Recueil du Symposium 2016 de Statistique Canada
Croissance de l'information statistique : défis et bénéfices

Les mégadonnées dans la perspective de la recherche par enquête

Reg Baker[1]

Résumé[2]

Le terme mégadonnées peut signifier différentes choses pour différentes personnes. Pour certaines, il s'agit d'ensembles de données que nos systèmes classiques de traitement et d'analyse ne peuvent plus traiter. Pour d'autres, cela veut simplement dire tirer parti des ensembles de données existants de toutes tailles et trouver des façons de les fusionner, avec comme objectif de produire de nouveaux éléments de connaissance. La première perspective présente un certain nombre de défis importants pour les études traditionnelles de marché, recherches sur l'opinion et recherches sociales. Dans l'un ou l'autre cas, il existe des répercussions pour l'avenir des enquêtes, qu'on commence à peine à explorer.

1. Introduction

En 1987, pour commémorer le 50e anniversaire de la publication, les responsables de *Public Opinion Quarterly* ont demandé à 16 universitaires et praticiens d'enquête bien connus de soumettre leur vision de l'avenir de la recherche sur l'opinion publique (Bogart, 1987). Dans sa réponse, James Beniger mentionne qu'« un ensemble de nouvelles technologies rendra possible le contrôle de masse en temps réel du comportement individuel... la recherche par enquête donnera de plus en plus lieu à des mesures plus directes du comportement, rendues possibles par les nouvelles technologies informatiques. »

Près de trois décennies plus tard, cette perspective, bien qu'elle ne soit pas encore concrète, est à tout le moins plus clairement perceptible. Il s'agit du monde des mégadonnées, et selon le contexte dans lequel on se trouve et le type de recherche que l'on effectue, c'est un rêve devenu réalité ou l'apocalypse.

Le présent document commence par une définition de ce que nous entendons par le terme « mégadonnées ». Il porte ensuite sur certains des défis auxquels nous faisons face lorsque nous utilisons le paradigme des mégadonnées. Il conclut avec certaines réflexions concernant les répercussions des mégadonnées sur l'avenir des enquêtes.

2. Définition des mégadonnées

Qu'entendons-nous exactement lorsque nous parlons de « mégadonnées »? La définition la plus souvent entendue est celle des trois V – volume, variété et vélocité (Laney, 2001). Même s'il s'agit d'un sommaire des défis que posent les mégadonnées, il ne s'agit pas du tout d'une définition. Ward et Barker (2013) ont examiné les définitions les plus souvent utilisées par divers intervenants de l'écosystème des mégadonnées, qui est constitué d'experts-conseils des

[1] Marketing Research Institute International, 5073, Red Fox Run, Ann Arbor, Michigan, États-Unis, 48105

[2] Le présent document est fondé sur un chapitre portant le même titre dans Paul Biemer, Edith de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars Lyberg, Clyde Tucker et Brady West (rédacteurs), *Total Survey Error in Practice*, qui doit être publié par Wiley, au début de 2017. En raison d'une entente concernant les droits d'auteur avec l'éditeur, nous ne pouvons pas présenter ce chapitre ici, mais uniquement un bref résumé.

TI, de fabricants de matériel, de développeurs de logiciels et de fournisseurs de services. Ils ont noté que la plupart des définitions comprennent trois attributs primaires : taille, complexité et outils. Ils ont proposé cette définition :

> « Les mégadonnées désignent l'entreposage et l'analyse d'ensembles de données importants ou complexes, au moyen d'une série de techniques, y compris, sans s'y limiter : NoSQL, MapReduce et l'apprentissage automatique. »

Plus simplement, nous pourrions aussi dire que :

> Le terme « mégadonnées » renvoie à des ensembles de données qui sont si importants et complexes qu'ils ne peuvent pas être traités et analysés au moyen de logiciels conventionnels.

