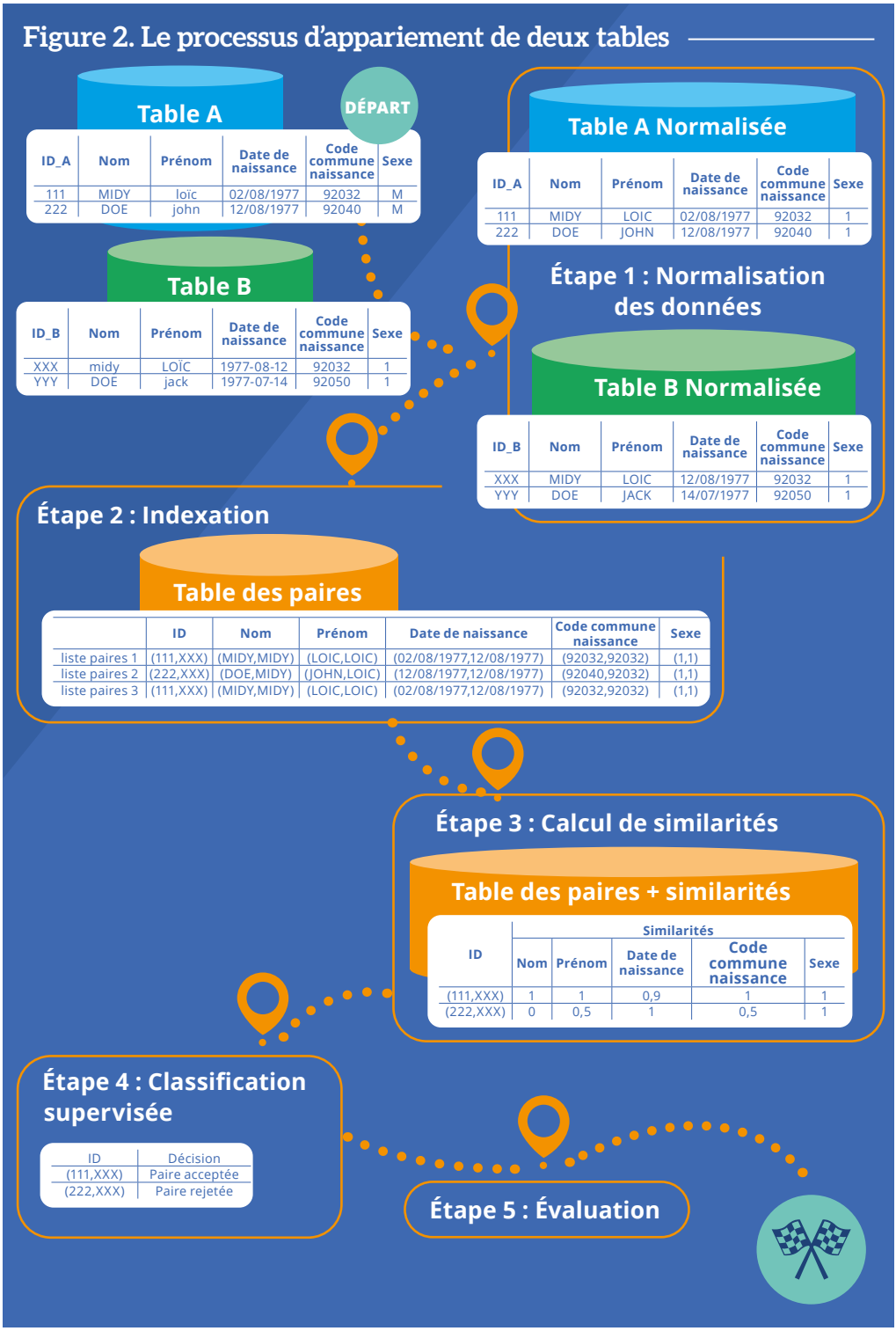
Les identifiants indirects utilisés dans l'appariement se présentent sous des formats hétérogènes dans les différentes sources mobilisées dans InserJeunes. La normalisation des données, première étape du processus d'appariement, consiste à les recoder selon une structure commune afin de faciliter les traitements ultérieurs.

Pour les noms et les prénoms, les traitements principaux suivants sont réalisés :

- mettre en majuscule[15] ;
- éliminer les caractères spéciaux ;
- remplacer les lettres spéciales par leur valeur canonique[16] ;
- éliminer les noms/prénoms d'une lettre et certains noms/prénoms de deux ou trois lettres afin de supprimer les termes peu informatifs comme DE et LE.

14. C'est-à-dire que l'unité d'observation est l'individu (ici élève ou apprenti).
15. Ce qui supprime au passage les accents.
16. Par exemple : Ç devient C, Ï devient I.

Figure 2. Le processus d'appariement de deux tables



Par ailleurs, jour, mois et année des dates de naissance sont stockés dans des variables différentes afin de pouvoir mener des calculs sur les dates de naissance même lorsqu'elles ne sont que partiellement renseignées.

