

BIGDATA (/TAG/BIGDATA/)



Anonymisation des données

👤 Jérôme Mainaud (/author/jmainaud/) 📅 15 Nov 2016

💬 0 Commentaires (/2016/11/15/anonymisation-des-donnees/#disqus_thread)



(/content/images/2016/10/anonymat.png)

Aujourd'hui, pratiquement toutes les applications possèdent des informations à caractère personnel, ne serait-ce qu'un identifiant de connexion ou une adresse électronique auxquels peuvent être attachées des informations plus ou moins sensibles. Les entreprises responsables chercheront à protéger ces données pour différentes raisons. Si les raisons purement morales ne sont que rarement suffisantes pour justifier des dépenses de sécurisation, les conséquences en termes d'images et le coût associé sont beaucoup plus moteurs. De plus, la législation européenne se durcissant, des conséquences judiciaires ou pénales ne peuvent plus être écartées. Ainsi, le règlement général sur la protection des données (https://fr.wikipedia.org/wiki/R%C3%A8glement_g%C3%A9n%C3%A9ral_sur_la_protection_des_donn%C3%A9es) (RGPD), règlement européen en vigueur depuis le 24 mai 2016 et applicable à partir du 25 mai 2018, prévoit des sanctions pouvant aller jusqu'à 4 % du chiffre d'affaires annuel mondial et 20 millions d'euros.

Le sujet de la protection des données étant un vaste sujet, nous allons nous attaquer aujourd'hui à deux techniques cousines qui permettent de réduire la sensibilité des données : l'anonymisation et le pseudonymat.

Anonymisation

L'anonymisation est une opération qui consiste à transformer des données personnelles de façon à ne plus permettre l'identification de la personne concernée. Cette transformation doit être irréversible. C'est-à-dire qu'il ne doit pas exister de méthode directe ou indirecte permettant de rattacher les données à la personne d'origine.

L'anonymisation peut-être motivée par des contraintes légales. Le considérant 26 du RGPD libère l'entreprise de ses obligations sur les jeux de données anonymisées. De même, les données comportant des informations à caractère personnel peuvent être conservées au-delà des délais de détention autorisées par la CNIL si elles sont convenablement anonymisées.

Cette technique est utile pour transmettre tout ou partie du jeu de données à un tiers qui a besoin de travailler sur les données réelles sans avoir besoin des données nominatives. Les cas sont nombreux :

- Exploitation des données pour des traitements statistiques.
- Création d'un jeu de tests réalistes pour les environnements hors-production.
- Reconstruction du jeu de production sur un environnement pour étudier un incident.

L'anonymisation sera importante dès lors qu'un être humain est susceptible d'accéder directement aux données sans passer par l'application usuelle et de son mécanisme de gestion de droit ou que les données sont transmises à un tiers, prestataire ou client, voire ouvertes au public sur une plateforme de données ouvertes.

Ainsi, un client qui gère des comptes de paiements souhaite que les développeurs n'aient accès qu'à une copie anonymisée de la base de production pour les analyse de documents pour éviter qu'ils regardent les données de personnalités connues, ou tout simplement de leurs collègues (ceux-ci faisant partie des premiers utilisateurs).

Pseudonymation

Il arrive parfois qu'on veuille rendre les données personnelles illisibles tout en conservant la possibilité de lever le secret. L'anonymat étant, en toute rigueur, irréversible, on parle dans ce cas de pseudonymat.

Un des intérêts du pseudonymat est de réduire la surface du risque. La masse des données étant pseudonymisées, leur stockage et leur traitement ne nécessite pas de protection supplémentaire par rapport à la sécurité conventionnelle. Seuls les données et le mécanisme de lever du pseudonymat ainsi que les traitements qui les utilisent doivent bénéficier d'un surcroît de sécurité. Par exemple, un data scientist peut explorer des données de navigation et en extraire une liste de prospects sous la forme d'une collection de pseudonymes d'adresse email. C'est uniquement au moment de l'envoi de la campagne qu'un composant spécifique et protégé remplacera le pseudonyme par la valeur réelle.

