Note relative au processus d'anonymisation des données issues de la plate-forme FUN

La plate-forme de MOOC de France Université Numérique se base sur la technologie open source Open edX. Cette technologie permet de récolter un certain nombre de données quant à l'utilisation que les apprenants font des ressources, les interactions sur les forums, etc. Ces données sont utilisées dans le cadre de travaux de recherche, nécessaires pour améliorer les dispositifs pédagogiques mis en place. Avant d'être partagées avec des chercheurs et exploitées dans le cadre de leurs travaux, ces données doivent être anonymisées conformément à la législation en vigueur. Dans ce document, nous revenons sur le processus d'anonymisation à mettre en place.

Il faut dans un premier temps distinguer deux types de données.

* Les données portant sur le cours, les utilisateurs, et les actions des utilisateurs au sein de la plate-forme, forums exclus.
* Les données issues des forums.

Les données du premier type sont facilement anonymisables. Quant au second type de données, il existe des processus d'anonymisation des données de forums. Ces processus sont imparfait, le problème de l'anonymisation des forums restant une question ouverte de recherche. Nous reviendrons dans le cadre de ce document sur les méthodes d'anonymisation de ces deux types de données.

# I Données issues de la plate-forme, forums exclus

Il est possible de télécharger 7 fichiers de données à partir de la plate-forme edX. Cinq sont au format csv, le fichier de log est au format json, le fichier portant sur les données issues des forums est au format mongodb. Ces fichiers portent les noms suivants :

1. auth\_user.csv
2. auth\_userprofile.csv
3. certificates\_generatedcertificate.csv
4. courseware\_studentmodule.csv
5. student\_courseenrollment.csv
6. tracking\_log.json

Ces fichiers et les données afférentes sont disponibles à cette adresse :

 <http://data.edx.org/en/latest/internal_data_formats/sql_schema.html>

Nous allons revenir en détail sur chacun de ces fichiers en expliquant les données sensibles du point de vue de l'anonymisation.

## I.1 auth\_user.csv

Le fichier auth\_user.csv contient des données de base sur l'apprenant :

|  |  |
| --- | --- |
|  *id* | Clef primaire d'identification |
| username | Pseudonyme |
| first\_name | Prénom |
| *last\_name* | Nom de famille |
| *email* | Adresse email |
| *password* | Mot de passe |
| *is\_staff* | Statut au sein de la plate-forme |
| *is\_active* | Statut du compte (activé ou non) |
| *is\_superuser* | droits d'administration |
| *last\_login* | Date de dernière connexion à la plate-forme |
| *date\_joined* | date de création du compte |

Au cours du processus d'anonymisation, les colonnes username, first\_name, last\_name, email et password sont effacées. Le id est remplacé par un identifiant unique qui sera commun à toutes les bases de données.

## I.2 auth\_userprofile.csv

Le fichier auth\_userprofile.csv contient un certain nombre de données plus personnelles sur l'apprenant.

|  |  |
| --- | --- |
|  *id* | Clef primaire d'identification |
| user\_id | Clef étrangère d'identification |
| name | Nom et prénom |
| *meta* | données optionnelles sous la forme de texte libre |
| gender | genre |
| mailing\_address | adresse mail |
| year\_of\_birth | année de naissance |
| level\_of\_education | Diplôme le plus élevé obtenu |
| goals | objectifs à l'inscription, champ libre |
| allow\_certificate | Le participant est-il autorisé à recevoir un certificat |

Au cours du processus d'anonymisation, les colonnes name, mailing\_address et meta sont effacées. Les colonnes id et user\_id sont remplacées par un identifiant unique. Nous gardons les données sur le niveau de diplomation et l'année de naissance.

## I.3 student\_courseenrollment.csv

Une ligne de cette table représente l'engagement d'un étudiant dans un cours donné

|  |  |
| --- | --- |
|  *id* | Clef primaire d'identification |
| user\_id | Clef étrangère d'identification |
| course\_id | identifiant du cours |
| created | date d'inscription dans le cours |
| is\_active | le compte de l'utilisateur est-il actif dans le cours ? |

Au cours du processus d'anonymisation, les colonnes id et user\_id sont remplacés par l'identifiant unique.