Dutcher (2014) a demandé à 40 leaders d'opinion différents leurs définitions et a obtenu près de 40 réponses différentes. Ces réponses sont résumées dans le nuage de mots de la figure 1.

Figure 1
Nuage de mots tiré de Dutcher (2014).

Nous pouvons aussi élaborer davantage sur le sujet en soulignant trois sources principales :

1. Les **données transactionnelles**, qui décrivent un événement, par exemple, une interaction entre une personne et une entreprise ou une entité gouvernementale. Les données des clients et les données administratives des gouvernements sont deux exemples courants.
2. Les **données des médias sociaux** tirées (ou recueillies) à partir de réseaux de médias sociaux, de blogues, d'études sur le commerce électronique, de recherches sur le Web et d'autres sources de texte en format libre et de données non traditionnelles, comme des photographies et des enregistrements vidéo.
3. L'**Internet des objets (IdO)**, soit les données recueillies à partir de dispositifs interconnectés qui englobent une vaste gamme d'objets, y compris les automobiles, les appareils électroménagers, les scanneurs, les feux de circulation, les caméras de sécurité, les capteurs portables, les GPS, etc. L'IdO est la source la moins déployée, mais celle qui pourrait avoir la plus grande portée parmi ces trois sources de mégadonnées.

3. Certains défis

Les chercheurs qui s'intéressent à l'exploitation du potentiel de ces données font face à au moins trois défis principaux.

3.1 Technologie

Les mégadonnées sont un monde de téraoctets, de pétaoctets, d'exbioctets et de zettaoctets. C'est Walmart qui saisit plus d'un million de transactions de clients à l'heure et qui les télécharge dans une base de données comptant plus de trois pétaoctets (SAS 2102). C'est le Weather Channel, qui recueille 20 téraoctets de données au moyen de capteurs partout dans le monde, chaque jour (Henchen, 2013). C'est le nombre ahurissant de transactions de données qui sont générées chaque minute dans les médias sociaux et par les appareils intelligents interconnectés, comme les scanneurs et les technologies portables. En comparaison, les études de marché, les recherches sur l'opinion et les recherches sociales évoluent encore principalement dans un monde de gigaoctets. Nous ne sommes tout simplement pas habitués à travailler à l'échelle des mégadonnées.

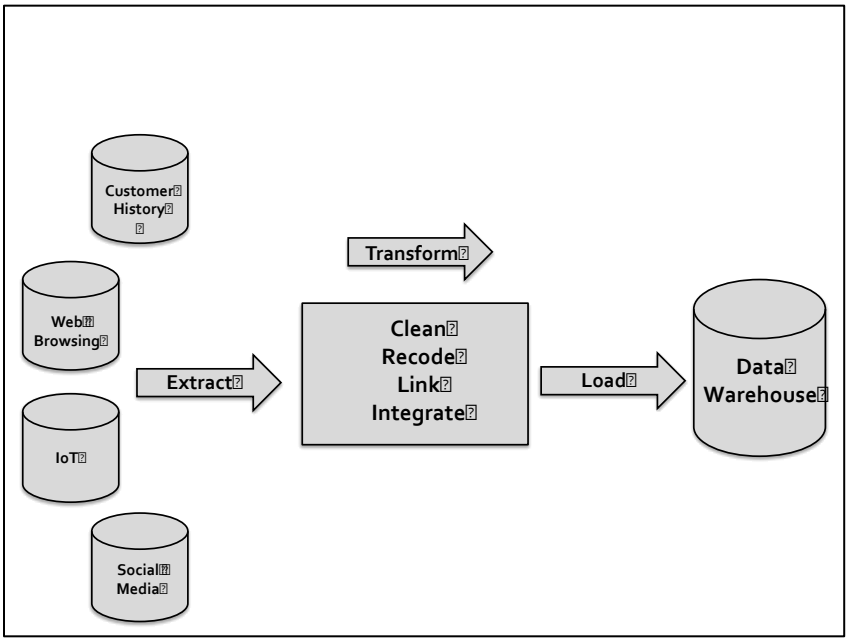
L'exploitation pertinente des mégadonnées nécessite aussi un investissement important dans les personnes, ainsi que dans la technologie. Les mégadonnées vont au-delà du recrutement d'un scientifique des données. Le rapport de l'AAPOR sur les mégadonnées (2015) comporte un résumé utile des compétences et des technologies nécessaires pour travailler à partir des mégadonnées. Même si le rapport ne quantifie pas l'investissement, celui-ci va probablement bien au-delà de ce que peuvent se permettre les organismes de recherche, sauf les très importants.