Un autre cas de pseudonymat, un peu particulier mais très courant, est celui utilisé pour la gestion des mots de passe. Au lieu de conserver les mots de passe en clair dans la base (ce qui est MAL™ !), on conserve une empreinte — si vous ne le faites pas, revoyez tout de suite vos pratiques. Comme le but n'est pas de retrouver le mot de passe d'origine, mais de vérifier la validité du mot de passe, on calcule l'empreinte du candidat qui est comparé avec celle de la base. Évidemment, le mode d'opération du calcul d'empreinte sera adapté à cet usage (bcrypt (<https://fr.wikipedia.org/wiki/Bcrypt>), PBKDF2 (<https://fr.wikipedia.org/wiki/PBKDF2>) ou le nouveau ARGON2 (<https://password-hashing.net/>)) et on ajoutera un peu de *sel* pour ne pas pouvoir détecter les valeurs identiques.

Des opérations difficiles

Éviter la réidentification d'une donnée est plus compliqué qu'il n'y paraît. Il n'est pas rare de sous-estimer la capacité d'un attaquant à retrouver l'identité d'origine. Prenons deux exemples qui s'appuient sur deux moyens différents.

Retrouver un individu particulier

Une technique parfois présentée comme utilisable pour anonymiser les données est le remplacement d'une donnée par une empreinte cryptographique. L'algorithme rend très difficile la recherche de la valeur initiale. Par contre, il est facile à calculer à partir de la valeur sans autre artifice et permet de garder la possibilité de recouper les données anonymes, ce qui peut être utile pour croiser les données anonymes. Seulement, s'il est difficile de lever l'anonymat, il est facile de vérifier si une donnée est présente dans la base. Il suffit de calculer l'empreinte de la donnée qu'on cherche et de regarder si on la retrouve.

Par exemple, voici un extrait d'un fichier fourni par WorldWideSourcing4JeanMarie™ aux recruteurs. Pour aguicher les clients, une version anonymisée (pseudonimisée, en fait) est distribuée gratuitement. Mais pour recevoir les informations nominatives (nom et email du candidat), il faut renvoyer la référence et payer 1000\$. Mais saurez-vous me dire si Jérôme Mainaud travaille chez Ippon ?

ref,nom,email,employeur ... 57259349-1ae7-457a-9650-c37ae322a11a,4aa94118991267f28e83475e912c5ce3260350f4,bf0ee06e3d54e67e2493e1937c8b5a4c1940f2ce,lppon Technologies f388246c-c9b0-4b55-ba99-c6565d579921,da9e74fb88524793da8a31a444f298c3448bb3a7,e5264e8cfe61517787196cd172ffe2c677b247c3,lppon Technologies 69c92375-a5be-4bc0-b639-79da04435d8b,45eef417a3465a75b8dc0f6dbaa7e603aaaba5b6,a2218710dca449c2888b1b2710ed29ec121fcbd4,lppon Technologies 7e12b5ef-60aa-460d-a0ff-e03168c97949,a6d88762a417bdb8c8e7926acbab152df1f77976,46429ed9d5c79b3a74003b12e42ed197cfe1de47,lppon Technologies ...

L'utilisation d'un sel est inutile ici, il ralentira juste le test et fera perdre la capacité à recouper les informations qui nous motivait à l'origine. Seule l'utilisation d'une fonction de hachage avec clé secrète dont on oublie la clé après la procédure peut être envisagée.

()Réidentification

Dans le premier exemple, la personne réalisant l'anonymisation a sous-estimé la capacité à effectuer une attaque partielle en ne levant pas l'anonymat en général, mais seulement celui de la personne qui l'intéresse.

Une autre erreur est de sous-estimer la capacité à identifier les individus avec très peu d'information. Il n'est pas rare de conserver dans les fichiers, le sexe, la date de naissance et la commune dans laquelle une personne habite à des fins statistiques. Ces données étant peu précises, les conserver semble parfaitement légitime. Vraiment ? Pourtant, en 2010, Latanya Sweeney a montré dans son étude « Uniqueness of Simple Demographics in the U.S. Population, Laboratory for International Data Privacy Working Paper, LIDAP-WP4 (2000) » que ces données suffisent à identifier 87 % de la population américaine. Autrement dit, il suffit d'avoir un fichier contenant le nom de la personne, sa date de naissance, son sexe et la ville où elle habite pour la réidentifier dans 87 % des cas. Et vu que ces données sont les premières qu'on vous demande lorsque vous vous inscrivez quelque part, un tel fichier ne doit pas être difficile à trouver.