## I.4 courseware\_studentmodule.csv

 Ces tables de données donnent des indications quant à l'utilisation des modules au sein des cours. Pour la description de ce qu'est un module, nous vous référons à la description donnée par le consortium edX.

|  |  |
| --- | --- |
|  *id* | Clef primaire d'identification |
| module\_type | type de module |
| *module\_id*  | clef d'identification du module |
| *student\_id*  | Clef étrangère d'identification |
| *state*  | éléments de description du module |
| *grade*  | note associée au module |
| *created*  | date de création de la ligne de cette base |
| *modified*  | date de dernière modification de la ligne |
| *max\_grade* | note maximale qu'il est possible d'atteindre avec ce module |
| course\_id  | Clef d'identification du cours |

Au cours du processus d'anonymisation, seul la colonne id est remplacé par l'identifiant unique commun à toutes les bases de données.

## I.5 certificates\_generatedcertificate.csv

Le fichier certificates\_generatedcertificate.csv est une base de données portant sur les certificats.

|  |  |
| --- | --- |
|  *id* | Clef primaire d'identification |
| user\_id  | Clef étrangère d'identification |
| *download\_url*  | url de téléchargement du certificat |
| *grade* | Note attribuée au moment de la délivrance du certificat |
| *course\_id*  | Clef d'identification du cours |
| *status* | statut du certificat |
| name | Nom présent sur le certificat |
| created\_date  | Date de création du certificat |
| modified\_date | Date de modification du certificat |

Au cours de ce processus d'anonymisation, la colonne "name" est supprimée, et la colonne user\_id est remplacée par un identifiant unique.

## I.6 tracking\_log.json

Les logs répertorient l'ensemble des actions réalisées sur la plate-forme: consultation des ressources, réalisation des tests, actions sur les forums. Ils permettent de dresser un historique détaillié des actions réalisées au sein de la plate-forme, utilisateur par utilisateur.

|  |  |
| --- | --- |
| *agent* | navigateur utilisé |
| *event* | caractéristiques de l'évènément |
| *event\_source* | source de l'événement (navigateur ou serveur) |
| *event\_type* | Type d'événement |
| *ip* | Adresse ip de l'utilisateur ayant déclenché l'événement |
| *page* | URL de la page sur laquelle se trouvait l'utilisateur au moment de l'événement |
| *session* | Clef d'identification de la session de l'utilisateur |
| *time* | Heure et date de l'événement (GMT) |
| *username* | Pseudonyme de l'utilisateur |

Au cours du processus d'anonymisation, le pseudonyme de l'utilisateur est remplacé par l'identifiant unique, l'adresse ip est remplacé par la localisation géographique. Les autres données sont conservées.

Ces données correspondent au données brutes, et sont difficilement exploitables en tant que tel dans le cadre d'un projet de recherche. Le format de base de données qui s'est imposé a été développé au MIT et se nomme MOOCdb (<http://francky.me/mit/moocdb/all/>). Il est désormais utilisé par le consortium edX, Coursera, et une partie conséquente de la communauté de recherche. L'anonymisation des données doit être fait de telle sorte qu'elle ne menace pas le processus de jointure entre tables et donc la constitution de MOOCdb. Par ailleurs, au moment d'attribuer un identifiant unique à chaque étudiant, il ne faut pas oublier que les étudiants s'inscrivent dans plusieurs cours, et qu'il est nécessaire de pendre en compte cet état de fait au moment d'attribuer un identifiant commun.

# II Données issues des forums

## II.1 Format des données issues des forums

Les données issues des forums sont plus difficiles à anonymiser que les données issues des logs. Elles sont stockées dans des bases de données au format MongoDB, sous la forme de collections JSON.