Les sources mobilisées dans InserJeunes sont de bonne qualité. En effet, il n'y a pas de doublons dans les sources principales, il y a très peu de valeurs manquantes sur les identifiants indirects, et le code COG[17] de la commune de naissance, information plus précise que le libellé de la commune, est généralement fourni. Cela est dû au fait que pour tous les élèves, l'immatriculation au Répertoire national des identifiants élèves, étudiants et apprentis a déjà nécessité que l'ensemble des variables identifiantes soient fournies. De même, chaque salarié dans la source DSN fait l'objet d'une procédure de certification du NIR[18] associé, ce qui assure également un niveau de qualité élevé des variables identifiantes.

INDEXER LES DONNÉES : L'APPROCHE NAÏVE

Pour la deuxième étape, celle de l'indexation des données, une première approche naïve consiste à analyser tous les croisements possibles entre les deux tables. Mais le temps de traitement augmente de manière quadratique avec le nombre d'observations des tables à apparier et donc en pratique, cette méthode n'est plus applicable au-delà d'un certain seuil.

Dans le cas de l'appariement qualité, environ 315 000 apprentis sont rapprochés des 7,5 millions de salariés ayant un contrat actif en décembre de l'année considérée. L'analyse exhaustive de chacune des 2,3 billions de paires possible prendrait plusieurs jours voir plusieurs dizaines de jours[19].

> *« La méthode d'indexation retenue doit conjuguer deux objectifs en apparence contradictoires : élaborer une liste de paires la plus petite possible tout en veillant à obtenir le plus possible de paires relatives au même individu. »*

Par ailleurs, analyser l'ensemble des paires ne présente aucun intérêt. En effet, lorsque les noms ou les prénoms ou les dates de naissance sont très différents, il est extrêmement peu probable que la paire soit acceptée.

L'étape d'indexation a pour objectif d'établir une liste de paires « potentiellement intéressantes » de taille raisonnable. La méthode d'indexation retenue doit conjuguer deux objectifs en apparence contradictoires : élaborer une liste de paires la plus petite possible tout en veillant à obtenir dans cette liste le plus possible de paires relatives au même individu.

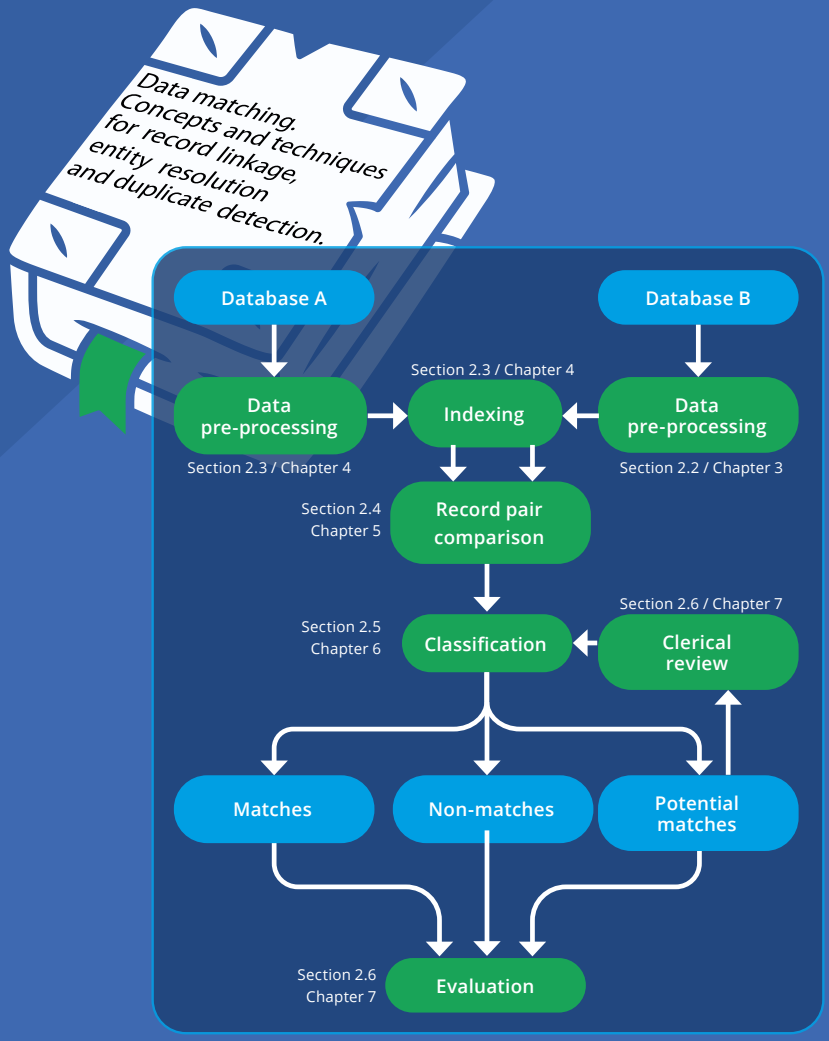
17. Le Code officiel géographique (COG) identifie chaque commune de France.
18. Numéro d'inscription au répertoire national d'identification des personnes physiques (RNIPP).
19. Si chaque ligne était analysée très rapidement, mettons en un cent millième de seconde, le temps de traitement total serait de 273 jours. Même en parallélisant les calculs sur un certain nombre de cœurs, cela resterait trop long.