Les solutions techniques

Voyons maintenant quelques techniques d'anonymisation et de pseudonymisation.

La suppression

Une méthode efficace pour anonymiser un jeu de données est d'effacer les données sensibles. Dans une base de données, on remplacera la valeur par `NULL`, ce qui signifie l'absence de valeur. Si toutes les valeurs permettant la réidentification ont bien été supprimées, alors l'anonymisation est garantie.

Cette solution ne peut pas toujours être mise en place lorsque les champs à anonymiser sont marqués comme obligatoire. Il n'est en effet pas possible de supprimer le contenu d'une colonne `NOT NULL`, alors qu'on peut vouloir garder cette contrainte sur les données légitimes. Il faudra alors choisir entre supprimer cette contrainte ou choisir une autre technique comme la mise à blanc.

Les données sensibles étant supprimées, le recoupement entre deux jeux de données n'est plus possible.

La mise à blanc

Lorsqu'on ne peut pas supprimer la valeur, par exemple parce que la présence d'une valeur est obligatoire ou que l'absence de valeur risque de perturber les affichages, il est possible de la remplacer par une valeur constante et non signifiante.

Il peut s'agir de blancs « », de soulignés « » ou de croix « ». Cette forme correspond à l'anonymisation par gommage ou par rayure au marqueur noir. L'utilisation de croix a l'avantage d'être visible sur un écran ou dans un édition.

Seul point d'attention, selon les cas, on souhaitera conserver la taille du mot remplacé. Ce ne doit être fait que si cette taille ne révèle pas d'information sur la donnée effacée.

Les données étant réécrites par une valeur constante, le recoupement entre deux jeux de données n'est plus possible.

Troncature

La troncature consiste à réduire la précision de l'information. Cela consiste à ne conserver que les premières lettres d'un nom ou à ne conserver que l'année d'une date de naissance.

Attention cependant, cette forme présente des risques de réversibilité importants. De simples initiales suffisent: si je dis « JFK a été assassiné par LHO » ou « NS et FH seront candidats », il y a de fortes chances que ayez une idée de qui on parle. De même, il n'y a pas beaucoup de JM qui travaillent au Pôle Conseil chez Ippon.

La réécriture des données ne permet pas le recouplement entre deux jeux de données.

Cette forme d'anonymisation est intéressante pour les données dont la forme tronquée reste utile. Une date de naissance réduite à l'année ou une adresse réduite à la ville permettent de réaliser des statistiques en limitant la quantité d'information transmise. Attention cependant à ce qu'on ne puisse réidentifier les personnes.

Substitution

La substitution consiste à remplacer les données par une autre valeur qui n'a aucun rapport avec la valeur d'origine. Elle présente de nombreuses variantes selon les objectifs.

Dans un contexte d'anonymisation, on remplacera les données sources par des données de même forme. L'utilisation d'un dictionnaire de valeurs dans lesquelles on tire au sort un substitut est une pratique courante : dictionnaire de noms, prénoms, noms de villes ou de rues. Les valeurs numériques ou les numéros de téléphone peuvent être simplement tirés au sort. Cette forme d'anonymisation est la plus efficace quand le jeu de données doit être utilisé pour des tests et des recettes. Les données sont proches de celles de la production par leur forme et leur distribution. Cela permet un fonctionnement cohérent de l'environnement de test. Il serait difficile de tester un moteur de recherche si tous les noms contiennent « xxxxxxxx » ou « Jean Dupont ».

En général, cette forme de substitution ne permet pas l'unicité du pseudonyme pour une valeur, mais ce n'est que très rarement un problème lorsque les bases sont normalisées et que les pseudonymes ne sont pas utilisés dans les clés primaires.

Dans un contexte de pseudonymisation, le but sera généralement de protéger la donnée lors d'un traitement et d'en permettre la levée à terme. Dans ce cas, on ne s'intéressera plus à la ressemblance des données (le type de pseudonyme pouvant être différent du type d'origine), qu'à la capacité de remplacer une valeur par un pseudonyme unique. La valeur sera typiquement remplacée par un nombre ou un UUID. Un dictionnaire bidirectionnel fera le lien entre la valeur et son pseudonyme.

