

de transformer nos vies autant que l'invention de l'écriture ou d'Internet.

Bien sûr, cette transition vers une société orientée par les données est un défi. Dans un monde de données illimitées, la méthode scientifique à laquelle nous sommes habitués ne fonctionne plus : il y a tellement de connexions et de corrélations potentielles entre les données que les outils statistiques classiques fournissent souvent des résultats absurdes. L'approche scientifique classique donne de bons résultats quand l'hypothèse est claire et que les données sont conçues pour répondre à la question. Mais dans la complexité inextricable des systèmes sociaux à grande échelle, il existe souvent des milliers d'hypothèses raisonnables, et il est impossible d'ajuster les données à toutes ces hypothèses en même temps.

Ainsi, à l'avenir, nous devons analyser les données du monde réel pour faire émerger les lois sous-jacentes plus systématiquement et plus en amont que jusqu'ici.

Nous devons construire des « laboratoires vivants » où nous pourrions tester nos idées pour construire des sociétés orientées par les données.

Un exemple d'un tel laboratoire vivant est la Cité des données ouvertes que nous venons d'inaugurer à Trente, en Italie, avec la coopération des élus, de *Telecom Italia*, de *Telefonica*, de la fondation Bruno Kessler et de l'Institut pour la conception dirigée par

POUR LA PREMIÈRE FOIS DANS L'HISTOIRE nous pouvons observer la société avec une précision suffisante pour envisager de construire des systèmes sociaux plus performants.

les données (IDDD). Le but de ce projet est de démultiplier le flux d'idées au sein de la ville de Trente. Des logiciels tel que notre système openPDS (*Personal Data Store*), qui implémente les principes de la nouvelle donne pour les données, permettent aux individus de partager en toute sécurité des données personnelles (telles des informations relatives à leur santé ou à leur famille) en contrôlant

où va l'information et quelle utilisation en est faite. Par exemple, une des applications d'openPDS encourage le partage entre les familles des bonnes pratiques concernant l'éducation des jeunes enfants. Comment les familles dépensent-elles leur argent ? Quels sont la fréquence de leurs sorties et leur niveau de socialisation ? Une fois que l'individu a donné son accord, ses données peuvent être rassemblées, rendues anonymes et partagées automatiquement en toute sécurité avec d'autres familles.

Nous pensons que de telles expériences montreront que les avantages potentiels d'une société orientée par les données valent la peine de faire ces efforts et de prendre ces risques. Imaginez : nous pourrions anticiper les krachs financiers, détecter et prévenir les épidémies, utiliser au mieux les ressources naturelles et encourager la créativité pour prospérer. Ce rêve pourrait devenir bientôt une réalité... si nous évitons les écueils avec prudence. ■



Muséum national d'Histoire naturelle

Le stratotype Hettangien

Coordonné par Micheline Hanzo

Le temps du géologue est divisé en étages définis par les fossiles qu'ils contiennent. Afin d'établir des références universelles, certains sites où les couches géologiques correspondant à ces étages affleurent ont été qualifiés de « stratotypes ».

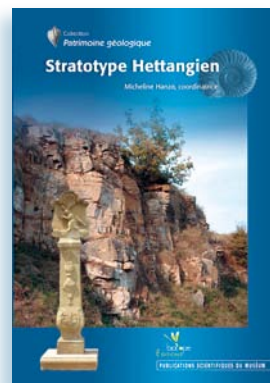
La collection *Patrimoine géologique* a pour objet de publier des synthèses sur chacun des stratotypes situés en France.

Le stratotype de l'Hettangien est donc une référence géologique internationale. Il est de ce fait aussi la célébrité d'Hettange-Grande, localité de Moselle où il a été défini.

Suite à bien des controverses, c'est Eugène Renevier qui, en 1864, en fut l'initiateur, séduit par les travaux de son prédécesseur Olry Terquem, sur des niveaux particulièrement riches en fossiles.

Ce stratotype historique méritait d'être protégé ! Il obtint le statut de Réserve naturelle nationale d'Hettange-Grande en 1985. *Bildstock*, pavés, pierres à four, etc., témoignent de l'exploitation du matériau gréseux du site stratotypique.

Bien documenté, richement illustré, cet ouvrage constitue une synthèse des connaissances acquises sur l'Hettangien au niveau de son stratotype, et de la Lorraine belgo-luxembourgeoise, en évoquant plus brièvement d'autres sites du même âge en France.



