Christine L. Borgman

Qu'est-ce que le travail scientifique des données ?
Big data, little data, no data

OpenEdition Press

Préface

DOI : 10.4000/books.oep.14717
Éditeur : OpenEdition Press
Lieu d'édition : OpenEdition Press
Année d'édition : 2020
Date de mise en ligne : 18 décembre 2020
Collection : Encyclopédie numérique
EAN électronique : 9791036565410

Préface

Les *big data* suscitent aujourd'hui un vif intérêt, mais les *little data* sont tout aussi essentielles à la recherche scientifique. Plus le volume des données augmente, plus notre capacité à examiner des constatations individuelles diminue. L'observateur doit sans cesse s'écarter des phénomènes qu'il étudie, il lui faut de nouveaux outils et perspectives. Néanmoins, les mégadonnées ne sont pas nécessairement de meilleures données. Plus l'observateur s'éloigne de son point de départ, plus on éprouvera de difficultés à déterminer la signification de ses observations : comment elles ont été recueillies, manipulées, réduites et transformées, avec quels présupposés et quels objectifs. Les scientifiques préfèrent souvent travailler sur un petit nombre de données que l'on peut inspecter de près. Parfois, elles et ils ne disposent d'aucune donnée, car ces dernières n'ont pas été découvertes ou ne peuvent pas l'être.

Les données de la recherche sont à la fois bien plus que de simples marchandises à exploiter et bien moins. Plans de gestion, critères de diffusion et autres politiques pleines de bonnes intentions que proposent les organismes de financement, les revues et les instituts de recherche tiennent rarement compte de la diversité des données et des pratiques disciplinaires. Rares sont les politiques qui s'efforcent de définir la notion de « donnée » autrement que par une liste d'exemples. Plus rares encore sont celles qui répondent aux motivations intrinsèques et extrinsèques divergentes des acteurs de la recherche. Les données peuvent revêtir simultanément des dimensions très variées selon les individus. Elles peuvent être un atout à maîtriser, à accumuler, à négocier, à combiner, à explorer et, éventuellement, à divulguer. Elles peuvent constituer une charge qu'il faut gérer, protéger ou détruire. Elles peuvent être sensibles ou confidentielles et leur diffusion peut comporter de grands risques. Leur valeur peut ne pas apparaître immédiatement ou n'être comprise que bien plus tard. Certaines données valent la peine d'être conservées indéfiniment tandis que d'autres n'ont qu'une utilité éphémère. En effet, en quelques heures ou quelques mois, certains types d'observations perdent leur intérêt sous l'effet des avancées technologiques et scientifiques.

Pour comprendre le rôle des données dans la recherche scientifique, il faut d'abord admettre qu'elles ne sont pas des objets naturels, possédant une essence propre. En réalité, elles sont des représentations d'observations, d'objets ou d'autres entités qui servent à mettre en évidence des phénomènes à des fins de recherche. Ces représentations varient au fil du temps et en fonction des scientifiques et des circonstances. Dans l'ensemble des sciences exactes, humaines et sociales, la communauté

scientifique crée des données, les utilise, les analyse et les interprète, souvent sans s'accorder sur ce qu'elles sont. Concevoir quelque chose comme une donnée est, en soi, un acte scientifique. La recherche repose sur la preuve, sur l'interprétation et sur l'argumentation. Les données sont ainsi un moyen, dont la fin est le plus souvent l'article de revue, l'ouvrage, la communication de colloque ou une autre production digne de reconnaissance universitaire. Le chercheur ou la chercheuse travaille rarement dans l'idée de les réutiliser.

Galilée émaillait ses carnets de croquis. Les astronomes du XIX^e^ siècle réalisaient des clichés sur des plaques de verre. Aujourd'hui, ces scientifiques captent des photons à l'aide d'instruments numériques. Si l'on peut mettre en correspondance des images du ciel étoilé photographiées par des appareils grand public avec celles prises par des missions spatiales, c'est parce que les astronomes ont convenu de représentations communes pour décrire et mapper les données. En effet, l'astronomie a investi massivement dans des normes, des outils et des archives afin de pouvoir rassembler des observations recueillies sur plusieurs siècles. L'infrastructure des connaissances astronomiques est cependant loin d'être achevée ou complètement automatisée. Les professionnels de l'information jouent un rôle clé dans l'organisation de l'accès aux données, qu'elles soient astronomiques ou autres.