La description des données portant sur les forums est disponible à cette adresse :

<http://data.edx.org/en/latest/internal_data_formats/discussion_data.html>

|  |  |
| --- | --- |
| id | identifiant unique du post |
| type | type de post (commentaire ou nouveau fil de discussion) |
| anonymous | Le pseudonyme de la personne qui a posté commentaire est-il ou non anonyme |
| author\_id | clef étrangère d'identification de l'auteur du post |
| body | texte du post, encodé en UTF-8 |
| course\_id | Identifiant du cours dans lequel se trouve le post |
| created\_at | Date de création du post |
| updated\_at | Date de modification du post |
| votes | Nombre de votes associés au post |

Un certain nombre d'informations sont également associées aux fils de discussion et aux commentaires. Le tableau suivant représente les données associées aux fils de discussions.

|  |  |
| --- | --- |
| closed | Le commentaire a-t-il ou non été fermé par un modérateur |
| comment\_count | type de post (commentaire ou nouveau fil de discussion) |
| commentable\_id | identifiant associé au module sur lequel porte la discussion |
| last\_activity\_at | date de la dernière activité sur le post |
| title | titre du post, encodé en UTF-8 |

Les données associées aux commentaires sont représentées dans la table suivante :

|  |  |
| --- | --- |
| endorsed | Vaut 1 si un modérateur a approuvé la validité du commentaire (exemple = réponse correcte à une question posée par un autre utilisateur) |
| comment\_thread\_id | identifiant du fil de discussion sur lequel est posté le commentaire |
| parent\_id | identifiant du commentaire sur lequel porte le commentaire (dans le cas où le commentaire ne porte pas directement sur le fil de discussion) |
| parent\_ids |  |

Au cours du processus d'anonymisation, la colonne author\_id est remplacée par l'identifiant unique. Le véritable problème associé à l'anonymisation des forums concerne le contenu du texte (colonne body), qui peut contenir des informations susceptibles d'identifier les auteurs du post. Le fait de supprimer la colonne *body*, et la colonne *title* (titre du fil de discussion) permet d'anonymiser complètement les données.

Dans le cas où le travail de recherche inclurait l'analyse des corpus d'interaction, il faudrait procéder avant transmission des données à une anonymisation selon des méthodes semi-automatiques que nous explicitons par la suite. Ce travail d'anonymisation devra se faire en amont de toute transmission de donnée. L'anonymisation des corpus d'interaction à des fins de recherche est une question ouverte de recherche. Nous vous proposons un texte de Christophe Reffay détaillant une méthode d'anonymisation semi-automatique des corpus d'interactions, et qui sera appliquée dans ce contexte.

## II.2 Méthode interactive d’anonymisation semi-automatique de corpus d’interactions

Extrait de :

Christophe Reffay, François-Marie Blondel, Stéphane Allaire et Emmanuel Giguet (2012). **Anonymisation semi-automatique de corpus d’interactions éléments pour une méthode interactive**. Actes de la conférence JOCAIR'2012, Sidir, Baron Bruillard éditeurs. JOurnées Communication et Apprentissage Instrumentés en Réseau, Amiens, France, 7-9 septembre 2012, pp. 131-154. [Preprint in Archive](http://edutice.archives-ouvertes.fr/index.php?view_this_doc=edutice-00720211&version=1&halsid=ms3iagbetki4ga46j2p7nlpv33), [(Diaporama](http://www.stef.ens-cachan.fr/annur/reffay/RBAG_2012.ppt)).

MOTS-CLÉS : Anonymisation, Corpus, Partage de données de recherche, Fouille de graphies personnelles.

