JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

ALAIN GOY

L'appariement sécurisé de fichiers d'étudiants grâce au « hachage » des identifiants

Journal de la société française de statistique, tome 146, n° 3 (2005), p. 39-46

<http://www.numdam.org/item?id=JSFS_2005__146_3_39_0>

L'APPARIEMENT SÉCURISÉ DE FICHIERS D'ÉTUDIANTS GRÂCE AU «HACHAGE» DES IDENTIFIANTS

Alain GOY *

RÉSUMÉ

L'existence d'un identifiant dans le fichier SISE (Système d'Information sur le Suivi de l'Étudiant) des étudiants universitaires semblait ouvrir de belles perspectives pour l'étude des cursus d'étudiants, aussi bien par les services d'étude ou de recherche que par les universités elles-mêmes. Mais la CNIL [1] a considéré que cet identifiant donnait un caractère nominatif à ce fichier, et s'est opposé à sa diffusion tel quel. La technique de « hachage » des identifiants permet d'assurer que l'on ne peut pas remonter à l'identité de l'étudiant, tout en permettant de le suivre d'une année sur l'autre. L'article décrit comment elle a été mise en œuvre, ce qui a permis d'obtenir un accord de la CNIL. Cette première expérience ouvre la voie à une généralisation de l'utilisation des fichiers individuels d'élèves et d'étudiants pour la recherche et les études.

Mots clés : appariement, cursus, hachage, identifiant, trajectoire.

ABSTRACT

The inclusion of an identifier in the French database on university students (Système d'Information sur le Suivi de l'Étudiant : SISE) seemed to offer bright prospects for the analysis of students' academic records by study and research units as well as by universities themselves. But the French National Commission on Information Technology and Civil Liberties (CNIL) ruled that the identifier turned SISE into a nominal database; as a result, CNIL blocked the dissemination of the data in their original form. By "hashing" the identifiers, one can prevent students' identities from being traced while enabling their careers to be monitored year to year. The article describes how the hashing technique was implemented, resulting in CNIL approval. This initial experience paves the way for a widespread use of individual data files on school pupils and university students for research and analysis purposes.

Keywords : academic records, hashing, identifier, merging.

* Au moment de la rédaction de cet article, Direction de l'Évaluation et de la Prospective, adjoint à la directrice pour les statistiques, Ministère de l'Éducation nationale, de l'enseignement supérieur et de la recherche. Avec la collaboration de Claire Teissier et Pauline Girardot, sous-direction des études statistiques de la DEP.
Adresse actuelle : INSEE, 18 Bd Adolphe Pinard 75675 Paris cedex 14.
Courriel : alain.goy@insee.fr

1. Commission nationale de l'informatique et des libertés.

1. Le contexte de ce travail

La direction de l'Évaluation et de la prospective du ministère de l'Éducation nationale, de l'enseignement supérieur et de la recherche collecte et élabore de l'information pour accomplir ses deux missions : éclairer le débat public sur l'école et fournir des indicateurs pour le pilotage du système éducatif. La fonction dite d' « aide à la décision » nécessite une information très fine car chaque échelon de décision (administration centrale, rectorats, chefs d'établissement, inspecteurs) doit disposer de ses propres indicateurs de pilotage et les contrats d'objectifs passés entre ces différents échelons s'appuient aussi sur des indicateurs.

La source principale de cette information détaillée réside dans les systèmes de gestion, qu'il faudrait plus justement nommer systèmes de gestion et de pilotage. En effet les statisticiens sont intervenus en amont, lors du développement de ces systèmes, pour que les bonnes variables d'analyse y figurent ; les statisticiens gèrent également toutes les nomenclatures de ces systèmes de gestion.

Actuellement, chaque année, sont collectés des fichiers individuels sur 5,5 millions d'élèves du secondaire et 1,4 millions d'étudiants universitaires. Il reste néanmoins toute une partie du système éducatif pour laquelle l'information est encore collectée sous forme de tableaux par établissement ou école.

Des panels classiques aux suivis d'élèves dans les fichiers administratifs

Très tôt, les panels d'élèves sont apparus comme un outil indispensable pour analyser les cursus scolaires et ce qui les influence ou les détermine. Ils sont actuellement de deux types : les panels d'étudiants procèdent par interrogation annuelle directe des étudiants (questionnaire postal et relance téléphonique) ; pour les élèves du secondaire et du primaire, en revanche, on collecte un maximum d'information par voie administrative (soit par recherche dans les fichiers d'élèves – pour le secondaire – soit par interrogation des chefs d'établissement ou des directeurs d'école). De plus, des enquêtes sont faites directement auprès des familles. L'information a donc une origine mixte (administrative ou par enquête).

