Une littératie des données?

Jérôme Denis
Centre de sociologie de l'innovation— Mines ParisTech

Retours sur la troisième conférence «Data Literacy»
5 et 6 octobre 2018, Aix-en-Provence

En guise de reprise de cette journée passionnante, je voudrais revenir sur le titre même de la conférence et en particulier sur le terme de *littératie*, qui fait certes un peu jargonneux en français, mais qui est malgré tout utile, et qui me semble particulièrement adapté pour mettre en lumière une partie des enjeux contemporains autour des données, notamment dans les entreprises et dans les administrations. Pourquoi ? Tout simplement parce que, comme Jane Crofts, la première intervenante, l'a rappelé, les données sont une affaire de langage. Et plus encore, elles sont affaire d'écrit. Parler de littératie des données, c'est donc affirmer, ou rappeler, deux choses. Tout d'abord, que les données relèvent de la lecture, mais aussi — c'est un point crucial, comme l'ont montré Evelyn Münster et Christoph Nieberding dans leur présentation — de l'écriture. Manipuler des données, c'est lire et c'est écrire. Deuxième aspect : cette écriture et cette lecture ne vont pas de soi. Elles passent par un apprentissage. C'est d'ailleurs dans sa version « négative » que le terme littératie résonne en français, puisque la question de l'illettrisme est un phénomène identifié depuis longtemps, qui fait l'objet de politiques ambitieuses, organisées à l'échelle internationale.

Ce geste me semble aussi particulièrement important parce qu'il permet de résister à la tentation de présenter l'omniprésence de données aujourd'hui comme le résultat, ou le symptôme, d'une révolution, d'une rupture historique. Sur le plan anthropologique, on peut en effet rattacher les données à ce que Jack Goody a appelé la « raison graphique » à propos de la naissance des premières pratiques scripturales. C'est notamment ce qu'ont fait les historiens et les sociologues des sciences (au premier rang desquels Steven Shapin, Simon Schaffer, et Bruno Latour) en montrant que la production et la circulation des données scientifiques étaient tout entières composées d'opérations de lecture et d'écriture : de fabrique de traces qui se standardisent progressivement. De même, l'histoire des administrations et du travail dit « intellectuel » dans les entreprises (je pense aux travaux de Delphine Gardey JoAnn Yates) nous rappelle que l'invention des données est déjà ancienne, et qu'elle s'est jouée au fil d'un mouvement de rationalisation généralisée dans la reconfiguration des pratiques scripturales qui avaient cours depuis longtemps dans les bureaux.

À l'écoute des présentations de la journée, je vois également un deuxième intérêt à utiliser le terme de literacy, plutôt par exemple que celui de « culture des données ». En rattachant les données au domaine du langage, il permet en effet d'insister non seulement sur le fait qu'il y a un enjeu à maîtriser ce langage et ses subtilités, mais aussi que, en tant que langage, les données font partie des choses qui nous aident à penser le monde, à l'appréhender. Et, plus encore, si l'on accepte l'idée qu'aucun langage n'est neutre, que les mots que l'on utilise ne font pas que « représenter » ce qui

nous entoure, mais qu'il participent à faire exister d'une certaine manière la réalité dans laquelle nous vivons, on peut même comprendre, grâce à cette idée de littératie, les données comme des instruments de transformation du monde. Ce qu'a bien montré la présentation de Kim Albrecht à propos des enjeux complexes de la visualisation des données des sciences de l'univers.

Enfin, associer les données à la question du langage permet de rappeler l'importance de prendre compte le caractère situé, particulier, de tout jeu de données, comme l'ont fait la plupart des exposés de la journée. Comme les langues et leurs usages, les données sont ancrées. Elles prennent sens dans le milieu spécifique qui les engendrent. Autrement dit, parler de littératie est aussi un moyen de lutter contre toute tentation de voir dans les données une sorte de langage universel, transparent, univoque. Les données ne se produisent ni ne se comprennent « naturellement ». Elles ne circulent pas non plus sans friction d'un contexte d'usages à l'autre, comme Paul Edwards et ses collègues l'ont montré à propos de pratiques scientifiques dont on aurait pu penser pourtant qu'elles représentaient un environnement qui faciliterait l'échange sans couture de jeux de données. Même dans les domaines où l'on croit pouvoir se comprendre sans effort, la circulation de données, leur « ouverture », ne se font pas sans ajustements. C'est crucial dans les entreprises et dans les administrations où pullulent des jeux de données dites « métier » qui, s'ils sont manipulés sans aucun problème par celles et ceux qui les utilisent quotidiennement, restent hermétiques et difficiles à appréhender sans que soient mis en place des échanges et généralement des modifications importantes qui produisent, ou reproduisent, l'intelligibilité des données elles-mêmes. Passer d'un usage local de la « même » langue à un autre implique des précisions, des explicitations, et passer d'une langue à une autre nécessite des opérations de traduction. C'est de cela aussi qu'il est question lorsque l'on parle de littératie des données.

Revenons à la question de l'écriture. Il me semble qu'il y a un enjeu tout particulier aujourd'hui, dans le cadre des initiatives dédiées au à la littératie des données, à poser plus clairement encore la question de la production des données. C'est un point sur lequel Evelyn Münster et Christoph Nieberding ont notamment insisté en intégrant dans leur modèle non seulement ce qu'ils appellent les « décodeurs » des données, mais également les « codeurs ». Développer la *data literacy*, ça n'est pas seulement apprendre à tout le monde à lire correctement des données, c'est aussi interroger la fabrique même des données, les moments et les acteurs de leur *écriture*. Et c'est, dans le meilleur des cas, outiller et former à cette écriture. C'est d'autant plus important, que nous faisons face depuis plusieurs années à une rhétorique de la ressource naturelle, qui a cours aussi bien dans le domaine des données massives que dans celui des données ouvertes, qui laisse croire que les données sont des entités naturellement disponibles, qu'il suffirait de « libérer », de « récolter » puis de « traiter » pour produire de la connaissance.

