LA NUMÉRISATION DU MONDE

5/7 – PHILOSOPHIE DE LA DONNÉE

Selon les études de Bruno TEBOUL [1], la numérisation du monde a pour conséquence de placer les données numériques au centre des échanges économiques et sociaux. Nous produisons une quantité croissante de données, que des technologies de plus en plus sophistiquées permettent de faire circuler, de sécuriser, d'analyser. Une économie de la donnée tente de rendre compte de la valeur de ces opérations. Des politiques de la donnée sont mises en œuvre par les États et les grandes corporations. Un business se développe, notamment autour du Big Data. Mais la nature précise de la donnée reste floue. Une approche philosophique, comme celle menée par Luciano **FLORIDI**, peut en affiner la définition.

Qu'est-ce qu'une donnée ? L'avènement du numérique a amené à reprendre et préciser cette question longtemps négligée, qui émerge à la fin du XIX^{e} siècle, quand la philosophie se partage entre la prise en compte des « *données immédiates de la conscience* » (Bergson) et ***une méthode expérimentale privilégiant la mesure indirecte de données « objectives »***.

Le connu et l'inconnu

La philosophie des sciences, avec **QUINE**, mettra vite en doute cette prétention à l'objectivité.

Mais toute pensée pratique soucieuse de précision et de réalisme, d'un côté, toute approche scientifique, de l'autre, ne s'en appuie pas moins, jusqu'à aujourd'hui, sur des données.

Le terme même de « donnée » suggère un don qui est en réalité une contrainte : la donnée s'impose. C'est précisément ce qui lui permet de servir de socle à une réflexion partagée, de permettre le développement technique, de fonder des politiques publiques ou d'accroître les connaissances scientifiques. L'ingénieur, l'économiste, le physicien, le botaniste, l'agronome, le chimiste raisonnent données en main.

La donnée, dans ce contexte, c'est du connu, sur lequel on peut s'appuyer pour aller vers l'inconnu.

Le connu suppose un accord, une reconnaissance, bref un régime d'évidence ou de convention qui fait que l'ensemble des participants à une réflexion s'entendent sur ce qui est « donné ». On parle ainsi de données scientifiques, techniques, ou sociologiques : données intelligibles et valides, qui sont publiques ou à tout le moins partageables au sein d'une communauté d'utilisateurs qui reconnaissent leur valeur.

Dans cette perspective les données quantitatives ont un avantage indéniable sur les données qualitatives. Il est plus facile de s'entendre sur les nombres que sur les qualités. Aussi une bonne partie de la science et de l'ingénierie modernes cherche-elle à quantifier le qualitatif, à le décomposer en chiffres. Entre le noir et le blanc, il n'y a pas le gris, mais du noir à 25 %, à 70 %, etc. Une image se décompose en pixels, et chaque pixel peut se voir indexer sur l'axe qui va des infrarouges aux ultraviolets.

La numérisation du monde, ainsi, ne commence pas au « numérique », à la traduction de signaux en séries de 0 et de 1. Elle caractérise la vaste traduction du monde sensible – et en son sein de l'humanité – en séries de données, traduction entamée au début de l'ère moderne et qui connaît aujourd'hui une accélération vertigineuse.

Mais cette accélération porte en elle une seconde évolution, celle qui passe de l'analyse à la stochastique, de la réduction précise et rigoureuse des mécanismes à la révélation *ex post*, au moyen du datamining, de lois statistiques. Cette révolution, qui vient de commencer, pourrait se caractériser comme le triomphe de la pensée inductive sur la pensée déductive. Dans des domaines de plus en plus nombreux, la connaissance est produite à partir des corrélations extraites de grandes masses de données. Il s'agit moins de prouver que de voir apparaître des lois.

La statistique et l'algorithmique s'affirment ainsi désormais comme les outils fondamentaux de la connaissance, mais aussi de la décision.

Cette révolution, dans laquelle nous sommes plongés, oblige à interroger le statut de ce qui la nourrit, de ces *data* innombrables, stockées dans de gigantesques centres de stockage. En commençant par des questions aussi simples que radicales, comme celle-ci : les données sont-elles de l'information ?

