

Description des données exportées

A373 15C9 CC98 5190 59B6 560E 01ED 55A3

1 Données de sortie

Signature MD5 : A373 15C9 CC98 5190 59B6 560E 01ED 55A3 Sortie réalisée 27 avril 2023 suivie de modifications cosmétiques le 28 avril : renommage d'une colonne et type numérique pour l'effectif des cohortes.

2 Données sources

Les données sources sont les données SISE sur les inscriptions et les réussites entre 2010 et 2021. Pour se conformer à la demande que nous avons formulé nous sommes limités aux données concernant les académies de Paris, Creteil et Versailles. Nous n'avons pris en considération que les parcours individuels ayant au moins une inscription entre 2017 et 2021.

3 Objectif

L'objectif du projet Oracclle est de produire des données ouvertes montrant les parcours d'études possibles en Île-de-France dans leur plus grande variété de façon à enrichir le dialogue d'orientation. Les données produites s'attachent préserver cette richesse des parcours tout en préservant l'anonymat des personnes les ayant suivies, quitte à rendre les données inexploitable pour la production d'indicateurs quantitatif sur les flux et les réussites.

4 Fichier formations.csv

Tableau de l'ensemble des formations ayant reçu des inscriptions d'étudiantes d'après les données SISE restreinte à l'Île-de-France. Mis à part le code formation, les champs reprennent la nomenclature SISE.

formation code construit en concaténant l'identifiant de composante (code UAI), l'identifiant de diplôme (code SISE) et, lorsque l'information existe, le degré d'étude (DEGETU des données SISE)

- acaeta** académie de l'établissement tel que fourni dans les données SISE (code à deux chiffres)
- compos** le code UAI de la composante (code établissement ou code d'une composante de l'établissement)
- cursus-lmd** L, M ou D selon le type de cursus
- cycle** 1, 2 ou 3 selon le cycle concerné
- degetu** degré d'étude (le n du bac + n)
- diplom** code SISE du diplôme. Le code par un chiffre entre 1 et 8 inclus. Formation LMD sur 3 caractères 210 (licence), 220 (master ind), 221 (master pro), 222 (master rech), 420 (doctorat), 230 (licence), 240 (licence pro), 250 (master).
- etabli** code UAI de l'établissement si différent du code composante
- groupe** groupe disciplinaire selon la nomenclature SISE
- sectdis** code à deux chiffres représentant le secteur disciplinaire du diplôme (52 valeurs dans la nomenclature SISE)
- typ-dipl** type de diplôme (codé sur deux caractères)
précaution d'anonymisation : aucune.

5 Fichier cohortes.csv

Une cohorte est caractérisée par deux données, d'une part une séquence d'inscriptions principales dans des formations, une par année d'inscription et organisées selon les années croissantes et d'autre part un ensemble de formations réussies, ensemble organisé en séquence croissante selon l'ordre lexicographique. Le nombre de personnes ayant suivi exactement cette séquence d'inscriptions avec exactement cet ensemble de réussites est l'effectif de la cohorte. Si la cohorte avait un effectif inférieur à 5 son effectif n'est pas communiqué. Il est remplacé par un zéro (pour conserver l'homogénéité du type numérique pour la colonne). La séquence d'inscriptions ainsi que la liste des réussites sont conservés.

- cohorteid** identifiant numérique de la cohorte
- effectif** nombre de personnes dans la cohortes, ou zéro si ce nombre est inférieur à 5.
- inscriptions** séquence des inscriptions dans des formations, représentée par la concaténation des codes formation au sens de la première colonne du fichier formations.csv joints par le caractère "+". en cas d'années sans données entre deux inscriptions, le texte "trou" remplace le code formation.
- réussites** ensemble des réussites dans des formations, codifiées par des code formation et organisées par ordre lexicographique sous la forme d'une chaîne utilisant "+" comme séparateur.

6 Attributs des cohortes

Les personnes composant chaque cohorte étaient porteuses d'attributs :

- sexe
- type de bac obtenu
- année d'obtention du bac
- année de début de la séquence d'inscriptions dans la cohorte (première année d'inscription au sens des données SISE en Île-de-France)
- spécialités de terminale au bac général (introduites en 2021)

Ces attributs sont transformés en attributs du groupe entier qui compose la cohorte. Ainsi la corrélation entre eux est détruite. On peut savoir qu'il y a 30 femmes et 40 hommes dans une cohorte parmi lesquels 20 personnes ont obtenues leur bac en 2015, mais on ne sait pas combien il y avait d'hommes et de femmes parmi ces 20 personnes ayant obtenues leurs bac en 2015.