3.2 Qualité des données

Les chercheurs qui utilisent des mégadonnées soulignent souvent que leur qualité n'est pas celle à laquelle nous sommes habitués. Il arrive souvent que les données aient été recueillies pour une autre fin que la recherche, et que l'attention accordée à l'exactitude des éléments individuels, à leur complétude globale, à leur uniformité au fil du temps, à leur documentation complète, et même à leur signification, pose des défis graves en ce qui a trait à leur réutilisation. Les lecteurs qui sont familiers avec le modèle de l'erreur totale d'enquête (ETE) décrit par Groves (1989) reconnaîtront que les mégadonnées sont vulnérables à toutes les mêmes lacunes que les enquêtes, lacunes dans la couverture, données manquantes, mesures inappropriées, etc. La différence clé est que les chercheurs d'enquête, en théorie à tout le moins, conçoivent et contrôlent le processus d'élaboration des données d'une façon autre que les utilisateurs des mégadonnées.

Une part importante de la valeur des mégadonnées a trait à leur potentiel de fusionner plusieurs ensembles de données (p. ex., données de transactions de clients et données de médias sociaux ou données de l'IdO). Il s'agit d'un processus difficile et coûteux (voir la figure 2), et il s'agit aussi d'un point où des erreurs peuvent facilement se produire. Au cœur de ce processus de fusion figurent des bits de code informatique appelés ETC (extraire, transformer, charger), qui précisent les données qui sont extraites de bases de données de source, la façon dont elles sont vérifiées et transformées pour assurer la cohérence, puis fusionnées dans la base de données de sortie, habituellement un type d'entrepôt de données.

Figure 2
Aperçu simplifié du couplage de données



Prenons un moment pour examiner la difficulté de la spécification de l'ensemble de ces règles. Si vous avez déjà rédigé des spécifications de vérification pour un ensemble de données d'enquête, vous avez une petite idée de la difficulté. Tenons maintenant compte du fait que, dans une fusion de données à partir de plusieurs sources, vous pouvez avoir la même variable comportant différents codes; le même nom de variable utilisé pour mesurer des choses différentes; des règles différentes pour déterminer lorsqu'un élément est légitimement manquant et lorsqu'il ne l'est pas; des règles détaillées pour l'appariement d'un enregistrement d'une source de données avec un enregistrement d'une autre; des différences d'entités (clients, produits, magasins, coordonnées de GPS, gazouillis) qui doivent être résolues; etc. Il s'agit d'un travail difficile, laborieux, fastidieux et sujet à des erreurs. S'il n'est pas bien effectué, les conséquences sont désastreuses.

3.3 Analyses

Et il ne s'agit là que de la question des outils. La plupart des chercheurs d'enquêtes frappent habituellement un mur lorsqu'ils utilisent les mégadonnées. Il ne s'agit pas seulement de traiter des fichiers à l'échelle des pétaoctets. Il faut aussi tout un ensemble d'outils, qui dépendent à peu près tous d'un traitement en parallèle massif, ce qui va bien au-delà de ce que la plupart d'entre nous pouvons même concevoir.

Les responsables des études de marché, des recherches sur l'opinion et des recherches sociales, y compris ceux qui produisent des statistiques officielles, exécutent des travaux intéressants et utiles au moyen de ce que l'on peut décrire plus précisément comme des « données secondaires ». Il est légitime de se demander s'il s'agit réellement de mégadonnées. Et même si c'était le cas, la plupart n'ont pas encore saisi l'importance des nouvelles analyses requises pour exploiter réellement le potentiel des mégadonnées.