Parce que les relations entre publications et données sont multiples, les données de la recherche gagnent à être examinées dans le cadre de la communication savante. Leur production peut se faire délibérément et sur le long terme, créant ainsi une mine de ressources dont la valeur va en augmentant. Elle peut se faire *ad hoc*, voire au petit bonheur, en se saisissant des indicateurs des phénomènes présents au moment de la collecte. Même si le protocole de recherche est parfaitement défini, qu'il s'agisse d'astronomie, de sociologie ou d'ethnographie, le recueil des données peut s'avérer aléatoire ; les découvertes faites à un stade influent sur le choix des données au stade suivant. Quel que soit le domaine, il faut, pour devenir scientifique, apprendre à évaluer les données, à déterminer leur fiabilité et leur validité et à s'adapter aux conditions de son laboratoire, de son terrain ou de ses archives. Les publications visant à exposer des découvertes situent celles-ci dans le contexte de la discipline, en se fondant sur l'expertise du lectorat. On y présente les informations utiles à la compréhension de l'argumentation, des méthodes et des conclusions. En revanche, les détails nécessaires à la reproduction de l'étude sont souvent omis, car on part du principe que la cible est familière des méthodes du domaine. Les notions de réplication et de reproductibilité, bien qu'elles constituent un argument courant en faveur de la publication de données, ne sont pertinentes que dans certaines disciplines et ardues à mettre en œuvre même en leur sein. Le plus difficile reste de déterminer quelles productions scientifiques méritent d'être préservées.

Les politiques de gestion, de diffusion et de partage des données occultent la complexité des rôles qu'elles occupent dans la recherche et ignorent largement la diversité des pratiques intradisciplinaires et interdisciplinaires. La notion de donnée varie grandement au sein des sciences exactes, humaines et sociales et à l'intérieur de chaque spécialité. Dans la plupart des disciplines, on apprend la gestion des données plus qu'on ne l'enseigne, ce qui conduit à recourir à des solutions *ad hoc*. Les chercheurs et chercheuses éprouvent d'ailleurs souvent de grandes difficultés à réutiliser leurs propres données. Les rendre exploitables par des inconnus est plus épineux encore. Le partage de données n'est la norme que dans de rares disciplines, car il est très délicat à mettre en œuvre et nécessite d'importants investissements dans l'infrastructure de la connaissance.

Le présent ouvrage est destiné au vaste lectorat que sont les parties prenantes des données de la recherche, parmi lesquelles les scientifiques, les chercheurs et chercheuses, les chefs et cheffes d'établissement, les organismes de financement, les maisons d'édition, les bibliothèques, les archives et les responsables politiques. La première partie dresse un panorama des données et de la recherche scientifique tout au long de quatre chapitres, conduisant à une analyse des notions de donnée, de recherche et d'infrastructure de la connaissance, ainsi que de la diversité des pratiques scientifiques. Dans la deuxième partie, nous examinons l'usage des données dans le cadre des sciences exactes, des sciences sociales et des sciences humaines à travers trois chapitres constitués d'études de cas, dont la structure parallèle permettra la comparaison. La dernière partie traite en trois chapitres des politiques et des pratiques en matière de données ; nous nous y penchons sur l'origine des nombreux écueils auxquels se heurte le travail scientifique des données, parmi lesquels la diffusion, le partage et la réutilisation, l'attribution, le crédit et la découverte, et, enfin, les données à conserver et pourquoi.

L'histoire de la recherche et celle des données sont étroitement liées depuis longtemps. Ni l'une ni l'autre ne sont des notions nouvelles. Ce qui est nouveau, c'est la volonté d'extraire les données des processus scientifiques pour les exploiter à d'autres fins. Les coûts, les bénéfices, les risques et les récompenses associés à l'utilisation des données de la recherche sont en train d'être redistribués entre des acteurs en concurrence. L'objet de ce livre est de susciter un débat bien plus riche et plus éclairé entre ces parties prenantes. C'est l'avenir de la recherche qui est en jeu.

Christine L. Borgman
Los Angeles, Californie
Mai 2014