INDEXER LES DONNÉES : L'APPROCHE CLASSIQUE PAR CLÉ DE BLOCAGE

La méthode la plus souvent utilisée pour indexer les données consiste à ne conserver que les paires qui partagent la même modalité d'une ou de plusieurs variables indirectement identifiantes, qu'on appelle les **clés de blocage** (Christen, 2012 ; Jabot et Treyens, 2018). Par exemple si la clé de blocage est le code de la commune de naissance, cela veut dire que seules les paires partageant ce code seront retenues.

Cette méthode n'a pas été choisie pour InserJeunes, car elle présente plusieurs inconvénients. Tout d'abord, elle produit une liste de paires « potentiellement intéressantes » qui reste encore trop importante. En outre, elle conduit à écarter à tort certaines paires. Par exemple,

Figure 3. Les cinq étapes définies par Christen



si la clé de blocage est le code de la commune de naissance et si cette variable est mal renseignée pour un individu, alors il ne sera jamais apparié. Enfin, il n'est pas possible d'appliquer de clé de blocage sur les noms ou les prénoms car à la moindre faute de frappe ou d'orthographe, les modalités de la clé de blocage seront différentes. Pour résoudre ce problème, on peut certes remplacer les noms et prénoms par leur version phonétisée. Il existe pour ce faire de nombreux algorithmes phonétiques ***(encadré 1)***. Prenons l'exemple d'une paire avec les prénoms « christina » et « kristina ». Ces deux prénoms auront la même version phonétisée avec *Phonex*, (soit *c623)*. *A contrario*, « peter » et « pedro » ont la même version phonétisée avec *Soundex* (soit *p360*), ce qui veut dire qu'on conservera, selon les algorithmes, une liste trop importante de paires « potentiellement intéressantes ». De plus, ces algorithmes phonétiques ont été initialement développés pour la langue anglaise et ils n'ont pas tous été adaptés pour la langue française.

INDEXER LES DONNÉES : L'APPROCHE RETENUE DANS INSERJEUNES

Compte tenu des inconvénients de l'approche classique par clé de blocage, une méthode d'indexation spécifique a été développée pour InserJeunes.

Dans un premier temps, un appariement exact entre les deux tables est réalisé sur l'ensemble des champs suivants : le premier nom, le premier prénom, le jour, le mois et l'année de naissance, le code de la commune de naissance et le sexe. Dans le cadre de l'appariement qualité, cette étape permet d'apparier environ 84 % des apprentis avec la source DSN. Une fois cette étape franchie, il ne reste à apparier qu'environ 50 000 apprentis avec 7,2 millions de salariés ayant un contrat actif en décembre. Le volume de travail est ainsi déjà divisé par un facteur six[20].

Dans un second temps, l'union (sans doublons) des trois listes de paires suivantes est établie :

- les paires qui ont une distance faible entre les premiers noms, une distance faible entre les premiers prénoms, même département de naissance et même année de naissance ;
- les paires qui ont même date de naissance et même département de naissance ;
- les paires qui ont même premier nom et même premier prénom.

InserJeunes utilise la **distance de Levenshtein** pour les noms et les prénoms. Celle-ci correspond au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'un nom/prénom à l'autre. L'union des différentes requêtes permet de bien couvrir tous les cas rencontrés fréquemment et ainsi de conserver presque toutes les paires « potentiellement intéressantes ». De plus, comme chaque requête est relativement précise, le nombre de paires « potentiellement intéressantes » retenues n'est pas trop élevé. Pour l'appariement qualité, cette méthode d'indexation conduit à retenir 1 million de paires soit un nombre raisonnable qui pourra être traité suffisamment rapidement lors des étapes ultérieures du processus.

Cette méthode d'indexation fonctionne très bien sur la volumétrie d'InserJeunes mais ne serait peut-être pas adaptée dans le cas de l'appariement de tables de plusieurs dizaines de millions de lignes.

20. Soit 315 000/50 000.

CALCULER LES SIMILARITÉS

La troisième étape consiste à enrichir chacune des paires déterminées lors de l'indexation, de cinq variables « similarités » sur nom, prénom, date et commune de naissance et sexe. Chaque similarité est une mesure du degré de ressemblance des identifiants indirects considérés. Trois natures de similarités différentes sont utilisées dans InserJeunes selon la nature des variables mobilisées comme identifiants indirects.

