Association Nationale des Dirigeants en Sciences de l'Information
www.andsi.fr

Les inscriptions à l'université en Open data par Pierre BOUDES, VP SI de Paris 13

Compte rendu rédigé par ANDSI

En bref…

Pierre BOUDES, Vice-Président des Systèmes d'Information de l'université Paris 13, a ouvert les données d'inscription et de parcours des étudiants, afin de cerner les flux entre universités, et procurer aux étudiants un « GPS de l'orientation », les aidant à déterminer le meilleur chemin entre départ et arrivée… souhaitée. Ce service est offert en respectant des règles strictes d'anonymisation des données.

L'Association Nationale des Directeurs des Systèmes d'Information organise des débats et en diffuse des comptes-rendus, les idées restant de la seule responsabilité de leurs auteurs. Elle peut également diffuser les commentaires que suscitent ces documents.

Ouverture des données, pourquoi ? Plusieurs raisons peuvent présider à cette démarche :

- la publication fait norme et, par besoin d'objectivité, suscite confiance et l'adhésion ;
- pour un effet de levier, pour que d'autres les améliorent, cependant cet objectif est difficile à atteindre ;
- pour l'interne ; amélioration des processus internes et du dialogue entre services ;
- pour renforcer la culture des données dans l'organisation ;
- pour créer un écosystème territorial ou par activité, ce qu'Henri Verdier (cf. réunion ANDSI sept 2014) appelle une « infrastructure essentielle » ;
- pour l'image cependant des risques (communication peu contrôlée…) existent ;
- et pourquoi pas pour éviter de les sauvegarder ?

Le contexte

L'université Paris 13 (P13) compte 24 000 étudiants, sur 5 sites + la maison des sciences de l'homme et le campus de Condorcet, en banlieue parisienne (essentiellement Seine St Denis), 30 laboratoires de recherche et de multiples partenariats.
Elle se caractérise également par la maison des sciences du numérique, dirigée par Younès BENNANI (cf. réunion ANDSI d'avril 2018) et par l'université numérique Île-de-France (UNIF) (500 000 étudiants), dirigée par notre collègue Dominique BASCLE, Administrateur de l'ANDSI et DSI de P13.
Les premières motivations pour l'ouverture des données sont un besoin d'objectivité dans le regard sur P13, un effet de levier sur le SI en stratégie d'innovation ouverte, ainsi que le renforcement de la culture des données (en qualité, non ressaisie…) et ainsi initier un cercle vertueux entre culture et qualité des données.
Le contexte externe est changeant ; COMmunautés d'UniversitEes (COMUES) et demande d'affirmer la légitimité de l'Université, la réussite des diplômes délivrés…
P13 fait partie des universités pilotes, qui échangent des données afin de notamment mesurer les flux inter-établissements. Ces mesures conditionneront l'intérêt du développement, par l'UNIF, d'outils de suivi des étudiants d'une université à l'autre.

Les premières données publiées, objet de cette conférence, sont 10 ans d'inscription des étudiants (avec lycée d'origine, nationalité, Bac), offrant ainsi la possibilité de retracer les parcours d'étude dans le supérieur.

Pourquoi partir de ce jeu de données de plus de 200 000 entrées ? :

- pour partie car ces données existent, et elles reflètent la production réelle de formation (et non l'offre de formation) ;
- et également car la démarche répond à des enjeux politiques en « osant la sincérité » de manière vérifiable, tout en étant très conservateur sur la protection des données.

La source est une application de gestion de scolarité, ancienne (plus de 25 ans), permettant de tracer les parcours, en anonymat. Cette préservation de l'anonymat a été éprouvée par un Hackaton (k anonymisation avec k=5 et suppression de la cardinalité si k<10), avec évaluation du risque de réidentification et généralisation (remplacer le pays par le continent) si nécessaire ou suppression des attributs trop singuliers qui sinon auraient permis une réidentification des personnes.

La publication a été faite en avril 2017 sur data.gouv.fr, avec 213 000 lignes de 2006 à 2015 pour 106 000 étudiants, puis élimination des valeurs rares (800 lignes). L'amélioration de la qualité des données a fait partie du process de publication.

Deux sortes de jeux de données sont produits : en sélectionnant des sous-ensembles d'attributs intéressants de la donnée initiale puis en anonymisant ces projections, ou bien en reconstituant de différentes façon les parcours d'études (dits traces) suivis par les étudiants durant les 10 années considérées. Une personne s'occupant d'une étape de diplôme (une année d'un parcours diplômant) peut aisément voir quelles sont les traces incluant cette étape, avec une grande richesse de représentations graphiques, développées en interne.
La plateforme de publication choisie a été data.gouv.fr qui fonctionne avec le logiciel udata développé par Etalab. La plateforme du ministère de l'enseignement supérieur et de la recherche n'étant pas accessible pour publication directe par les établissements. Cette dernière utilise le logiciel Opendatasoft plus complet en fonctionnalités que udata.
Le format choisi a été le CSV, mais à l'avenir, avec l'intégration de données plus hétérogènes dont des données des laboratoires de recherche, nous souhaitons compléter cette publication dans un format linked open data (des données ouvertes 5 étoiles, cf. Tim Berners-Lee). Ce format permet des requêtes raisonnablement complexes sur plusieurs sources en SPARQL ou GraphQL.

L'objectif est actuellement de croiser les données entre universités de façon mieux suivre les étudiants et étudier les flux inter-établissements.

Débat

Intervenant : les 800 suppressions ont-elles concerné les étudiants ou les années du cursus.
Pierre BOUDES : les années du cursus.

Int : Vous avez parlé d'un « GPS de l'orientation », qu'entendez-vous exactement par cela ?
P.B. : il s'agit de faire comme Google map : présenter les alternatives pour aller d'un point A (diplôme acquis) à un point B (diplôme désiré), en toute autonomie ou en préalable à un rendez-vous avec les conseillers d'orientation. D'autres données pourraient enrichir le service, comme la distance domicile/université, le coût des loyers autour de l'université… ceci afin de permettre un choix plus éclairé du parcours.

Int. : Les interruptions d'études sont-elles représentées ?
P.B. : Elles le sont uniquement sous forme de traces particulières où l'année d'étude a été conservée.

Int. : Dans le cursus, quand un étudiant part à l'étranger et revient, cela est-il géré ?

P.B. : Non, Les étudiants partant par exemple au Canada pour une année (cas non singulier) sont en général inscrits dans une étape de diplôme en France cette année-là et sont confondus avec les étudiants restés en France.

Présentation de l'orateur

Pierre BOUDES est enseignant chercheur à l'université Paris 13 au Laboratoire d'Informatique de Paris-Nord (LIPN-CNRS). Il s'intéresse notamment à la théorie de la démonstration et de la programmation. Il est actuellement vice-président des systèmes d'information à l'université Paris 13.