317 p. • 232 figs • 35,90 € TTC • format 16,5 x 24 cm
ISBN 978-2-85653-688-9

ouvrage en vente à la librairie des Publications
Scientifiques face à la Grande Galerie de l'Évolution

PUBLICATIONS SCIENTIFIQUES DU MUSÉUM

Commandes et renseignements :

Muséum national d'Histoire naturelle
Publications Scientifiques

Casé postale 41 • 57 rue Cuvier • 75231 Paris cedex 05
Tél. : 01 40 79 48 05 • Fax : 01 40 79 38 40
diff.pub@mnhn.fr • <http://www.mnhn.fr/pubsci>

Comment préserver l'anonymat

Tristan Allard, Benjamin Nguyen et Philippe Pucheral

Presque tous nos comportements laissent des empreintes numériques. Les informaticiens conçoivent et développent des techniques pour que l'analyse de ces gisements de données ne compromettent pas la vie privée.

Santé, déplacements, achats, appels téléphoniques, réseaux sociaux, recherches d'information : toutes ces facettes de nos vies laissent des empreintes numériques sur Internet ou sur divers appareils. Elles sont stockées et classées dans des bases de données gigantesques, telles celles des moteurs de recherche, qui gardent l'historique des requêtes associées à une adresse IP (qui identifie un appareil connecté à Internet) pendant plusieurs années. Ces masses d'informations représentent une manne sans précédent pour les sociétés humaines (voir l'article d'Alex Pentland dans ce numéro). Dans le domaine de la santé, par exemple, on peut analyser les caractéristiques individuelles de chaque patient, afin de mieux le prendre en charge, ou réaliser des études de santé publique : l'analyse du nombre de fois où le mot « grippe » est tapé dans un moteur de recherche dans une certaine région

(reconstituée approximativement avec les adresses IP) renseigne ainsi sur l'importance locale de l'épidémie. De même, les données de mobilité pourraient être enregistrées par les systèmes de géolocalisation (GPS) et aider à optimiser l'aménagement urbain. Si l'analyse de données sur les individus est pratiquée depuis des millénaires, notamment lors des recensements démographiques, la quantité et la diversité des informations disponibles aujourd'hui en multiplient l'intérêt potentiel.

Cette manne d'informations fait cependant planer une menace, elle aussi sans précédent, sur la vie privée des individus. La plupart du temps, les données ne sont pas publiques, mais elles circulent tout de même entre différents acteurs : des ensembles de données médicales sont transmis à des organismes de recherche, les sites en ligne peuvent vendre (légalement ou non) une partie des données personnelles de leurs

utilisateurs, etc. Certaines données sont même accessibles à tous sur Internet : le moteur de recherche de Facebook (uniquement disponible en anglais pour l'instant) permet, par exemple, de trouver les utilisateurs qui aiment tel sport, ont tel âge, habitent telle région... Autre exemple, le site Netflix, qui recommande des films en fonction des goûts de chacun, a publié des données de ses utilisateurs à l'occasion d'un concours visant à optimiser les algorithmes qui déterminent les films à recommander.

Un compromis entre utilité et protection

Or de nombreuses études montrent qu'il est souvent possible d'identifier la personne associée à un jeu de données, même quand ce jeu ne contient pas son nom, ni ses coordonnées. Des pans entiers de la vie privée peuvent être dévoilés, avec



des préjudices multiples : discrimination au crédit bancaire ou à l'assurance selon l'état de santé, discrimination à l'emploi selon l'orientation sexuelle ou le groupe ethnique, etc. Protéger les données personnelles sans les rendre inexploitable est donc un enjeu majeur à l'ère du tout numérique. La diffusion de ces données se révèle être un art de l'équilibre, où l'on recherche le meilleur compromis entre utilité des données et protection des individus. C'est cet « art » que nous examinerons ici.

Le souci de la protection des données s'est accru pendant la seconde moitié du XX^e siècle, à mesure que l'informatique se généralisait. Il a fait naître un domaine de recherche spécifique dans les années 1970, visant à garantir un anonymat plus ou moins complet aux titulaires des données. Un ensemble de statistiques sur des données personnelles peut révéler beaucoup d'informations sur un individu précis. Par exemple,

L'ESSENTIEL

■ De nombreuses données personnelles sont publiées sur Internet ou transmises à divers organismes en indiquant leurs titulaires par un pseudonyme.

■ Cela ne suffit pas à garantir l'anonymat des données, d'où une menace sur la vie privée.

■ De nombreuses techniques sont développées pour protéger les données sensibles sans les rendre inexploitable.

si un recensement liste le revenu moyen par habitant avec un découpage du territoire en carrés de 200 mètres de côté, le propriétaire d'un grand domaine risque d'être seul dans la bande de territoire considérée ; un individu qui le sait peut donc déduire son revenu de l'analyse des statistiques. On voit ici que la quantité de renseignements extractibles de données publiées dépend de ce que l'on sait par ailleurs. Le développement des ordinateurs ayant rendu les informations disponibles plus nombreuses et plus accessibles, le danger s'est multiplié.