> *« Chaque similarité est une mesure du degré de ressemblance des identifiants indirects considérés. »*

Tout d'abord, la **similarité de Jaro-Winkler** est mise en œuvre pour les noms et prénoms. Cette dernière est une adaptation de la similarité de Jaro développée par le statisticien Winkler qui ajoute une « bonification » lorsque les deux chaînes que l'on compare débutent par un préfixe commun. L'algorithme de calcul de la similarité de Jaro entre deux chaînes de caractères est le suivant :

- on commence par calculer un *facteur éloignement* qui est égal à la longueur de la chaîne la plus longue divisée par 2 moins 1 (exemple : si on compare DWAYNE et DUANE, l'éloignement est de 2) ;
- puis, on établit la *liste des caractères correspondants* c'est-à-dire les caractères qu'on retrouve dans les deux chaînes avec un éloignement inférieur ou égal à la valeur calculée précédemment (si on compare **D**W**A**Y**NE** et **D**U**ANE**, les caractères correspondants sont D, A, N et E) ;
- ensuite il faut calculer le *nombre de transpositions entre les caractères correspondants*, c'est-à-dire le nombre de fois (divisé par deux) où le $i^{ème}$ caractère correspondant de la première chaîne est différent du $i^{ème}$ caractère correspondant de la seconde chaîne (dans l'exemple qui précède, le nombre de transpositions est de 0) ;
- enfin, on calcule **la similarité de Jaro**, somme pondérée des trois termes suivants :
  - nombre de caractères correspondants divisé par la longueur de la première chaîne de caractères (soit 4/6 dans notre exemple) ;
  - nombre de caractères correspondants divisé par la longueur de la seconde chaîne de caractères (soit 4/5 dans notre exemple) ;
  - nombre de caractères correspondants moins nombre de transpositions, divisé par le nombre de caractères correspondants (soit 1 dans notre exemple). La similarité de Jaro vaut donc ici 1/3 x 4/6 + 1/3 x 4/5 + 1/3 x 1 = 0,822.

Il existe de nombreuses autres mesures de la similarité entre noms et prénoms. Par exemple, certaines sont basées sur la comparaison de *bigrammes ou trigrammes*[21] entre chaînes de caractères, comme la similarité de Jaccard. Cependant il semble qu'il n'en existe pas qui donne des résultats nettement meilleurs que la similarité de Jaro-Winkler, sur les noms et prénoms (Christen, 2006).

Ensuite, une similarité spécifique à InserJeunes a été élaborée pour les dates de naissances. Par exemple, lorsque deux dates de naissance ne diffèrent que sur le jour de naissance, la similarité est de 0,9 si la différence ne porte que sur un des deux chiffres et de 0,8 sinon.

21. Par exemple les bigrammes de DWAYNE sont DW, WA, AY, YN et NE et les trigrammes de DUANE sont DUA, UAN et ANE.

Si les deux dates de naissance ont la même année et que le jour d'une date correspond au mois de l'autre et réciproquement (exemple : **01**02**02**2005 et **02**01**01**2005) alors la similarité est de 0,65.

Enfin, pour la variable sexe, une similarité binaire est utilisée. Pour la commune de naissance, la similarité est de 1 lorsque les codes COG sont identiques. S'ils sont différents, la similarité est de 0,5 si le code département est identique et de 0 sinon.

CLASSIFIER LES PAIRES: UNE AFFAIRE DE *MACHINE LEARNING?*

La quatrième étape consiste à statuer sur chaque paire, en utilisant les similarités calculées à l'étape précédente. Lorsque les similarités sont élevées, c'est-à-dire proches de 1, la paire est acceptée. Dans le cas contraire, la paire est rejetée.

Encadré 1. Introduction aux algorithmes phonétiques

Un algorithme phonétique est un algorithme conçu pour indexer les mots selon leur prononciation. Par exemple l'algorithme **Soundex** procède comme suit :

1. Conserver la première lettre de la chaîne.

2. Supprimer toutes les occurrences des lettres : a, e, h, i, o, u, w, y (à moins que ce ne soit la première lettre du nom).

3. Attribuer une valeur numérique aux lettres restantes de la manière suivante (dans la version pour les noms en anglais) :

- 1 = B, F, P, V
- 2 = C, G, J, K, Q, S, X, Z
- 3 = D, T
- 4 = L
- 5 = M, N
- 6 = R

4. Si deux lettres (ou plus) avec le même nombre sont adjacentes dans le nom d'origine, ou s'il n'y a qu'un h ou un w entre elles, alors on ne retient que la première de ces lettres.

5. Renvoyer les 4 premiers éléments. S'il y a moins de 4 éléments compléter par des zéros.

Voici des exemples d'application d'algorithmes phonétiques :

Nom de départ	Nom phonétisé			
	Soundex	Phonex	NYSIIS	Double Metaphone
christina	c623	c623	chra	krst
kristina	k623	c623	cras	krst
peter	p360	b360	pata	ptr
pedro	p360	b360	padr	ptr

(Sources : https://fr.wikipedia.org/wiki/Algorithme_phon%C3%A9tique et https://fr.wikipedia.org/wiki/Soundex).

Une première approche simple consiste à calculer une similarité globale pour chaque paire, fonction strictement croissante des similarités des différents champs. Les paires dont la similarité globale est supérieure à un certain seuil sont acceptées, les autres étant rejetées. La fonction et le seuil retenus sont choisis de manière empirique *via* l'analyse d'un échantillon de paires dont le statut (accepté ou rejeté) a été annoté manuellement. Cette méthode présente l'avantage de la simplicité mais le choix de la fonction et du seuil demeurent arbitraires, donc rien ne garantit que ces choix soient optimaux.

