

ANNEXE 5
AU CAHIER DES CLAUSES TECHNIQUES
PARTICULIERES (CCTP)

Réalisation des prestations de tierce maintenance applicative des applications "Répertoire national des identifiants élèves, étudiants et apprentis" (RNIE) et « Base nationale des identifiants élèves » (BNIE) pour le compte du ministère chargé de l'éducation nationale, de l'enseignement supérieur et de la recherche.

L'algorithme d'immatriculation de la BNIE et du RNIE

L'algorithme d'immatriculation de la BNIE et du RNIE

Le service d'immatriculation des élèves du système éducatif français repose en grande partie sur des traitements automatisés implémentés au sein de la BNIE et du RNIE. C'est notamment le cas pour le traitement des demandes de création d'INE, d'admission ou de ré immatriculation¹ qui est fondé sur un algorithme d'immatriculation développé par la DEPP au milieu des années 2000. Les autres types de demandes ne font pas appel à cet algorithme puisque l'INE de l'élève est déjà connu.

L'algorithme d'immatriculation intégré à la BNIE traite les demandes reçues par **ONDE**, le système de gestion de scolarité du premier degré. Ce même algorithme est également intégré dans le RNIE et traite les demandes reçues par SYSCA² et INES³ qui sont respectivement le système de gestion de scolarité du second degré et l'interface qui permet, pour l'enseignement supérieur, de rechercher un INE et de demander l'immatriculation d'un nouvel étudiant au RNIE.

1. Fonctionnement de l'algorithme

Dans son fonctionnement, l'algorithme d'immatriculation utilise les cinq critères d'état civil que sont le nom, prénom(s), sexe, date et lieu de naissance. ONDE, SYSCA ou INES transmettent donc pour chaque élève recherché toutes ces informations afin qu'elles soient comparées aux mêmes informations stockées dans la BNIE ou le RNIE.

Par la suite, le processus d'immatriculation se fait via les étapes de normalisation des données, de calcul de clés d'appariement, d'extraction de ressemblants, de notation et enfin de décision.

1.1 Normalisation des données

Dans un premier temps, les données d'état civil des élèves sont normalisées afin de pallier tout problème lié à la qualité des données. Cela passe notamment par :

- la suppression des accents,
- le traitement des caractères spéciaux,
- le traitement des particules,
- le traitement des espaces,
- la transcription en majuscule.

La date de naissance est également normalisée afin de définir la valeur 00 lorsque les jours et mois de naissance sont inconnus.

1.2 Calcul des clés d'appariement

Les clés d'appariement sont des encodages numériques de certaines variables parmi lesquels on a :

- noms (nom de famille, nom d'usage),

¹ Type de demande où il est nécessaire de donner un nouvel INE à un élève, malgré le fait qu'il en possède déjà un. Cela fait généralement suite à un traitement de doublons, où deux élèves différents se partagent le même INE.

² Système d'information statistique consolidé académique

³ Identifiant National dans l'Enseignement Supérieur

- prénom(s),
- jour de naissance,
- mois de naissance,
- année de naissance,
- lieu de naissance (département, commune, pays),
- sexe.

On appelle **clés de recherche** les clés calculées à partir des données des élèves recherchés, et **clés d'accès** celles calculées pour les élèves présents en base. On compte également deux types de recherches :

- **une recherche exacte**, comparant les données normalisées entre elles,
- **une recherche approchée**, prenant compte des variantes orthographiques ou phonétiques. Cette étape repose sur des clés phonétisées et vise à corriger les erreurs de saisie ou les différences d'usage (Ferrera/Ferreira par exemple).

NB : toute recherche passe impérativement par une recherche exacte, suivie d'une recherche approchée si la première n'aboutit pas.

Des dictionnaires sont utilisés par les algorithmes au moment de la phonétisation :

- dictionnaire des caractères spéciaux,
- dictionnaire des mots vides,
- dictionnaire des particules.

1.3 Extraction et sélection

Ces étapes consistent en l'extraction, pour chaque élève recherché, de l'ensemble des élèves présents en base partageant les mêmes clés que lui. En fonction du volume, tous les élèves extraits peuvent ne pas être sélectionnés. Si leur nombre est important, seuls ceux extraits par au moins 2 clés peuvent être sélectionnés. Les autres sont rejetés.

1.4 Notation et décision

La notation est régie par un système de pondération, attribuant un poids plus important au nom (usage, famille) qu'aux prénoms et aux autres informations. A chaque variable utilisée pour la notation est attribuée une note. La note globale de ressemblance est une moyenne pondérée de ces notes.

Deux seuils sont nécessaires :

- seuil de ressemblance, en deçà duquel l'élève extrait n'est pas sélectionné,
- seuil d'identification, à partir duquel l'élève recherché a été identifié (mais il peut y avoir des doublons, notamment si deux élèves sont présents au-delà du seuil d'identification).

L'algorithme attribue un INE à partir de la note globale sur 900, selon les règles suivantes :

- **note de 900** : l'INE existant est attribué automatiquement,
- **note < 800** : un nouvel INE est créé automatiquement,
- **note entre 800 et 900** : le cas est mis en litige.

Pour les cas en litige :

- si la note est inférieure à 865, la décision revient au gestionnaire,
- si la note est supérieure ou égale à 865, l'algorithme considère qu'il s'agit très probablement du même élève, mais laisse la validation finale au gestionnaire.

2. Evolutions à venir

L'algorithme d'immatriculation fait actuellement l'objet d'une refonte visant à améliorer la qualité de l'appariement et à réduire le nombre de litiges créés. Ce nouvel algorithme suit les mêmes étapes que celui actuellement en place dans la BNIE et le RNIE. Les améliorations portent sur la précision de la comparaison, la recherche des cas d'inversion des noms et prénoms ou encore l'« assouplissement » des critères de décision.