Cette forme de pseudonymisation est très intéressante dans un contexte Big Data. L'unicité du pseudonyme est garantie et il est plus facile de sécuriser les données et les pseudonymes, relativement petit plutôt que la masse des données stockées sur des environnements Hadoop. De plus, l'anonymisation ultérieure est facile à effectuer en oubliant la correspondance dans le dictionnaire. <http://blog.ippon.fr>

Discovery to Delivery

Chiffrement

Le chiffrement est une forme de pseudonymisation qui remplace la valeur d'origine par une valeur illisible sans la clé de déchiffrement. Contrairement à la substitution, le secret se limite à la seule clé de déchiffrement, ce qui en simplifie encore la gestion et la sécurisation. L'anonymisation est possible en oubliant la clé de déchiffrement, mais elle ne peut être partielle.

L'unicité du pseudonyme, et donc la capacité à recouper les données, sont perdues par les modes d'opération fiables. Pour garder cette propriété, il faut choisir un mode faible comme ECB ([https://fr.wikipedia.org/wiki/Mode_d%27op%C3%A9ration_\(cryptographie\)#Dictionnaire_de_codes:_C2.AB_Electronic_codebook](https://fr.wikipedia.org/wiki/Mode_d%27op%C3%A9ration_(cryptographie)#Dictionnaire_de_codes:_C2.AB_Electronic_codebook)) avec tous les risques associés.

Hachage

Déjà vu dans les exemples ci-dessus, le calcul d'empreinte (ou hachage) est une forme de pseudonymisation particulière.

Elle ne permet pas de retrouver la valeur originelle sans conserver un dictionnaire. Elle est donc souvent considérée à tort comme une forme anonyme de la donnée. Mais sa forme par défaut présente deux propriétés importantes :

1. Elle est facile à calculer à partir de la valeur d'origine.
2. Elle garantit l'unicité du pseudonyme.

Résultat, il est possible de chercher les informations associées à une valeur particulière et de recouper avec des valeurs identiques. Des variantes permettent de tuer ces propriétés si on le souhaite.

1. Les algorithmes basés sur un secret (HMAC) rendent difficiles le calcul de l'empreinte et empêchent la recherche d'une valeur particulière.
2. L'ajout d'un sel aléatoire supprime l'unicité du pseudonyme.

Notez qu'en utilisant un sel et un secret en même temps, le mécanisme de hachage perd son intérêt, à l'exception des cas d'usage similaires à la conservation d'un mot de passe.

Pour aller plus loin

Quelques liens pour continuer :

- L'anonymisation de données en masse chez Bouygues Telecom (<http://www.ossir.org/jssi/jssi2011/1B.pdf>)
- Data Anonymization and Re-identification: Some Basics Of Data Privacy (<http://whimsley.typepad.com/whimsley/2011/09/data-anonymization-and-re-identification-some-basics-of-data-privacy.html>)
- Broken Promises Of Privacy: Responding To The Surprising Failure Of Anonymization (https://epic.org/privacy/reidentification/ohm_article.pdf)

Vous avez trouvé cette publication utile? Cliquer sur 

NEWSLETTER

Pour recevoir nos actualités, abonnez-vous à notre newsletter! Vous pouvez vous désabonner à tout moment.

Adresse Email *

Prénom *


Nom *

Entreprise

Je m'inscris


POST RÉCENTS

MIXIT 2018

 May 4, 2018


(/2018/05/04/mixit-2018/)

Arduino Odyssey

 May 2, 2018

(/2018/05/02/arduino-odyssey/)

Peut-on être Agile sans être agile ?