Compte tenu des limites de la première approche, des classifications mobilisant des **algorithmes de *machine learning* supervisés** ont été testées (forêts aléatoires et machines à vecteurs de support ou séparateurs à vaste marge (SVM) ***(encadré 3)***). La démarche consiste à entraîner l'algorithme sur un échantillon de paires dont le statut a été renseigné manuellement, puis à l'appliquer sur les autres paires. Les paramètres optimaux de chaque algorithme sont déterminés par validation croisée en maximisant la **métrique *f-measure***.

Dans le cas de l'appariement qualité, l'approche simple, les forêts aléatoires et les SVM ont tous donné des résultats similaires et excellents donc, au final, la classification artisanale a été retenue pour InserJeunes.

Comment expliquer ce résultat, *a priori* surprenant ? Une façon de représenter notre problème est de considérer que chaque paire est un point dans un espace à 5 dimensions, celles des 5 similarités (nom, prénom, date de naissance, commune de naissance et sexe). La variable sexe étant très peu discriminante, elle pourrait être éliminée de l'analyse ce qui restreindrait l'espace à 4 dimensions. Dans chaque dimension, les similarités prennent des valeurs entre 0 et 1.

Le problème consiste donc à trouver une frontière de séparation entre les points/paires acceptés et les points/paires rejetés dans un espace $[0 ; 1]^4$, soit un espace de toute petite taille. De plus, la zone « proche » du point (1,1,1,1) correspond à la zone dans laquelle se trouvent presque toutes les paires qu'on doit accepter. Ainsi, les points correspondant aux paires qu'il faut accepter ne sont pas trop « mélangés » avec les points correspondant aux paires qu'il faut rejeter. Il est donc relativement facile de résoudre ce type de problème, ce qui explique que toutes les méthodes testées donnent de manière équivalente de très bons résultats.

La classification probabiliste, développée originellement par (Fellegi et Sunter, 1969) est une autre méthode développée spécifiquement dans le cadre des appariements sur identifiants indirects et fréquemment citée dans la littérature. Cette méthode n'a pas été investiguée par l'équipe InserJeunes, car il n'en existe pas, à notre connaissance, d'implémentation rapide dès que le volume de données est assez important.

ÉVALUER POUR VALIDER LES CHOIX OPÉRÉS

Étant donné le caractère central des appariements sur identifiants indirects dans InserJeunes, il convenait d'évaluer leur qualité. C'est donc la cinquième et dernière étape du processus.

L'évaluation nécessite de disposer d'un échantillon de paires dont le statut (accepté ou rejeté) a été annoté manuellement et qui n'a pas été utilisé lors de l'étape de classification. Sur cet échantillon, la prédiction issue de la classification supervisée est comparée avec le véritable statut de la paire, c'est-à-dire celui établi manuellement, ce qui permet d'obtenir dans un premier temps quatre grandeurs :

- les vrais positifs (VP) ;
- les faux positifs (FP) ;
- les vrais négatifs (VN) ;
- et les faux négatifs (FN).

Par exemple, une paire faux négatif est une paire rejetée par l'algorithme de classification mais acceptée par l'humain qui a réalisé l'annotation. À partir de ces quatre grandeurs il est possible d'établir plusieurs mesures de la qualité globale.

La mesure la plus connue est l'*accuracy*, soit (VP+VN)/ nombre total de paires. Mais comme toute mesure qui utilise les vrais négatifs, elle n'est pas adaptée. Pourquoi ? Parce que les données sont déséquilibrées : il y a beaucoup de paires dont le vrai statut est rejeté et peu de paires dont le vrai statut est accepté. Dans le cas de l'appariement qualité, environ 40 000 paires sont acceptées sur 1 million de paires donc au minimum 950 000 paires ont pour véritable statut « rejeté ». Un classifieur naïf qui rejette 100 % des paires a donc une *accuracy* d'au moins (0+950 000)/ (1 000 000) soit 95 %.

Encadré 2. Les démarches juridiques

Les appariements sur identifiants indirects se font souvent sur des données à caractère personnel (DCP). Or leur usage est encadré depuis 2018 par le règlement général sur la protection des données (RGPD)*.

Le dispositif InserJeunes a donc fait l'objet d'une déclaration au registre des traitements suivi par le délégué à la protection des données du ministère de l'Éducation nationale, conformément à l'article 30 du règlement. Une analyse d'impact relative à la protection des données (AIPD) a également été réalisée. La réalisation d'une AIPD est obligatoire d'une part pour certains types de traitements (cette liste ayant été établie par la Cnil**) et d'autre part lorsqu'au moins deux critères parmi une liste de neuf s'appliquent au traitement.

InserJeunes remplit les trois critères suivants :

- collecte de données personnelles à large échelle ;
- croisement de données ;
- et personnes vulnérables (patients, personnes âgées, enfants, etc.).

Schématiquement, une AIPD comporte trois parties :

- une description du traitement mis en œuvre ;
- l'évaluation de la nécessité et de la proportionnalité de collecte de DCP ;
- une analyse des risques de sécurité ainsi que leur impact potentiel sur la vie privée ;

Le RGPD impose de limiter au strict nécessaire, compte tenu des finalités du traitement, la collecte et la conservation de DCP.