Mais il est aussi envisageable d'utiliser des sources purement administratives pour établir des cursus d'élèves ou étudiants dès lors qu'elles contiennent un identifiant permettant d'apparier les fichiers de plusieurs années. C'est le cas des fichiers individuels d'élèves du secondaire ou d'étudiants universitaires, qui contiennent un identifiant INE (Identifiant national élève ou étudiant). Certes, on n'aura pas la richesse des panels évoqués ci-dessus mais on aura une information plus exhaustive et donc exploitable à un niveau géographique ou fonctionnel fin. L'INE est un numéro attribué depuis 1995 aux élèves dans l'enseignement secondaire, et ils le conservent tout au long de leurs études. Dans le cas d'étudiants n'ayant pas effectué leur cursus secondaire en France, ou lorsqu'un étudiant a perdu son numéro INE, les universités disposent d'un algorithme (fourni par la DEP) permettant d'immatriculer les étudiants. On

verra ci-dessous qu'un projet est en cours pour généraliser l'immatriculation des élèves dès l'école primaire et utiliser ce numéro dans tout le système éducatif.

Mais il existe des freins à l'utilisation de fichiers nationaux comportant l'INE

La CNIL considère que la constitution de fichiers centraux avec un identifiant tel que l'INE présente un risque. Elle estime en effet qu'il est actuellement facile de se procurer l'INE d'un élève particulier, et que la multiplication de fichiers d'études nationaux comportant l'INE pourrait permettre trop aisément à des personnes mal intentionnées d'accéder à des informations individuelles nominatives. Elle juge que ce risque ne peut être couru que si, moyennant bien sûr toutes les précautions nécessaires, c'est indispensable à l'accomplissement des missions d'un service. Ainsi les autorités académiques, qui ont en charge directement les élèves du second degré, sont-elles autorisées à constituer des fichiers avec INE, mais pas la DEP au niveau national. Les études de trajectoires scolaires lui sont donc interdites par cette voie. En revanche, l'administration de l'enseignement supérieur relevant du niveau national, la DEP est bien autorisée à constituer des fichiers avec INE. Mais il n'est pas permis de les mettre à disposition des universités ou des chercheurs.

Il fallait trouver une solution pour permettre un suivi anonyme des parcours scolaires et universitaires

La situation semblait bloquée du point de vue des statisticiens et chercheurs. Or, il existait un projet à moyen terme de constituer des fichiers individuels d'élèves et d'étudiants de tout le système éducatif. Ce projet ne pouvait avoir de sens que si l'on trouvait une solution aux problèmes bloquant l'utilisation de l'INE. Cette note s'attache plus particulièrement à décrire l'expérimentation d'une solution sur un champ particulier mais significatif : les fichiers d'étudiants SISE (Système d'Information sur le Suivi de l'Étudiant). Cette expérimentation a aussi ouvert la voie à des réflexions pour mettre en place des solutions plus élaborées et pérennes, qui seront évoquées brièvement.

2. L'expérimentation conduite à partir des fichiers SISE sur les étudiants

La richesse des données SISE

Au ministère de l'ducation nationale, la direction de l'évaluation et de la prospective (DEP), et plus particulièrement le bureau des études statistiques sur l'enseignement supérieur, est chargé de réaliser chaque année un recensement des inscriptions prises dans les universités et établissements d'enseignement supérieur.

Depuis la rentrée universitaire 1994-1995, ce recensement s'effectue, pour ce qui concerne les universités et certains établissements, dans le cadre de l'opération SISE. Le système est alimenté par deux remontées d'informations annuelles, l'une concernant les inscriptions universitaires, la seconde portant sur les résultats aux diplômes.

Les données sont extraites des bases de gestion des services de scolarité des universités. Elles sont transmises par les universités sous la forme d'un fichier informatique dont le contenu et le format ont été définis en commun lors du développement du projet SISE. Les fichiers contiennent un enregistrement par inscription prise par un étudiant (si un étudiant est inscrit à la préparation de plusieurs diplômes, il apparaîtra autant de fois qu'il aura pris d'inscriptions). Chaque enregistrement est composé d'une quarantaine de variables décrivant la scolarité actuelle de l'étudiant (diplôme préparé, régime d'inscription, unité de rattachement...), sa scolarité antérieure (série et année d'obtention du baccalauréat, année d'entrée dans l'enseignement supérieur...), sa situation socio-démographique (identifiant national étudiant, sexe, année de naissance, nationalité, catégorie socioprofessionnelle des parents...).