 Apr 27, 2018

WE ARE IPPON



(<http://www.ippon.fr/nous-rejoindre/>)

Partagez cet article:

f (<https://www.facebook.com/sharer/sharer.php?u=http://blog.ippon.fr/2016/11/15/anonymisation-des-donnees/>)

t ([https://twitter.com/share?](https://twitter.com/share?text=Anonymisation%20des%20donn%C3%A9es&url=http://blog.ippon.fr/2016/11/15/anonymisation-des-donnees/)

[text=Anonymisation%20des%20donn%C3%A9es&url=http://blog.ippon.fr/2016/11/15/anonymisation-des-donnees/](https://twitter.com/share?text=Anonymisation%20des%20donn%C3%A9es&url=http://blog.ippon.fr/2016/11/15/anonymisation-des-donnees/))

g+ (<https://plus.google.com/share?url=http://blog.ippon.fr/2016/11/15/anonymisation-des-donnees/>)

d (<http://www.digg.com/submit?url=http://blog.ippon.fr/2016/11/15/anonymisation-des-donnees/>)

r (<http://reddit.com/submit?url=http://blog.ippon.fr/2016/11/15/anonymisation-des-donnees/&title=Anonymisation des données>)

in (<http://www.linkedin.com/shareArticle?mini=true&url=http://blog.ippon.fr/2016/11/15/anonymisation-des-donnees/>)



st (<http://www.stumbleupon.com/submit?url=http://blog.ippon.fr/2016/11/15/anonymisation-des-donnees/&title=Anonymisation des données>)



(/author/jmainaud/)

Jérôme Mainaud (/author/jmainaud/)

Ippon

Ippon est un cabinet de conseil en technologies, créé en 2002 par un sportif de Haut Niveau et un polytechnicien, avec pour ambition de devenir leader sur les solutions Digitales, Cloud et BigData.

Ippon accompagne les entreprises dans le développement et la transformation de leur système d'information avec des applications performantes et des solutions robustes.

Ippon propose une offre de services à 360° pour répondre à l'ensemble des besoins en innovation technologique : Conseil, Design, Développement, Hébergement et Formation.

Nous avons réalisé, en 2017, un chiffre d'affaires de 31 M€ en croissance organique de 30%. Nous sommes aujourd'hui un groupe international riche de plus de 320 consultants répartis en France, aux USA, en Australie et au Maroc.


FRANCE **WEBSITE** ([HTTP://WWW.IPPON.FR](http://www.ippon.fr))

in **LINKEDIN** ([HTTPS://WWW.LINKEDIN.COM/COMPANY/IPPON-TECHNOLOGIES](https://www.linkedin.com/company/ippon-technologies))

[Post précédent](#)**RxJS**[\(/2016/11/16/rxjs/\)](/2016/11/16/rxjs/)[Poste suivant](#)**TamTam - Sortie de Scala 2.12**[\(/2016/11/14/tamtam-sortie-de-scala-2-12/\)](/2016/11/14/tamtam-sortie-de-scala-2-12/)

0 Commentaires

Blog Ippon Technologies


 S'identifier ▾ Recommander Partager

Les meilleurs ▾



Commencer la discussion...

S'IDENTIFIER AVEC

OU INSCRIVEZ-VOUS SUR DISQUS 

Nom

Soyez le premier à commenter.

ÉGALEMENT SUR BLOG IPPON TECHNOLOGIES

OSRM : le chemin le plus court pour tracer la route !

1 commentaire • il y a 2 mois

**Olivier FARLOTTI** — Article très clair et didactique !
Merci !**La dette technique tue les grenouilles**

2 commentaires • il y a 2 mois

**fracolo** — L'histoire de la grenouille est une fable.
Expérimentalement, les grenouilles ne réagissent pas comme ça....**Big Data 2.0 ?**

2 commentaires • il y a 2 ans

**Christophe PARAGEAUD** — Bonjour, la question est aussi vaste que le sujet. Liste ressources Big ...**Node-Red : l'IoT à portée de tous**

1 commentaire • il y a 7 mois

**Simon** — Bonjour, Cet article est très intéressant. Du coup, j'ai testé Node Red qui est effectivement très simple de prise en main. ...

IPPON

(<http://www.ippon.fr/>)

Discovery to Delivery

Fiers de notre passion pour la technologie et l'expertise dans les systèmes d'information, nous nous associons à nos clients pour offrir des solutions innovantes pour leurs projets stratégiques.

Grâce à la communication, à la curiosité et à l'esprit d'équipe, nous proposons les meilleures solutions à nos clients.

© 2018 IPPON (<http://blog.ippon.fr/>). Tous les droits sont réservés.