Atomes d'information

La distinction importe, et elle s'inscrit dans une chaîne allant des faits au savoir.

Sven Ove **HANSSON**, professeur de philosophie à l'Institut royal de technologie de Stockholm, résume dans un article de 2002 le jeu des différences entre données, information et savoir : « Les données diffèrent de l'information en ce qu'elles n'ont pas à se présenter sous une forme qui se prête à l'assimilation. Si au lieu de l'ouvrage [de sociologie que je suis en train de lire], j'avais sur mon bureau les dix mille questionnaires sur lesquels il repose, j'aurais des données au lieu d'information. En résumé, il faut que des données soient assimilables pour pouvoir constituer de l'information et qu'elles soient assimilées pour pouvoir constituer du savoir. »

HANSSON reprend une distinction déjà faite par Roger BOHN dans un article de la Sloan Review of Management (1994) entre donnée, information et connaissance.

– Les données sont des éléments provenant des capteurs, elles sont relatives au niveau mesuré d'une variable quelconque.

– L'information consiste en des données organisées dans une structure donnée et qui, placée dans un contexte, est dotée de sens.

– La connaissance va plus loin : elle permet de faire des prédictions, d'établir des liens de causalité ou de prendre des décisions. Ce qui a de la valeur, c'est la connaissance. Mais comme le note **BOHN** l'information est plus facile à stocker, à décrire et à manipuler. La même chose est-elle vraie des données ? En termes numériques, oui : une donnée serait en quelque sorte un atome d'information, une mesure minimale, à un instant et en un point

de l'espace. Bref, quelque chose qui peut se réduire à un 0 ou un 1.

En termes philosophiques, la donnée est également plus facile à caractériser que l'information. C'est un concept plus simple, moins glissant. La donnée serait la traduction la plus immédiate, la plus brute, d'un fait. Elle n'est pas le fait, mais l'unité minimale d'observation qui permet de le caractériser.

Il serait illusoire de prétendre à son objectivité, et de soutenir qu'il n'y a pas d'intention ni de projet dans la donnée. La mesure, en elle-même, procède d'une discrimination entre toutes les données mesurables d'un phénomène : vous choisissez de mesurer telle variable plutôt que telle autre, et par ce filtre vous définissez ainsi une réalité, celle, en quelque sorte, que vous avez besoin de connaître.

Mais dans le cas de la donnée, typiquement produite par un capteur, une apparence d'objectivité est retrouvée, par deux voies : la très faible quantité d'information contenue dans la donnée, et la présence d'autres capteurs (qui permettent de construire une représentation plus riche du phénomène observé, comme par exemple l'état du pneumatique avant droit de votre voiture : la chaleur, les vibrations, la pression de l'air, l'âge du pneumatique, la durée d'utilisation aujourd'hui, permettent à votre ordinateur de bord de construire une information extrêmement fiable).

Une définition sémantique de la donnée

Les considérations qui précèdent peuvent être renversées. D'un côté, elles aboutissent à assumer la nécessité d'opérer des choix, de filtrer les données et ainsi de reconstruire une représentation très réductrice de la réalité.

D'un autre côté, la recherche d'objectivité, la multiplication des captures de données et la croissance exponentielle de la masse de données recueillies ouvrent vers ce fantasme scientifique d'une représentation complète des phénomènes, d'une numérisation absolue du monde. Version scientifique : quand on étudie telle pathologie osseuse, on mobilise 100 000 ensembles de données très complètes provenant de 100 000 patients différents, et on se donne ainsi une chance inédite de comprendre un phénomène, ou tout au moins de tout enregistrer, de ne rien laisser de côté. Version quotidienne, c'est l'homme connecté : votre pression sanguine est analysée chaque seconde par des capteurs, votre position dans l'espace est captée en permanence, etc. Vous vous transformez en un producteur de masses de données toujours plus abondantes. Avec à la clé une interrogation sur ce que peuvent valoir ces données. En d'autres termes, qu'est-ce qui peut leur permettre d'entrer dans la chaîne qui va des faits au savoir, en passant par l'information ?