Les informations contenues dans SISE sont très riches et permettent de traiter de nombreux sujets. À la fin de chaque remontée, la DEP publie rapidement une note d'informations de six pages décrivant essentiellement les grandes évolutions d'effectifs par cycle d'études, par discipline et par université et intègre les données détaillées (mais sans identifiant) dans la « base centrale de pilotage » du ministère.

Les universités veulent utiliser elles-mêmes les données SISE

Les universités ont bien accès aux données agrégées et aux publications, mais les données disponibles dans SISE permettent des analyses beaucoup plus détaillées et plus fines. En effet, les données étant individuelles, tous les croisements entre variables sont possibles. De plus, le système contient la localisation géographique précise de la formation des étudiants. De ce fait, il est possible à partir de SISE de traiter de nombreux sujets d'études intéressants pour les universités. Tous ne peuvent pas être traités par la DEP ; en particulier, les études menées au niveau national peuvent difficilement se décliner par académie et *a fortiori* par université. Certaines universités sont donc demandeuses, depuis plusieurs années, d'un retour du fichier national SISE.

Depuis 2001-2002, les universités reçoivent chaque année un cédérom contenant des données nationales anonymes extraites de SISE au format du logiciel BEYOND, qui permet de faire très facilement des tableaux. Ces cédéroms ont permis aux établissements de comparer leur population d'étudiants à celle des autres établissements, de connaître l'offre de formations des autres établissements, etc. Cependant, cette solution n'est pas entièrement satisfaisante : les universités veulent, par exemple, comprendre pourquoi certains étudiants quittent leur établissement en fin de premier ou deuxième cycle pour s'inscrire ailleurs. Elles souhaitent pouvoir suivre les étudiants qui s'inscrivent dans leur établissement, observer leurs parcours, les migrations inter-universités, afin de

connaître les débouchés de leur formation mais aussi d'adapter éventuellement leur offre de formation.

Toujours un blocage sur la constitution de fichiers nominatifs... mais la CNIL suggère elle-même une solution

Afin d'effectuer un suivi longitudinal de leurs étudiants, les universités ont, à plusieurs reprises, sollicité la DEP pour obtenir un fichier national avec INE. Ce genre d'informations est sensible dans la mesure où il permettrait, par exemple, la fabrication de *curriculum vitae* électroniques des étudiants. Les contacts informels avec la CNIL ont confirmé cette idée. Elle considérait que la présence dans un fichier de l'identifiant national élève-étudiant suffisait à le rendre nominatif, et amenait à interdire sa diffusion large. Mais elle a suggéré que l'on tire parti des travaux déjà effectués, notamment par le CHU [2] de Dijon [3] et par la CNAMTS [4] en matière de cryptage sans retour (hachage) des identifiants.

La DEP a rencontré un contexte très favorable puisqu'il s'est trouvé que la Société française de statistique (SFdS, groupe statistiques économiques et sociales) a organisé en janvier 2003 une formation sur la « pratique des appariements sécurisés », avec la participation du Dr Catherine Quantin du CHU de Dijon et de Mme Vulliet-Tavernier, chef de service de la CNIL. Les nouvelles méthodes de cryptage désormais autorisées, notamment les méthodes de cryptage sans retour (par hash coding), appliquées aux identifiants, permettent en effet de concilier les exigences apparemment contradictoires de l'anonymat et de l'appariement de fichiers. L'identifiant « haché » permet des appariements de fichiers mais ne permet pas de retrouver le nom ni les coordonnées de la personne.

La CNIL a finalement autorisé la diffusion aux universités des fichiers SISE avec INE « haché » (donc sans possibilité de retour arrière), à condition aussi de limiter le nombre de variables diffusées afin d'éviter des identifications indirectes.