### Anonymiser : une nécessité légale pour le partage de données dans le respect de l’éthique

Que ce soit pour la consignation à long terme ou le partage au-delà d’un groupe de recherche, l’anonymisation des corpus d’interaction est indispensable pour des raisons éthiques et légales. Elle doit se faire à l’écart de la plateforme où les acteurs interagissent : sur une copie de ces interactions. Il faut bien distinguer les données identifiant les acteurs *dans le système d’information* (auteurs, destinataires, lecteurs, répondants, etc.) des extraits de contenus écrits par ces auteurs dans les interactions et qui désignent, par des données personnelles, sous des formes infiniment plus diverses, les acteurs impliqués ou ceux de leur environnement proches. Les identifiants peuvent être aisément supprimés remplacés ou masqués puisqu’ils sont nécessairement écrits à l’identique dans tout le corpus. En ce qui concerne les extraits pouvant dévoiler l’identité des acteurs, la tâche est bien plus complexe car elle ne revêt pas le caractère systématique des identifiants utilisés par le système. Étrangement, nous n’avons pas trouvé d’outils dans le champ du traitement automatique du langage naturel. La raison principale est que les techniques de ce champ utilisent la syntaxe et la structure des phrases ; ce qui les rend difficiles à utiliser dans le cas des productions plurilingues d’une part mais aussi dans celui des productions d’apprenants (qui sont rarement des textes parfaitement orthographiés et structuré). La méthode proposée dans (Reffay et al., 2012) aborde l’anonymisation des contenus d’interactions plurilingues produites par des apprenants.

Dans le cas des données d’interaction, anonymiser ne se limite pas à supprimer ou masquer les identifiants des personnes. Il s’agit bien de procéder à une *dépersonnalisation totale* (AFNOR 2000) des données afin de garantir « qu’une personne, étrangère au dispositif, ne puisse pas identifier l’un des acteurs de la formation ». Plus précisément, cela implique notamment de faire fi de toute information sensible qui pourrait permettre de retracer un individu, sans toutefois perdre d’éléments sémantiques à travers les interactions.

Après une présentation générale de la méthode, nous détaillons dans (Reffay et al., 2012) deux des techniques de fouilles et les illustrerons sur deux corpus très différents.

### Que cherche-t-on à anonymiser ? Pourquoi est-ce si difficile ? Illustration.

Il est très fréquent que les acteurs de la formation (enseignants, tuteurs, apprenants) utilisent dans leurs messages des éléments qui permettent de les identifier aisément. Ce sont par exemple des noms, prénoms, adresses de courriel, sites web personnels, adresses postales, numéros de téléphone, identifiants dans des outils de communication externes (Skype, Twitter, Facebook), lieux de résidence, institutions de rattachement, espaces fréquentés, etc. Toutes ces données doivent être masquées pour assurer l'anonymisation du corpus et donc la protection des acteurs. Le choix des éléments à repérer est une étape préliminaire à notre méthode. Ce choix doit être guidé par la règle suivante :

Chaque information conservée dans le corpus, prise séparément, doit pouvoir correspondre à de nombreuses personnes pour leur assurer l’anonymat. Même combinées en un faisceau d’informations, elles ne doivent pas permettre l’identification d’une personne physique. Pour illustrer cette règle, voici un exemple extrait du corpus « Nomades » de (Desprez, 2012) :

« Bonjour, je m'appelle **Kelly**. J'ai 16 ans, je suis une élève en **1ère S** dans le lycée **Rosa Luxemburg** à **Canet**, pas très loin de **Perpignan**. »

Kelly est un prénom fréquent (plus de 150 comptes facebook contenant cette graphie). Être en 1e S n’est pas rare en soit, mais préciser dans quel lycée et dans une ville de 3 000 habitants (qui n’a sans doute qu’un seul lycée) sont des informations bien trop précises qui permettent certainement de trouver la seule Kelly qui réponde à ces critères (en 2012). Dans ce cas précis, il faut impérativement masquer au moins le nom de l’institution et le nom de la (petite) ville.

Ce qui rend la tâche très difficile à automatiser entièrement relève de plusieurs aspects. Le premier vient du fait que ces marques (ex : « Paris ») sont susceptibles de variations, notamment syntaxiques, comme des erreurs (Pari, Parsi, paris, etc.) ou des altérations volontaires (Parigi en italien). Le deuxième provient de l’homonymie : une même graphie (i.e. : forme lexicale) peut représenter deux objets différents dans le monde physique. Exemple :

 « Sylvie **Paris** semble avoir développé une véritable addiction au PMU (**Paris** Mutuels Urbains). Elle fréquente assidument l’hippodrome de Longchamp à côté de **Paris**. »