De plus, pour une cohorte donnée et un type d'attribut donné, on a effacé de l'information dès que des valeurs de l'attribut concernaient des effectifs faibles, inférieurs à 5, au sein du groupe. Cet effacement tient compte du risque de recalcul de l'information masquée à partir de l'effectif total de la cohorte. Ce risque de secret secondaire est traité en effaçant de nouvelles valeurs jusqu'à ce qu'il ne soit plus possible d'inférer l'effectif d'une valeur initialement représentée. Ainsi :

- si une cohorte se compose de 70 individus, 40 hommes et 30 femmes aucune information n'est effacée concernant le type d'attribut sexe.
- si une cohorte se compose de 7 individus, 4 hommes et 3 femmes, l'attribut sexe n'est pas renseigné.
- si une cohorte se compose de 7 individus, uniquement des hommes, l'attribut sexe est renseigné avec une seule valeur homme, d'effectif égal à 7.
- si une cohorte se compose de 43 individus, 40 hommes et 3 femmes, l'attribut sexe n'est pas renseigné.
- si une cohorte se compose de 5 personnes ayant obtenu le bac en 2015, 6 personnes ayant obtenu le bac en 2016, 4 personnes ayant obtenu le bac en 2017, 3 personnes ayant obtenu le bac en 2018, alors l'attribut année du bac ne renseigne un effectif que pour les valeurs 2015 et 2016. Les valeurs 2017 et 2018 sont fournies, mais leur effectif est masqué par un zéro. La valeur 2019 (par exemple) initialement absente reste absente.
- si une cohorte se compose de 5 personnes ayant obtenu le bac en 2015, 6 personnes ayant obtenu le bac en 2016, 7 personnes ayant obtenu le bac en 2017, 3 personnes ayant obtenu le bac en 2018, alors l'attribut année du bac ne renseigne un effectif que pour les valeurs 2016 et 2017. Les valeurs 2015 et 2018 sont fournies, mais leur effectif est masqué par un zéro.

- si une cohorte se compose de 5 personnes ayant obtenu le bac en 2015, 6 personnes ayant obtenu le bac en 2016, 1 personne ayant obtenu le bac en 2017, 1 personne ayant obtenu le bac en 2018, l'attribut année du bac ne renseigne un effectif que pour la valeur 2016. Les valeurs 2015, 2017 et 2018 sont fournies, mais leur effectif est masqué par un zéro.
- si une cohorte se compose de 5 personnes ayant obtenu le bac en 2015, 1 personne ayant obtenu le bac en 2016, 1 personne ayant obtenu le bac en 2017, 1 personne ayant obtenu le bac en 2018, l'attribut année du bac n'est pas renseigné.

6.1 Fichier cohorte-bac.csv

cohorteid l'identifiant numérique de la cohorte auquel se rapporte la valeur de l'attribut

effectif l'effectif concerné par la valeur de l'attribut, ou zéro sur cet effectif a été masqué

bac-rgrp la valeur de l'attribut bac-rgrp (selon la nomenclature SISE)

6.1.1 Nomenclature SISE BAC-RGRP

code	libellé
1	Littéraire
2	économique
3	scientifique
4	techno stt
5	autres technos
6	professionnel
7	dispense
A	Bac général (2021)

6.2 Fichier cohorte-anbac.csv

cohorteid l'identifiant numérique de la cohorte auquel se rapporte la valeur de l'attribut

effectif l'effectif concerné par la valeur de l'attribut, ou zéro sur cet effectif a été masqué

anbac la valeur de l'attribut année du bac.

6.3 Fichier cohorte-sexe.csv

cohorteid l'identifiant numérique de la cohorte auquel se rapporte la valeur de l'attribut

effectif l'effectif concerné par la valeur de l'attribut, ou zéro sur cet effectif a été masqué

sexe la valeur de l'attribut sexe (selon la nomenclature SISE : 1 homme, 2 femme).

6.4 Fichier **cohorte-premiereins.csv**

cohorteid l'identifiant numérique de la cohorte auquel se rapporte la valeur de l'attribut

effectif l'effectif concerné par la valeur de l'attribut, ou zéro sur cet effectif a été masqué

annee la valeur de l'attribut première année d'inscription dans le supérieur en Île-de-France dans la limite de la fenêtre d'observation 2010-2021.

6.5 Fichier **cohorte-spe.csv**

Les deux spécialités de terminale au bac général forment une paire, que nous avons ordonné par ordre lexicographique.

cohorteid l'identifiant numérique de la cohorte auquel se rapporte la valeur de l'attribut

effectif l'effectif concerné par la valeur de l'attribut, ou zéro sur cet effectif a été masqué

bac-spe1 la première valeur de l'attribut paire de spécialités au bac

bac-spe2 la seconde valeur de l'attribut paire de spécialités au bac