L'aspect technique : le « hachage » des identifiants INE

Une convention a été passée avec le CHU de Dijon pour l'utilisation de son logiciel ANONYMAT, qui est lui-même fondé sur l'algorithme SHA (*Standard Hash Algorithm*) considéré comme un des plus sûrs du domaine public. Le choix de cet algorithme pour le développement du logiciel ANONYMAT a été validé par le service central de la sécurité des systèmes d'information (SCSSI), organisme interministériel. L'algorithme transforme de façon irréversible un identifiant de l'individu (dans notre cas l'INE, alors que pour les études d'épidémiologie c'étaient les nom, prénom, sexe et date de naissance des

2. Centre hospitalo-universitaire.

3. Quantin C., Bouzelat H., Allaërt F.A. *et alii.* Automatic record hash coding and linkage for epidemiological follow-up data confidentiality, *Meth Onform Med* 1998 ; 37 :271-7. Voir aussi l'article de Catherine Quantin *et coll.* dans ce même dossier.

4. Caisse nationale d'assurance maladie des travailleurs salariés.

patients). Mais, à lui seul, il ne garantit pas qu'on ne pourra pas retrouver les informations concernant une personne particulière, repérée par exemple par son INE. Ainsi, si quelqu'un a la possibilité de refaire passer l'algorithme sur cet INE avec la même clé, il obtiendra l'INE « haché » et pourra rechercher dans les fichiers statistiques des informations sur cette personne. C'est ce qu'on appelle une « attaque par dictionnaire ». C'est pourquoi le logiciel ANONYMAT prévoit deux clés différentes à deux niveaux : dans les études épidémiologiques, le centre de soins fait un premier « hachage » avec une clé K1 et transmet au service d'études qui fait un second « hachage » avec une clé K2. Dans le cadre de notre expérimentation il n'y a qu'un niveau, celui du service central de statistiques, et, comme nous le verrons ci-dessous, nous avons pris d'autres précautions pour prévenir les attaques par dictionnaire ou toute autre méthode de contournement.

La variable cryptée que donne le logiciel en sortie comporte quarante caractères (l'INE en comporte onze). Le risque de collision (deux numéros INE distincts donnant un même INE crypté) est quasi nul et, de fait, aucune n'a été observée.

L'organisation mise en place pour assurer la confidentialité

Le cryptage avec le meilleur des algorithmes ne peut à lui seul assurer la sécurité et la confidentialité. On a vu que si les clés de « hachage » n'étaient pas tenues secrètes [5] , des « attaques par dictionnaire » étaient possibles. De plus, s'il subsiste sur un disque dur un fichier comportant à la fois un INE et un INE « haché », cela permettra à une personne mal intentionnée de remonter à une information nominative. Et cela est même possible s'il existe côte à côte 2 fichiers triés dans le même ordre, l'un avec l'INE et l'autre avec l'INE « haché ».

C'est pourquoi le « hachage » de l'INE dans les fichiers SISE s'effectue dans le cadre d'une procédure bien précise. Le logiciel est installé sur un ordinateur protégé par un mot de passe connu des deux personnes habilitées à effectuer l'opération. Cet ordinateur n'est pas connecté au réseau. Les clés de cryptage ne sont pas installées sur le disque dur, mais sont conservées dans un coffre. Lors de l'exploitation, la variable « INE » non « hachée » ne se trouve pas dans le fichier en sortie. Ce fichier est immédiatement trié sur la variable « hachée » pour que l'ordre du fichier source ne soit pas conservé. Il est ensuite gravé sur CD-ROM et effacé du disque dur de l'ordinateur.

La dernière précaution consiste à limiter le nombre des variables pour éviter qu'on puisse identifier un individu par recoupements. Ainsi, les variables concernant les communes de résidence des étudiants et celles de leurs parents, ainsi que la nationalité détaillée des étudiants ne seront pas diffusées.

5. Signalons au passage qu'une clé est impossible à deviner car elle comporte plus de mille caractères...

Ce que les universités vont pouvoir faire... et ne pas faire

Les premiers fichiers avec INE « haché » ont été diffusés à l'automne 2004 aux établissements qui en avaient fait la demande (tous avaient préalablement été prévenus de cette possibilité, bien sûr). Il s'agit des fichiers SISE relatifs à cinq années universitaires : 1999-2000 à 2003-2004. Les numéros INE avaient en effet une fiabilité insuffisante avant 1999.

Cinq universités ont pu donc commencer à apparier ces fichiers pour constituer et étudier *statistiquement* les trajectoires d'étudiants : leurs cursus internes mais surtout externes. L'expérience de la DEP est qu'il ne s'agit pas d'une tâche facile : un étudiant peut avoir plusieurs inscriptions par année, et cela complique les appariements. Il peut y avoir des erreurs sur les numéros INE (ou, plus probablement, des ré-immatriculations d'un même étudiant). Bref, la tâche est lourde mais les Observatoires des parcours et de l'insertion professionnelle des étudiants ont la motivation et la connaissance du terrain qui leur permettront de bien exploiter ces données qui, si elles n'avaient pas pu être diffusées aux universités, auraient été gravement sous-utilisées.