Prenons le célèbre éditorial de Chris Anderson publié dans *Wired*, en 2008, « The end of theory: The data deluge makes the scientific method obsolete ».

> « En présence de données massives, cette approche scientifique — hypothèse, modèle, test — devient désuète... Les pétaoctets nous permettent de dire que « la corrélation suffit ». Nous pouvons cesser de chercher des modèles. Nous pouvons analyser les données sans hypothèse en ce qui a trait à ce qu'elles

> pourraient démontrer. Nous pouvons mettre des chiffres dans les grappes de calcul les plus importantes ayant jamais existé et laisser les algorithmes statistiques trouver des tendances, lorsque la science ne le peut pas. »

Il s'agit d'un énoncé assez juste de la perspective de la science des données et de sa confiance à l'égard de l'apprentissage informatique — l'utilisation d'algorithmes permettant de trouver des tendances dans les données, sans être guidé par un ensemble d'hypothèses analytiques concernant les rapports entre les éléments de données. Pour paraphraser Vasant Dhar (2013), nous avons l'habitude de nous poser la question suivante : « Ces données correspondent-elles au modèle? ». Les scientifiques des données se posent plutôt la question suivante : « Quel est le modèle qui correspond à ces données? »

Les plans de recherche et les approches analytiques qui sont au cœur des recherches par enquêtes ont été élaborés à une époque où les données étaient rares et coûteuses et où les outils analytiques à notre disposition étaient faibles et n'avaient pas suffisamment de puissance. La combinaison des mégadonnées et de la technologie informatique en expansion rapide a modifié ce calcul.

Cela peut sembler une hérésie pour nombre d'entre nous des sciences sociales. Toutefois, il existe aussi un argument de longue date dans la profession statistique concernant la valeur de ces méthodes d'analyse algorithmique. Par exemple, en 2001, le distingué statisticien Leo Breiman décrivait deux cultures dans la profession statistique.

> « Certains présument que les données sont produites grâce à un modèle de données stochastiques établi. D'autres utilisent des modèles algorithmiques et traitent le mécanisme de données comme étant inconnu.... si notre objectif en tant que champ d'études est d'utiliser les données pour résoudre des problèmes, nous devons nous éloigner de la dépendance exclusive à l'égard de modèles de données et adopter un ensemble plus diversifié d'outils. »

On retrouve des arguments similaires parmi les statisticiens jusque dans les années 60 (voir, par exemple, Tukey, 1962).

Il existe évidemment des dangers, et les débats concernant la corrélation et la causalité (ainsi que l'endogénéité) doivent être pris au sérieux. Il existe même un site Web (http://www.tylervigen.com/spurious-correlations) consacré à certaines des conclusions les plus divertissantes, mais complètement erronées, que l'on peut tirer de corrélations qui font fi des règles. Toutefois, tout scientifique de données sérieux notera rapidement que ce type d'analyse exige davantage que de bonnes compétences en mathématiques, une puissance informatique massive et une bibliothèque d'algorithmes d'apprentissage automatique. Une connaissance du domaine et un jugement critique sont essentiels. Ou, comme nous le rappelle Nate Silver (2012) « les prédictions fondées sur des données peuvent réussir et elles peuvent échouer. C'est lorsque nous laissons de côté notre rôle dans le processus que les risques d'échec augmentent. »

3.4 Éthique

Deux des piliers les plus importants de la fondation éthique des études de marché, des recherches sur l'opinion et des recherches sociales sont le consentement et la confidentialité. Le consentement signifie que les participants possibles à la recherche reçoivent une description de l'objectif de la recherche et de la façon dont leurs données seront utilisées. La confidentialité signifie que le chercheur fournira le niveau de protection des données nécessaire pour s'assurer que l'identité d'un participant n'est jamais divulguée à un tiers sans le consentement explicite de ce participant, et que toutes les données diffusées pour l'analyse seront anonymisées.