Le phénomène s'est encore accéléré depuis le début des années 2000, avec l'essor d'Internet, qui permet de consulter et de croiser instantanément ou presque de multiples sources d'information (réseaux sociaux, annuaires, listes de diplômés, etc.). Lors de la publication de données, il est devenu crucial de prendre en compte les connaissances annexes accessibles à un



1. LES REQUÊTES TAPÉES DANS UN MOTEUR DE RECHERCHES renseignent sur l'identité de celui qui les soumet. Ainsi, supposons que l'on sache que l'individu 6208 s'intéresse à l'arbre généalogique des Dupont et qu'il a cherché un coiffeur et un boulanger rue Férou. En croisant ces requêtes avec des données disponibles sur Internet,

telles les pages blanches, on découvre qu'un certain Marc Dupont habite rue Férou. On peut alors avoir accès aux requêtes de cet individu et s'apercevoir, par exemple, qu'il s'intéresse à la dépression (sans doute en souffre-t-il) ou qu'il a consulté des pages où l'on propose des rencontres discrètes. Sa vie privée est menacée.

individu mal intentionné. En outre, les analystes se contentent de moins en moins de statistiques sur les données personnelles et demandent d'accéder directement à celles-ci, afin d'augmenter la précision de leurs études.

La conjonction de ces deux facteurs a exacerbé la nécessité d'une protection robuste. Au cours des dix dernières années, de nombreux chercheurs ont étudié la façon de publier des données tout en préservant la vie privée des individus concernés. Ils ont développé des méthodes dites d'assainissement des données, qui consistent à dégrader leur précision de façon contrôlée, afin de réduire la probabilité que l'on puisse réidentifier la personne correspondante ou accéder à des informations sensibles la concernant.

Les faiblesses de la pseudonymisation

Comment rendre anonyme un jeu de données ? Le moyen le plus ancien et le plus intuitif est de remplacer le nom de l'individu concerné par un pseudonyme ou un nombre – on parle de pseudonymisation. C'est la méthode qu'a utilisée l'entreprise AOL, en août 2006, lorsqu'elle a mis en ligne 20 millions de requêtes adressées à son moteur de recherche par 650 000 utilisateurs sur une période de trois mois (officiellement pour les rendre disponibles aux chercheurs de divers domaines). Les noms des utilisateurs étaient remplacés par des nombres aléatoires, mais les mots clés des recherches étaient laissés en clair. Quelques jours plus tard, des journalistes du *New York Times* ont identifié la personne portant le pseudonyme 4417749.

■ LES AUTEURS



Tristan ALLARD est postdoctorant à l'Inria (Institut national de recherche en informatique et en automatique).



Benjamin NGUYEN est chercheur à l'Inria et maître de conférences à l'Université de Versailles Saint-Quentin-en-Yvelines.



Philippe PUCHERAL est chercheur à l'Inria et professeur à l'Université de Versailles Saint-Quentin-en-Yvelines.

Elle avait tapé des centaines de requêtes, tels « paysagiste à Lilburn, Géorgie », « individus dont le nom de famille est Arnold » ou « hommes célibataires de 60 ans ». Les journalistes ont croisé ces mots clés avec des données disponibles sur le Web et ont peu à peu reconstitué l'identité probable de l'individu correspondant, une veuve de 62 ans habitant Lilburn, qu'ils ont fini par retrouver.

De nombreux cas similaires illustrent l'échec de la pseudonymisation à rendre les données anonymes. Remplacer les seuls champs directement identifiants, tels que le nom et le prénom ou le numéro de sécurité sociale, ne prémunit pas contre une réidentification ultérieure (voir la figure 2). Pourtant, la pseudonymisation de données reste aux yeux de la loi une forme acceptable d'anonymisation, sans doute parce que ses failles sont mal comprises par les législateurs.

Prévenir les croisements de données

Au début des années 2000, Latanya Sweeney, de l'Université Carnegie Mellon, aux États-Unis, a proposé une méthode, nommée *k-anonymat*, pour prévenir les réidentifications *via* des croisements de données. Elle a d'abord montré qu'il était parfois possible de retrouver le titulaire d'un jeu de données médicales pseudonymisé à partir des champs (Sexe, Date de naissance, Code postal). En effet, ces champs sont aussi présents dans d'autres jeux de données comprenant le nom du titulaire (voir la figure 2). En outre, la combinaison des renseignements correspondants est