Signalons quand même une opération qui sera impossible localement : constituer, à partir de ces fichiers, un fichier de lancement d'enquête auprès des étudiants qui ont interrompu leurs études. Il sera possible en effet de séparer parmi les sortants de l'université ceux qui sont allés dans une autre université de ceux qui sont complètement sortis du fichier SISE, mais on ne pourra pas retrouver leurs nom et adresse, puisque tout a été fait pour qu'on ne puisse pas remonter aux données nominatives ! Patience, des solutions sont néanmoins à l'étude, toujours en concertation avec la CNIL...

3. Les perspectives à moyen terme

À plus long terme, la DEP prévoit d'élargir l'application d'une technique analogue à tout le système éducatif. L'objectif est de pouvoir suivre le cursus des élèves tout au long de leur scolarité, depuis l'enseignement primaire jusqu'à l'enseignement supérieur. C'est une option qui a été prise après mûre réflexion dans le cadre du programme statistique à moyen terme du Conseil National de l'Information Statistique (CNIS), car elle seule peut permettre de répondre à un besoin bien réel. Ce projet permettrait d'améliorer les travaux liés au pilotage du système éducatif (prévision d'effectifs, taux de scolarisation, sorties du système éducatif...) et d'avoir une vision plus claire des parcours de formation initiale d'un point de vue pédagogique et géographique.

Des travaux sont actuellement en cours pour généraliser l'usage de l'INE dans l'ensemble du système éducatif. Une *base nationale des identifiants élèves* permettra d'attribuer des numéros (dès l'école primaire) et de les retrouver en cas de besoin. Parallèlement, se constituent des fichiers d'élèves avec identifiant sur des champs non ou partiellement couverts actuellement : enseignement primaire, apprentissage, enseignement agricole, enseignement supérieur hors universités. Il est envisagé que ces fichiers soient transmis

au service statistique ministériel de l'ducation nationale, après « hachage » et cryptage de l'identifiant. Chaque année, serait ainsi constitué un ensemble de fichiers qui se prêteraient à des appariements, grâce à leur identifiant commun. Ces fichiers, à finalité d'étude et de recherche, seraient conservés sur le long terme et mis à disposition des utilisateurs autorisés. Ce projet porte le nom de FAERE (Fichiers anonymisés d'élèves pour la recherche et les études)

La réflexion sur le cryptage ou le « hachage » des INE s'est poursuivie de façon interactive avec la CNIL dans le cadre d'une « demande de conseil ». Pour la CNIL, l'utilisation de la technique de « hachage », qui interdit tout retour en arrière, était un point clé. Nous avons fait valoir que nous devions mettre en place un système sur le long terme et qu'il faudrait probablement un jour faire face à l'obsolescence de la technique que nous adoptons, quelle qu'elle soit. Pouvoir décrypter les données et les recrypter en utilisant une technique plus robuste nous paraissait une exigence incontournable. Finalement un compromis a été trouvé sur la base de ce que fait l'*Office fédéral de statistiques suisse*. La solution est plus compliquée que celle qui a été retenue pour l'expérience sur SISE. Elle combine « hachage » et cryptage : lors de l'envoi d'un fichier par un rectorat, par exemple, l'INE serait « haché » puis crypté par une clé générée automatiquement. La DEP décrypterait et recrypterait immédiatement ces fichiers en utilisant un mot de passe apporté par trois personnes d'horizons différents (par exemple la DEP, une autre direction du ministère et l'Insee). Si l'algorithme de cryptage devenait caduc, les trois mêmes personnes seraient mobilisées pour décrypter les INE et les recrypter avec une autre technique. Pour assurer la pérennité du système, nous développerions nos propres programmes d'exploitation, mais en utilisant les algorithmes *publics* réputés les plus solides. La spécification complète du système à mettre en place (technique, organisation), dans lequel chaque maillon compte, va faire l'objet d'un volumineux dossier qui devra être approuvé cette fois formellement par la CNIL.

Nous ne sommes pas au bout du chemin mais la progression est bien engagée. La constitution d'un vaste ensemble de données, avec identifiant, portant sur les élèves et les étudiants, débloquera ou facilitera, en interne, nombre d'études très attendues et ouvrira un champ d'investigation important aux chercheurs.