L'équipe InserJeunes a également fait auprès du Cnis des demandes d'accès aux sources de la Direction générale de l'enseignement et de la recherche*** et de la Dares utilisées dans le dispositif, au titre de l'article 7bis de la loi de 1951*.

** Voir références juridiques en fin d'article.*
*** Voir https://www.cnil.fr/sites/default/files/atoms/files/liste-traitements-aipd-requise.pdf.*
**** Direction du ministère de l'Agriculture.*

Trois autres mesures ont donc été retenues dans le dispositif InserJeunes :

- la **précision**, qui vaut VP/(VP+FP). Par exemple, si la précision est de 80 % alors cela veut dire que 80 % des paires acceptées le sont à bon escient ;
- le **rappel**, qui vaut VP/(VP+FN) : si le rappel est de 90 % alors cela veut dire que 90 % des vraies paires ont été détectées par l'algorithme de classification ;
- et la ***f-measure*** qui est la moyenne harmonique de la précision et du rappel : elle vaut donc 2 x (précision x rappel)/(précision+ rappel).

Dans le cas de l'appariement qualité, la précision vaut 95 % et le rappel 99 %. Au total, 97 % des apprentis sont appariés dans l'appariement qualité (84 % *via* l'appariement direct et 13 % *via* l'appariement approché) soit un taux d'appariement proche du taux théorique de 100 %.

LA MISE EN ŒUVRE INFORMATIQUE : LE CHOIX D'UN OUTIL SPÉCIFIQUE

Plusieurs logiciels d'appariement ont été testés dans le cadre du projet InserJeunes : *FEBRL* de Peter Christen, *matchID* développé au ministère de l'Intérieur[22] et deux librairies R. En R, la meilleure implémentation semble être la librairie R *fastlink*, mais d'après la documentation[23] elle met environ 8 heures pour traiter l'appariement de tables de 300 000 lignes et elle s'appuie sur la classification probabiliste de Fellegi et Sunter. Seul l'outil *matchID* répondait aux besoins mais sa mise en œuvre s'est avérée relativement complexe.

L'équipe InserJeunes a également étudié la documentation concernant d'autres logiciels. L'institut national de statistiques italien Istat a développé un outil nommé RELAIS[24] qui implémente notamment la classification probabiliste ; mais à partir de 100 000 observations, le temps de traitement est d'environ 1h15 (Eurostat, 2009). Aux États-Unis, le bureau du Census a de son côté conçu l'outil *bigMatch* spécifiquement pour traiter de gros volumes, mais il semble ne réaliser que la phase d'indexation des données et il est écrit en C, ce qui complique son intégration avec des outils de *machine learning*[25].

Suite à ce travail de comparaison, il a été décidé de développer un outil spécifique qui réponde à quatre grands besoins d'InserJeunes :

- les appariements doivent être **rapides** : l'outil InserJeunes réalise l'appariement qualité en 15 minutes ;
- l'outil d'appariement doit être **générique** c'est-à-dire facilement adaptable pour tous les cas d'appariements sur identifiants indirects. Pour ce faire, la spécification de chaque appariement (les champs comparés, la méthode de similarité choisie pour chaque champ, etc.) est décrite dans le langage de balisage XML qui est ensuite interprété par l'outil. Cela implique que le statisticien ou le *data scientist* qui produit le XML respecte une grammaire formelle : il doit décrire son appariement en respectant un formalisme et un vocabulaire spécifiques à l'outil[26] mais qui est très fortement inspiré par celui présenté dans l'ouvrage de Peter Christen (Christen, 2012). Cette façon de procéder permet également d'assurer une traçabilité complète de chaque appariement, les spécifications XML étant toutes sauvegardées ;

22. Dans le cadre d'un projet du programme Entrepreneurs d'Intérêt Général.
23. Voir (Enamorado, Fifield et Imai, 2019) page 362 figure 3 *Running Time Comparison*.
24. Pour plus de détail, voir (Istat, 2020).
25. L'équipe projet n'a pas connaissance de librairie de *machine learning* écrite en C, et faire cohabiter des briques écrites dans des langages différents demande un investissement plus conséquent.
26. Un langage spécifique a ainsi été créé à cette occasion pour le domaine appariement.

Encadré 3. Brève introduction aux classifications supervisées en *machine learning*

Les algorithmes de classification supervisée de *machine learning* sont, dans un premier temps, entraînés sur des données étiquetées c'est-à-dire pour lesquelles la variable à prédire est connue (dans le cas d'InserJeunes une variable qualitative binaire). On retient le paramétrage général de l'algorithme qui maximise une grandeur statistique à déterminer et qui dépend du problème traité.

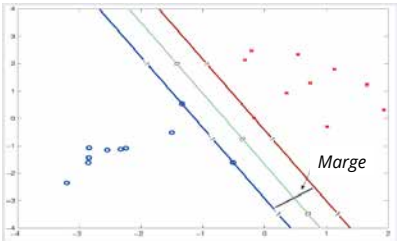


Figure A

En l'occurrence, c'est la métrique *f-measure* qui est maximisée. Ensuite, le modèle précédemment entraîné est appliqué sur de nouvelles données pour lesquelles la variable à prédire est inconnue. L'enjeu méthodologique principal consiste à s'assurer que l'algorithme a « bien appris » pendant l'entraînement afin qu'il fasse ensuite des prédictions correctes sur les nouvelles données.

