À DÉFAUT DE RÉDUIRE LA COLLECTE DES DONNÉES, COMMENT LES ALTÉRER?

LE 5 FÉVRIER 2011 ***HUBERT GUILLAUD***

Des chercheurs ont créé un algorithme qui altère des données génétiques ou médicales afin de les anonymiser tout en permettant aux chercheurs de les utiliser.

Toutes les données sont devenues personnelles, écrivions-nous il n'y a pas si longtemps, montrant combien anonymiser les données devenait difficile, à l'heure où les champs de données eux-mêmes génèrent de l'identifiabilité. Paul Ohm (**blog**, [en]), **dans un article important sur l'étonnant échec de l'anonymisation** [en] annonçait déjà, qu'il n'y aurait pas de solutions miracles : *"les mesures qui sont prises augmenteront la confidentialité ou réduiront l'utilité des données, mais il n'y aura aucun moyen de garantir à la fois une utilité maximale des données et une confidentialité maximale."*

Dans la ***Technology Review*** [en] on apprend que des chercheurs du **Laboratoire de protection des renseignements médicaux** [en] de l'université Vanderbilt ont créé un algorithme pour altérer des données génétiques ou médicales afin de les anonymiser tout en permettant aux chercheurs de les utiliser.

Les enregistrements médicaux comportent de nombreuses informations sur les patients, allant de leur âge à leur historique médical. Quand ces données sont utilisées par des chercheurs, elles sont "anonymisées", c'est-à-dire qu'on enlève les identifiants directs comme le nom ou l'adresse, mais pas bien sûr les diagnostics et leurs historiques. Le problème est qu'il n'est pas difficile d'utiliser ces historiques pour ré-identifier une personne. Dans l'article publié dans *Proceedings of the National Academy of Sciences* par **Bradley Malin** [en] et ses collègues, ceux-ci estiment qu'ils sont capables d'identifier 96 % des patients en se basant seulement sur leurs historiques médicaux.

Pour résoudre ce problème, l'équipe du Laboratoire de protection des renseignements médicaux a conçu un algorithme capable de chercher dans une base de données les combinaisons de diagnostic qui distinguent un patient d'un autre et de les substituer par d'autres. Ainsi, le code qui distingue une ostéoporose post-ménopause pourrait devenir une simple ostéoporose… L'algorithme injecte des informations altérées afin de rendre les enregistrements des patients non identifiables. L'algorithme serait également capable d'ajuster le niveau d'anonymisation aux besoins des chercheurs, selon leurs recherches.

Quelques limites

Cette nouvelle approche comporte néanmoins quelques limites estiment les chercheurs : le système fonctionne mieux quand les chercheurs ont un but précis, afin que les bonnes données, qu'ils cherchent à exploiter, soient préservées par le système. Ce qui signifie qu'une même extraction ne pourrait pas servir à plusieurs recherches. Inversement, accéder à plusieurs extractions d'un même ensemble devrait certainement permettre, en les croisant, de rétablir les données altérées…

Si l'avenir de la science dépend de la façon de tirer parti d'informations existantes dans des silos de données, l'anonymisation de l'information demeure une question primordiale. Comme souvent, les chercheurs semblent pragmatiques : il leur faut maximiser le bénéfice scientifique tout en contrôlant les risques quant à la vie privée.

L'intérêt en tout cas de l'algorithme mis au point par les chercheurs, est de permettre d'aller plus loin qu'une fausse anonymisation des données et de montrer que la science prend le problème au sérieux. Reste que plutôt que d'augmenter la confidentialité des données, on devine que c'est l'autre option sur laquelle tout le monde va travailler : trouver les moyens d'en réduire l'utilité à minima. C'est la piste que tracent ces premières recherches… Il est possible que ce ne soient pas les dernières.

—

Article initialement publié sur **InternetActu** en avril 2010

Image CC Flickr **sombraala**

1 ping

Les tweets qui mentionnent À défaut de réduire la collecte des données, comment les altérer ? » Article » OWNI, Digital Journalism -- Topsy.com le 5 février 2011 - 11:55

[...] Ce billet était mentionné sur Twitter par Romain Pigenel, Pascal, jean polochon, Jazz, Herve Le Duc et des autres. Herve Le Duc a dit: [OWNI] À défaut de réduire la collecte des données, comment les altérer ? http://bit.ly/hPAaBz [...]