Le monde en évolution des mégadonnées pose un problème sur ces deux fronts. Lorsque nous réutilisons les données recueillies à d'autres fins, nous avons l'obligation de déterminer si notre utilisation prévue est conforme aux modalités auxquelles les personnes ont convenu. En outre, les recherches récentes ont démontré que la somme de données disponible sur les personnes et la puissance de traitement pour les combiner et les analyser ont fait en sorte que les approches traditionnelles en matière d'anonymisation sont devenues désuètes (voir, p. ex., Sweeny, 2013; de Montjoye et coll., 2015). Pour une discussion plus approfondie, voir Lane et coll. (2014), AAPOR Report on Big Data (2015).

4. Mégadonnées et avenir des enquêtes

Les responsables des études de marché, de façon plus particulière, passent beaucoup de temps ces jours-ci à penser à l'avenir du grand secteur de la recherche et à faire des prévisions à ce sujet (voir, par exemple, Kaden et coll., 2012). Un consensus se forme concernant trois transitions qui, au fil du temps, transformeront ce que nous faisons.

Tout d'abord, il est largement reconnu que nous passons d'un monde où les données sont rares et coûteuses à un monde où elles sont abondantes et abordables. Il s'agit de la différence entre les enquêtes (coûteuses) et le monde créé par l'intégration de la technologie dans tous les aspects de notre vie au quotidien.

En deuxième lieu, la recherche sera moins axée sur la collecte de données au moyen de questions et davantage sur l'observation et l'écoute. L'acceptation croissante et généralisée des principes de la psychologie cognitive, comme la théorie à processus duaux (Kahneman, 2011), en a amené plusieurs à prétendre que le comportement est un prédicteur plus fiable des choix que les personnes font que les questions concernant les attitudes et les intentions, comme c'est le cas habituellement dans les enquêtes.

En troisième lieu, au fur et à mesure que les sources de données disponibles pour l'étude d'un sujet se multiplient, les chercheurs mettront l'accent sur la synthèse des données de plusieurs sources et méthodes par rapport à l'analyse d'un ensemble de données unique.

Dans ce contexte, les enquêtes ne deviennent que l'une des nombreuses façons d'effectuer de la recherche. Lorsqu'on a demandé à Joan Lewis, Consumer and Market Knowledge, chez Proctor and Gamble, si les médias sociaux remplaceraient les enquêtes, elle a répondu : « Nous devons être agnostiques en ce qui a trait à la méthodologie » (Neff, 2011). Elle voulait dire qu'il existe de nombreuses méthodes et sources de données qui peuvent être utilisées pour étudier un problème. Nous voyons déjà cela dans les statistiques officielles, les instituts nationaux partout dans le monde passant des recensements fondés sur des techniques de collecte de données d'enquête classiques à l'utilisation de données administratives.

La viabilité à long terme des enquêtes pourrait bien dépendre de leur caractère propre, ainsi que de leur capacité à cibler une population particulière, à préciser les données d'intérêt et à concevoir un processus de collecte qui réduit les erreurs. Il faut tirer parti au maximum des mégadonnées, et les problèmes de couverture, de signification et d'exactitude représentent une préoccupation constante. L'avenir des mégadonnées, et peut-être même de la profession d'enquête, dépend de la façon dont nous résoudrons ces problèmes.