Le processus de classification supervisée est illustré avec l'algorithme *support vector machine* (SVM) appliqué sur un exemple simplifié, dans lequel il n'y a que deux dimensions ***(figure A)***.

La première approche consiste à choisir la frontière (les lignes bleues et rouge) dont la marge est la plus large possible, c'est-à-dire qu'on veut que tous les points d'une couleur soit d'un côté de la frontière et les points de l'autre couleur soient de l'autre côté et on veut également maximiser la taille du *no man's land* entre les deux lignes c'est-à-dire la zone dans laquelle il n'y a aucun point. Cette approche s'appelle aussi « séparateur à vaste marge ».

Mais pour certains jeux de données ce type de frontière n'existe pas. De plus, si on impose à l'algorithme de séparer 100 % des points, alors ce dernier « collera » trop aux données d'apprentissage. Il risque alors de « sur-apprendre » et de mal généraliser sur de nouvelles données ***(figure B).***

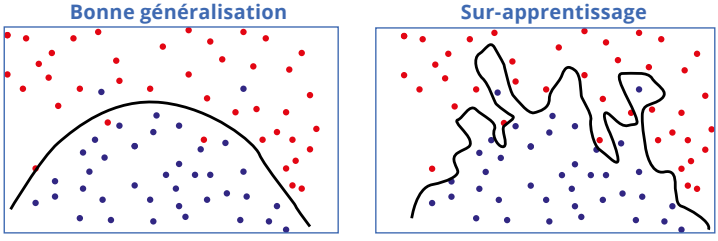


Figure B

Pour éviter cet écueil il faut accepter que le SVM ne classe pas correctement un petit pourcentage des paires. Il faut également évaluer la qualité de l'algorithme sur des données étiquetées qui n'ont pas été utilisées lors de la phase d'apprentissage, ce qui permet de vérifier qu'il n'y a pas eu de sur-apprentissage.

- étant donné que l'évaluation du processus d'appariement est fondamentale et que cela nécessite de disposer d'un échantillon de paires annotées manuellement, une interface d'annotation de paires **ergonomique** a été développée ;
- l'outil s'appuie sur plusieurs librairies *open source* ce qui a permis d'accélérer son développement (le cœur de l'outil a été développé en deux semaines[27]) et d'en faciliter la maintenance.

L'outil d'appariement d'InserJeunes sera mis à disposition en *open source* à l'été 2021. Cependant, pour des projets d'appariement sur des tables nettement plus volumineuses, l'outil *matchID* semble plus adapté notamment parce qu'il réalise l'indexation *via* des requêtes *elastic search* et non pas *via* des requêtes SQL.

PARTAGER L'EXPÉRIENCE ACQUISE AU COURS DU PROJET

À la lumière de l'expérience acquise au cours du projet InserJeunes, quels enseignements tirer ?

En premier lieu, il est manifeste que la qualité globale du processus d'appariement dépend très fortement de la qualité des variables identifiantes indirectes. Le contexte d'InserJeunes était favorable, car les variables sont bien renseignées dans les bases mobilisées. Le respect de chacune des étapes se révèle effectivement indispensable à la réussite de l'opération : bien normaliser les données afin de faciliter les traitements ultérieurs, consacrer du temps à déterminer la meilleure façon de calculer les similarités (ce travail s'appelle le « *feature engineering* » en *machine learning*) afin d'augmenter sensiblement la qualité de la classification, aucune de ces activités n'est superflue. L'évaluation, quand elle s'appuie sur un échantillon de paires annotées manuellement n'ayant pas été utilisées dans l'étape de classification, permet de garantir la qualité globale du processus et en particulier de vérifier qu'il n'y a pas eu de sur-apprentissage. Pour InserJeunes, pouvoir réaliser un « appariement qualité » annuel est une chance, mais tous les systèmes d'informations ne s'y prêteront pas.

Du point de vue informatique le fait de s'appuyer sur plusieurs librairies *open source* a permis de réaliser les développements dans des délais courts. Le choix de décrire chaque spécification d'appariement en XML, respectant un langage spécifique d'appariement a permis de spécifier puis d'intégrer rapidement dans la chaîne de production tous les appariements nécessaires à InserJeunes. Globalement, ces choix ont permis d'achever la mise en œuvre du cœur d'InserJeunes fin 2020, puis de diffuser début février 2021 les premiers résultats sur les jeunes sortants de voie professionnelle scolaire et par apprentissage à l'été 2018 et 2019[28].

27. L'indexation est réalisée en langage SQL sur une base de données PostgreSQL, en mobilisant le module fuzzystrmatch, le calcul des similarités de Jaro-Winkler et de Levenshtein mobilise la librairie Python jellyfish et les algorithmes de *machine learning* sont réalisés avec la librairie Python scikit-learn.
28. Voir (Collin et Marchal, 2021a ; 2021b ; 2021c).