N'importe quel responsable sérieux d'un programme de big data le sait, et la thèse de Samuel Goëta, l'un des deux fondateurs de datactivist, l'a bien montré à propos de l'open data : les données ne sont jamais à disposition dans les systèmes d'information, prêtes à l'emploi, attendant sagement d'être ramassées et rassemblées pour que leur puissance potentielle soit enfin libérée. Du point de vue de la littératie des données, cela implique au minimum de considérer deux principaux aspects : la prise au

sérieux des enjeux de la production des données, à contre-courant du vocabulaire de la collecte, et l'exposition dans les dispositifs de visualisation des données des conditions de cette production.

C'est sans doute en se penchant sur les cas des sciences citoyennes et de production contributive des données comme on a pu en voir plusieurs tout au long de la journée, tout en gardant en tête les leçons de l'histoire et de la sociologie des sciences, que l'on peut comprendre l'importance du premier point : toute « collecte » de données est une opération de production. Il ne s'agit jamais de cueillir des données, mais toujours de les générer. Si dans les entreprises et les administrations ces opérations sont, dans la plupart des situations, mal connues, voire expressément invisibilisées, ça n'est pas le cas dans les initiatives « participatives », qui au contraire prennent très au sérieux ce temps de la génération. Sans entrer trop dans les détails, on peut retenir que ces différents cas (autour de la cartographie avec OpenStreetMap, ou dans le domaine de la biodiversité, ou encore de la qualité de l'air pour ne citer que quelques exemples…) qu'ils assument que l'existence même de données repose sur un travail délicat, difficile, qui passe par un soin particulier, une attention à ce qui est fabriqué. Ils nous rappellent que, comme le proposait Bruno Latour il y a plus de vingt ans il est utile de ne pas toujours parler de « données », mais bien « d'obtenues ». On retrouve par ailleurs ici la question finalement classique de qualité des données. Question qui semble vite mise de côté par les prophètes de la données-pétrole ou de la donnée-déluge. Générer de belles données, des données pertinentes, intelligibles, a un coût, qu'il est problématique de négliger. En particulier si l'on se place du côté de celles et ceux qui demandent, voire exigent, qu'on leur transmette des données. Oublier que ces données doivent être obtenues, puis réajustées pour circuler, revient de fait à invisibiliser toutes ces opérations et à les reléguer au rang de sale boulot.

Dans le domaine des sciences citoyennes, on sait par ailleurs que ce travail n'a rien de neutre, qu'il est toujours en partie militant, voire explicitement politique. Le soin qui est consacré à la fabrique des données met dans certains cas clairement en lumière l'enjeu que représente la capacité de certaines données à transformer le monde, comme je l'évoquais plus haut, ou, *a minima*, à faire compter des choses qui, sans les données ne compteraient pas, ou pas tout à fait de la même manière : certaines espèces à un endroit particulier, le taux de présence de tel ou tel type de particules dans l'air, le niveau d'accessibilité, ou d'inaccessibilité, de tel quartier, etc.

Au cours de la journée, la question a été posée de savoir s'il était vraiment utile pour tout le monde de comprendre de quoi sont composés les jeux de données que l'on manipule, puisque finalement personne n'a vraiment besoin de savoir comment fonctionne une voiture pour s'en servir et se rendre d'un lieu à un autre. Prendre au sérieux la dimension générative et politique des données, dès les temps mêmes de leur fabrique, revient à réfuter très clairement la pertinence de cette métaphore. Les données ne sont pas des voitures. Elles font exister des pans entiers de notre réalité. Et si l'on en vient à les utiliser d'une manière ou d'une autre, on a tout à gagner à comprendre de quoi elles sont faites.

Ce qui m'amène à mon dernier point : comment peut-on faire progresser la littératie des données sur cet aspect ? Comment donner à comprendre ce qui se joue à même la fabrique des données ? Sans doute d'abord en passant par la pratique, l'un des principes fondateurs de la FING comme le rappelait Charles Nepotte en introduction. Je

pense en particulier à ce que Dorie Bruyas évoquait à propos des enfants qui comptaient leurs pas et qui en se confrontant à certains traitements des données qu'ils avaient produites réalisaient qu'il fallait vraisemblablement en générer de nouvelles, une fois que ce que chacun entendait par « un pas » aurait été plus clairement défini, au regard des résultats attendus. Mais on peut aussi poser cette question du côté de la visualisation des données ? Comment peut-on rendre visible une partie des conditions dans lesquelles les données mises en scène ont été produites ? Jusqu'où peut aller cette mise en lumière ? Peut-on par exemple donner à comprendre les limites connues qu'implique tel ou tel aspect de la production d'un jeu de données ? Comment donner à voir les incertitudes qu'il laisse ouvertes ?

Ce problème de l'appréhension des activités de production des données montre à quel point la question de la littératie peut être ambitieuse. L'histoire des sciences, tout comme celle des données administratives et commerciales, tend en effet à faire de l'effacement des conditions de leur production un gage de solidité des données. Comment peut-on sortir de cette logique de la boîte noire et valoriser les opérations d'inscription, de nettoyage, de formatage, de traduction qui ponctuent la génération des données ? Et comment reconnaître le rôle crucial des travailleuses et des travailleurs de la donnée ?