Pour l'heure, le point de vue le plus courant semble être celui qui souligne la complémentarité des mégadonnées et des enquêtes (Forsyth et Boucher, 2015; Macer, 2015). Même si cela peut être réconfortant, cela ne nous libère pas de la responsabilité d'élargir notre compréhension des sources de données et des méthodes, au-delà des enquêtes. Pour utiliser l'analogie de Couper (2013), les enquêtes sont comme des tournevis comportant différents usages et capacités. Toutefois, elles ne sont pas les seuls outils dont nous disposons. Nous avons aussi des marteaux, pinces, clés et perceuses, qui sont chacun conçus pour une tâche que l'on ne peut pas accomplir avec un tournevis. Pour bien faire un travail, nous devons les utiliser en combinaison. Le défi pour notre profession est d'apprendre quand et comment utiliser les bons outils pour le travail à effectuer. Il pourrait s'agir d'un défi de taille pour ceux qui sont habitués à toujours opter pour les enquêtes.

Bibliographie

AAPOR (2015) AAPOR Report on Big Data. Retrieved on March 30, 2015 from http://www.aapor.org/AAPORKentico/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15.pdf

Anderson, C. (2008) The end of theory: The data deluge makes the scientific method obsolete. *Wired*. 16(7). Retrieved on June 19, 2014 from http://www.wired.com/science/discoveries/magazine/16-07/pbtheory.

Bogart, Leo. 1987. "The Future of Public Opinion: A Symposium," in *Public Opinion Quarterly*, Vol. 51, Supplement, pp. S173-S191.

Brieman, Leo. (2001). Statistical Modeling: The Two Cultures**.** *Statistical Sciences***.** 16(3) 199-231.

Couper, M.P. 2013. Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys. *Survey Research Methods* 7(3): 145-146.

de Montjoye, Y., Radelli, L., Singh, V.K., and Pentland, A. 2014. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*. 347(6221), 536-539.

Dhar, Vasant. 2013. Data Science and Prediction. *Communications of the ACM*. 56, 12, 64-73.

Dutcher, J. 2014. What is Big Data? Retrieved on June 14, 2015 from http://datascience.berkeley.edu/what-is-big-data/

Forsyth, J., and Boucher, L. (2015). "Why Big Data is Not Enough." *Research World*. 50, January/February, 26-27.

Goves, R. M. (1989). *Survey Errors and Survey Costs.* New York: Wiley and Sons.

Henschen, D. (2013) "Big Data Reshapes Weather Channel Predictions" *Information Week* Retrieved on June 21, 2015 from http://www.informationweek.com/big-data/software-platforms/big-data-reshapes-weather-channel-predictions/d/d-id/1112776

Kaden,, R.J., Linda, M. and Prince, M. (2012) eds. *Leading Edge Marketing Research: 21st Century Tools and Practices*, Washington, DC: Sage.

Kahneman, Daniel (2011) *Thinking Fast and Slow*. New York: Farrar, Straus, and Giroux.

Lane, J., Stodden, V., Bender, S. and Nissenbaum, H. (2014) *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, New York: Cambridge University Press.

Laney, Douglas (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety. Retrieved on February 5, 2015 from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Macer, Tim. (2015). "Big Data Plus Research Means More Accurate Results." *Research World*. 52, May/June, 13-15.

Neff, Jack. (2011). "Will Social Media Replace Surveys as a Research Tool?" *Advertising Age*. Retrieved on April 12, 2015 from http://adage.com/article/news/p-g-surveys-fade-consumers-reach-brands-social-media/149509/

SAS (2012). Big data meets big data analytics. Retrieved on December 28, 2104 from http://www.sas.com/resources/whitepaper/wp_46345.pdf.

Silver, N. (2012) *The Signal and the Noise*. New York: The Penguin Press.

Sweeny, L. 2013. Matching Known Patients to Health Records in Washington State Data. Retrieved on July 17, 2015 from http://dataprivacylab.org/projects/wa/1089-1.pdf.

Tukey, J.W. (1962). "The Future of Data Analysis." *The Annals of Mathematical Statistics*. 33(1): 1-67.

Ward, Jonathan Stuart, and Adam Barker (2013). Undefined by data: a survey of big data definitions. arXiv: 1309.5821vi. Retrieved on November 12, 2014 from http://arxiv.org/pdf/1309.5821v1.pdf.