BIBLIOGRAPHIE

CHRISTEN, Peter, 2006. *A Comparison of Personal Name Matching: Techniques and Practical Issues.* [en ligne]. Septembre 2006. The Australian National University Research Publications. Joint Computer Science Technical Report Series, TR-CS-06-02. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://openresearch-repository.anu.edu.au/bitstream/1885/44521/3/TR-CS-06-02.pdf*.

CHRISTEN, Peter, 2012. *Data matching. Concepts and techniques for record linkage, entity resolution and duplicate detection.* 4 juillet 2012. Springer. ISBN 978-3-642-31163-5.

COLLIN, Christel et MARCHAL, Nathalie, 2021a. *Six mois après leur sortie en 2019 du système éducatif, 41 % des lycéens professionnels sont en emploi salarié.* [en ligne]. Février 2021. DEPP-MENJS. Note d'information n°21.06. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://education.gouv.fr/six-mois-apres-leur-sortie-en-2019-du-systeme-educatif-41-des-lyceens-professionnels-sont-en-emploi-309320*.

COLLIN, Christel et MARCHAL, Nathalie, 2021b. *Six mois après leur sortie en 2019 du système éducatif, 62 % des apprentis de niveau CAP à BTS sont en emploi salarié.* [en ligne]. Février 2021. DEPP-MENJS. Note d'information n°21.07. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://education.gouv.fr/six-mois-apres-leur-sortie-en-2019-du-systeme-educatif-62-des-apprentis-de-niveau-cap-bts-sont-en-309329*.

COLLIN, Christel et MARCHAL, Nathalie, 2021c. *Des lycéens professionnels et des apprentis mieux insérés 12 mois après leur sortie d'études en juillet 2020 que 6 mois après, malgré la crise.* [en ligne]. Mai 2021. DEPP-MENJS. Note d'information n°21.24. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://education.gouv.fr/des-lyceens-professionnels-et-des-apprentis-mieux-inseres-12-mois-apres-leur-sortie-d-etudes-en-323294*.

ENAMORADO, Ted, FIFIELD, Benjamin et IMAI, Kosuke, 2019. Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records. In : *American Political Science Review*. [en ligne]. N° 113, 2, pp. 353-371. [Consulté le 27 mai 2021]. Disponible à l'adresse : *http://imai.fas.harvard.edu/research/files/linkage.pdf*.

EUROSTAT, 2009. *Insights on Data Integration Methodologies*. [en ligne]. ESSnet-ISAD workshop, Vienne, 29-30 mai 2008, page 53. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://www.istat.it/it/files/2015/04/Insights-on-Data-Integration-Methodologies.pdf.*

EVAIN, Franck, 2020. Indicateurs de valeur ajoutée des lycées. Du pilotage interne à la diffusion grand public. In : *Courrier des statistiques*. [en ligne]. 31 décembre 2020. Insee. N° N5, pp. 74-94. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5008703/courstat-5-6.pdf*.

FELLEGI, Ivan P. et SUNTER, Alan B., 1969. A theory for record linkage. In : *Journal of the American Statistical Association.* Décembre 1969. Taylor & Francis Ltd.. Volume 64, n°328, pp. 1183-1210.

HUMBERT-BOTTIN, Élisabeth, 2018. La déclaration sociale nominative. Nouvelle référence pour les échanges de données sociales des entreprises vers les administrations. In : *Courrier des statistiques.* [en ligne]. 6 décembre 2018. Insee. N° N1, pp. 25-34. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647025/courstat-1-6.pdf*.

ISTAT, 2020. RELAIS (Record Linkage At Istat). In : *site de Istat.* [en ligne]. 19 novembre 2020. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://www.istat.it/en/methods-and-tools/methods-and-it-tools/process/processing-tools/relais*.

JABOT, Patrick et TREYENS, Pierre-Eric, 2018. Appariement de l'enquête Care par identification du plus proche écho. In : *site des 13es Journées de méthodologie statistique de l'Insee (JMS).* [en ligne]. 12-14 juin 2018. [Consulté le 27 mai 2021]. Disponible à l'adresse : *http://www.jms-insee.fr/2018/S20_1_ACTEv2_TREYENS_JMS2018.pdf*.

JAMES, Gareth, WITTEN, Daniela, HASTIE, Trevor et TIBSHIRANI, Robert, 2013. *An introduction to statistical learning with applications in R*. Springer. ISBN 978-1-4614-7138-7.

KILSS, Beth et ALVEY, Wendy, 1985. *Record Linkage Techniques – 1985.* [en ligne]. 1er décembre 1985. Workshop on Exact Matching Methodologies, Arlington, Virginia, May 9-10, 1985. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://rosap.ntl.bts.gov/view/dot/13855*.

FONDEMENTS JURIDIQUES

Loi n° 2018-771 du 5 septembre 2018 pour la liberté de choisir son avenir professionnel. In : *site de Légifrance*. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/dossierlegislatif/JORFDOLE000036847202/*.

Loi n° 51-711 du 7 juin 1951 sur l'Obligation, la coordination et le secret en matière de statistiques. In : *site de Légifrance*. [en ligne]. Mise à jour du 25 mars 2019. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000888573/*.

Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données. In : *site EUR-Lex.* [en ligne]. [Consulté le 27 mai 2021]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:32016R0679&from=FR*.