Une approche de la donnée comme atome d'information trouve ici ses limites, car elle ne dit rien de ce passage. Luciano **FLORIDI**, professeur de philosophie et directeur de recherche à l'*Oxford Internet Institute*, propose une réflexion qui permet de dépasser cette limite.

Il s'interroge sur la possibilité de fonder une théorie de l'information sur la donnée (« a data-based definition of information »). En d'autres termes, de définir sémantiquement la donnée, en se demandant ce qui lui permet de produire de l'information.

Il retient d'abord une définition rigoureuse de la donnée : « *une donnée est un fait supposé qui procède d'une différence ou d'un manque d'uniformité dans un contexte* » (*a datum is a putative fact regarding some difference or lack of uniformity within some context*).

Cette diaphore, cette différence dans le tissu du réel, ouvre sur la possibilité d'une

information, mais à certaines conditions. **FLORIDI** identifie trois réquisits. Il faudrait

– a) une ou plusieurs données ;

– b) que ces données soient bien formées (« *well-formed* »), c'est-à-dire assemblées selon certaines règles ;

– c) et qu'elles soient porteuses de sens (« *meaningful* »), c'est-à-dire à même d'être interprétées ou, si l'on préfère, aptes à être traduites, ou exprimées autrement.

Il s'ensuit, et c'est là le point essentiel, que la donnée peut se définir comme une entité relationnelle (*relational entity*).

L'éclairage théorique de **FLORIDI** permet de comprendre ce point, en faisant surgir toute la portée des notions de « différence » et de « manque d'uniformité ». Ces deux notions renvoient, dit-il, à ce que les Grecs nommaient « *diaphora* », un écart. **FLORIDI** poursuit en proposant une « définition diaphorique de la donnée » qui peut être appliquée suivant trois niveaux.

– Il distingue d'abord la donnée comme « *diaphora de re* », c'est-à-dire comme manque d'uniformité dans le monde réel. Il n'existe pas de nom spécifique pour de telles « données dans la nature ». Une suggestion possible est de se référer aux données comme des « *dédomen* » (traduction en grec ancien de « *données* »). On peut noter d'ailleurs d'un point étymologique, que le mot « datum » est apparu en latin à partir de la traduction d'une œuvre d'Euclide, *Dedomena*. On ne peut connaître directement la donnée, mais simplement l'inférer à partir de l'expérience. **FLORIDI** explique que les « *Dedomena* » sont des données pures ou des données proto-épistémiques, c'est-à-dire des données avant qu'elles ne soient épistémiquement interprétées. En tant que « *Fractures dans la fabrique de l'être* », elles ne sont jamais accessibles ni élaborés indépendamment d'un certain niveau d'abstraction. Elles ne sont pas épistémiquement expérimentées, mais leur présence est empiriquement déduite de (et requise par) l'expérience.

– À ces données proto-épistémiques s'ajoute la donnée comme « diaphora *de signo* », *c'est-à-dire le manque d'uniformité (ou la perception d'un manque d'uniformité) entre deux états physiques, comme le niveau plus ou moins élevé d'une batterie, un signal électrique dans une conversation téléphonique, ou un point dans l'alphabet morse.*

– *Vient enfin la «* diaphora *de dicto »*, c'est-à-dire le manque d'uniformité entre deux symboles, par exemple les lettres A et B dans l'alphabet latin. La notion centrale de diaphora, qui réunit ces trois versions de la donnée, renvoie à une divergence, un moment où quelque chose se met à différer, une différence qui appelle un sens. ***La donnée est l'entité symbolique qui code cette différence***. Elle est le lien entre cet écart, qu'on serait tenté de dire insignifiant, et le sens. Le point de passage de l'insignifiant au signifiant.

(à suivre)

[1] Senior VP, Science & Innovation, Groupe Keyrus, avril 2017.

Auteur de « La donnée n'est pas donnée - Stratégie & Big Data », de juin 2016.