Sur cet exemple fictif, il est facile de montrer combien il serait problématique de remplacer (à l’aveugle : automatiquement) toutes les occurrences de la graphie « Paris » par une même autre graphie (fusse-t-elle codée comme Dupond par exemple). D’abord, le texte perdrait de sa consistance comme dans l’explicitation du sigle : « PMU (Dupond Mutuels Urbains) ». Mais surtout, justement à cause de cette explicitation d’un sigle par ailleurs bien connu, ou pour la précision géographique en fin de phrase, il serait très facile pour un français de déduire que la graphie substituée par Dupond était « Paris » et donc, de ré identifier le patronyme de Sylvie. Dans ce cas, il faudrait pouvoir distinguer les occurrences de « Paris » qui représentent le patronyme de Sylvie des autres occurrences (PMU et ville) et n’effectuer le remplacement (par un pseudo) que lorsque « Paris » est utilisé comme patronyme désignant Sylvie.

Cette règle de transformation cesse de fonctionner quand le contenu relie explicitement une graphie personnelle (ex : patronyme) à un homonyme ou à une description (étymologie, signification, etc.) Pour reprendre l’exemple précédent, considérons l’extrait : « Je m’appelle Sylvie *Paris* (comme la *capitale française*)… ». La description du patronyme donnée entre parenthèses annule tout effet d’anonymisation. Dans cette situation, on peut choisir de ne pas transformer le patronyme « Paris » (qui peut être considéré comme suffisamment commun, on changera alors le prénom pour brouiller les pistes). On obtiendrait alors « Je m’appelle Sandrine *Paris* (comme la *capitale française*)… ». Ou bien, il faut envisager de modifier la description en rapport avec la graphie de remplacement ainsi : « Je m’appelle Sylvie *Dublin* (comme la *capitale irlandaise*) … »

Tous ces obstacles à une anonymisation entièrement automatisée nous ont conduits à envisager une méthode *interactive*, nécessitant l’intervention du chercheur, mais rendue *systématique* par l’assistance des outils de traitements qui s’y adossent.

### Une méthode d'anonymisation semi-automatique

Pour l’anonymisation de corpus d’interactions potentiellement multilingues, nous nous proposons de spécifier un outil adapté à ce type de données textuelles et répondant aux exigences de l’analyse de telles ressources par les chercheurs (a priori SHS). En proposant des techniques interactives de fouille et de marquage, nous visons une méthode interactive qui laisse le contrôle du niveau d’anonymisation au chercheur, mais qui le seconde dans les tâches complexes ou systématiques.

Dans notre méthode d’anonymisation, nous proposons de distinguer 7 étapes partant d’un corpus initial (à anonymiser) pour rendre, à la fin de la méthode, le corpus anonymisé :

1. Choisir les catégories d’informations (types d’entités) à identifier : noms, prénoms, institutions, villes, adresses de courriel, comptes facebook, MSN, numéros de téléphone, etc. ;
2. Catalogage : Lister toutes les graphies connues susceptibles de figurer dans le corpus et qu’il faut repérer et marquer (ex : prénom, patronyme, ville, courriel, et institution des participants) ;
3. Marquage : Marquer toutes les occurrences des graphies connues en les associant aux objets (entités réelles) qu’elles représentent (qui peuvent être différents en cas d’homonymie) ;
4. Fouille : Détection de nouvelles graphies ;
5. Définir pour chaque association (graphie - objet référencé) si la graphie doit être transformée au cours de la dernière étape (de substitution), et par quelle graphie de substitution ;
6. Vérifier la cohérence des graphies de substitution ;
7. Substitution de toutes les graphies (trop révélatrices) par leur graphie de substitution.

Les sept étapes de la méthode sont placées sur la figure 1 pour en expliciter l’enchaînement, les données nécessaires à chaque étape automatisable ou manuelle, ainsi que les données produites.

Les étapes 1 et 2 sont cruciales pour la qualité du résultat de l’anonymisation. Toutes les informations (externes au corpus) connues par le chercheur a priori sur les marques d’identification, sont autant d’indices permettant de trouver des références aux objets à repérer pour masquer. Plus ces indications sont précises et nombreuses, et plus large peut être la couverture des graphies repérées parmi celles susceptibles de ré-identifier les acteurs de la situation. Dans de nombreuses situations d’apprentissage, les chercheurs disposent d’une liste des participants avec quelques-unes de leurs caractéristiques : prénom, patronyme, institution, adresses, nationalité ou langue maternelle, etc. Ce sont précisément ces informations qu’il convient de recenser dans une première liste de graphies.

L’étape 3 de marquage peut être assistée d’outils (recherche dans un éditeur, concordancier) et doit permettre de retrouver à l’intérieur du corpus (dans leur contexte) toutes les occurrences des graphies listées au cours de l’étape 2. Pour chacune d’elles, le chercheur doit pouvoir choisir l’entité (objet du monde réel) à laquelle l’occurrence de cette graphie fait référence (ex : pour Paris : patronyme de Sylvie, PMU ou ville). Cette étape peut révéler de nouvelles entités (homonymes) à répertorier et à associer à la graphie correspondante. À l’issue de cette étape, chacune des occurrences des graphies listées est marquée et associée à une entité définie.

L’étape 4 est celle que nous détaillerons dans la suite de cette communication. Elle est constituée de processus (partie automatique) de fouille proposant de nouvelles graphies (potentiellement personnelles) au chercheur qui doit (partie manuelle) les retenir ou les écarter.

Lorsque l’étape 4 est fructueuse, i.e. elle a permis au chercheur de retenir de nouvelles graphies, il doit reprendre à l’étape 2 pour mettre à jour la liste des graphies dont il faudra marquer toutes les occurrences à l’étape 3. Si l’étape 4 ne détecte pas de nouvelles graphies, il passe à l’étape 5.

Ce n’est qu’à l’étape 5, quand le chercheur a une idée précise des entités représentées dans le corpus et des graphies (personnelles) recensées et marquées, qu’il est en mesure de décider avec finesse, quelles sont les graphies à remplacer pour rendre le corpus anonyme. Il doit alors choisir les pseudos qu’il peut utiliser pour ne pas dénaturer le corpus aux yeux des chercheurs susceptibles de l’analyser subséquemment.



Figure 1. Présentation générale de la méthode

L’étape 6 (entièrement automatique) doit vérifier certaines propriétés concernant les pseudos, et, le cas échéant, alerter le chercheur avant d’effectuer les substitutions (de l’étape 7) :

* Un même pseudo choisi pour deux graphies initialement différentes : ce qui risque d’engendrer des quiproquos qui n’avaient pas lieu d’être dans le corpus initial ;
* Un pseudo entre en collision avec une graphie (non modifiée) du corpus originel ;
* Une même graphie est remplacée par deux pseudos distincts : ce qui risque de rendre un quiproquo (dans le corpus d’origine) inconsistant dans le corpus anonymisé.

Le chercheur, ainsi alerté, peut bien sûr choisir de modifier certains pseudos en reprenant l’étape 5 ou d’ignorer les alertes et passer à l’étape 7 finale.

L’étape 7 de substitution est entièrement automatisable. Elle repose sur la table de transformation d’une part et le marquage de chaque occurrence d’autre part. La table précise pour chaque association (graphie – entité) par quel pseudo la graphie doit être remplacée. Dans le corpus marqué, chaque occurrence est associée à une entité. Ainsi, le processus parcourt le corpus (en une seule passe), et pour chaque occurrence marquée, vérifie dans la table si elle doit être remplacée. Dans l’affirmative, la graphie est substituée par le pseudo correspondant à l’association (graphie – entité) de cette occurrence.

La méthode rapportée ici n’a pas encore été implémentée dans un outil, mais nous pensons qu’elle est suffisamment systématique pour être développée pour peu qu’on trouve les ressources pour un tel travail.

Nous renvoyons le lecteur au texte (Reffay et al., 2012) qui met l’accent sur les détails concernant les techniques de fouille utilisées et leur application sur des exemples. Il y trouvera également la bibliographie qui a été supprimée dans cet extrait.