Rob Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*

Frédéric Clavert

Docteur en histoire contemporaine, Frédéric Clavert a étudié les sciences politiques et l'histoire contemporaine à Strasbourg et à Leeds. Ses recherches s'orientent aujourd'hui vers les relations entre les banquiers centraux et la construction européenne d'une part, sur les sources de l'historien.ne à l'ère numérique d'autre part. Il a été titulaire d'une bourse du DAAD et a obtenu l'aide à la publication du Prix Pierre Grappin. Après cinq ans comme chercheur (Histoire de l'intégration européenne et Humanités numériques) au Centre virtuel de la connaissance sur l'Europe (Luxembourg), il est désormais ingénieur de recherche pour le LabEx "Ecrire une Histoire Nouvelle de l'Europe"

23/12/2014

Décrire la « révolution des données » avec rigueur et esprit critique, loin du bruit qui se forme autour du sujet, tel est le but de cet ouvrage de Rob Kitchin, professeur à l'université nationale irlandaise Maynooth. Spécialiste de géographie urbaine, de ce que l'on appelle la « ville connectée », l'auteur en est venu à s'intéresser aux infrastructures et pratiques numériques et aux notions liées à celle de « donnée », dont il donne ici un panorama documenté et réaliste.

Par souci de rigueur, l'auteur s'attache dans son premier chapitre à définir précisément ce qu'est une donnée : « the raw material produced by abstracting the world into categories, measures and other representational forms […] that constitute the building blocks from which information and knowledge are created » (p. 1). Cette introduction est aussi exemplaire de l'un des soucis de l'auteur, rappelé à la fin de chaque chapitre et dans la conclusion du livre : mettre en lumière les recherches qui devraient être faites et qui ne le sont pas encore ou trop peu sur le sujet, particulièrement sur ce qui touche l'ontologie et l'épistémologie de la donnée et de ce qui l'accompagne - les grands assemblages de données, les infrastructures, l'open data et ses discours.

Après un passage en revue et une définition critique des différents types de données, l'auteur évoque la notion d'assemblages sociotechniques complexes de données, dont il analyse dans ce livre différents types : les données ouvertes (*open data*), les infrastructures de données (*data infrastructures*) et le *big data* (les données massives). Cette analyse repose sur trois éléments majeurs : le besoin de développer des moyens conceptuels et philosophiques pour donner du sens et interpréter les données ; la prise de conscience qu'une révolution des données est en cours, moment-clé dans l'évolution et la mutation des assemblages de données ; le besoin urgent, au vu des problématiques techniques, éthiques et scientifiques soulevées par cette révolution, de développer notre compréhension des

assemblages de données en train d'émerger. Sur cette base, le livre, divisé en tout en onze chapitres, peut être vu comme structuré en cinq parties : la première (chapitre 2) sur les *small data* et leurs infrastructures, la seconde sur la notion d'*open and linked data* (données ouvertes et reliées) (chapitre 3), la troisième, nettement plus longue, autour de la notion de *big data* (chapitres 4 à 6), la quatrième sur le discours englobant le *big data* (chapitre 7 à 9) et la dernière sur les conséquences de la révolution des données ainsi décrite (chapitres 10 et 11).

Le message que veut faire passer Rob Kitchin sur les petits ensembles de données fondés sur l'échantillonnage – les défenseur du *big data* revendiquant l'exhaustivité – est qu'ils sont toujours pertinents, bien que limités en volumes et statiques, car ils restent efficaces pour répondre à des questions précises de recherche nécessitant des données de très bonne qualité. Néanmoins, la recherche fondée sur ces petits ensembles de données gagne à les fondre dans des infrastructures qui permettront de les conserver sur le long terme, afin de réaliser des économies d'échelle, de mettre ces données à disposition d'autres utilisateurs et de les associer à des outils et des ressources pédagogiques autorisant à en faire des usages nouveaux. Cela explique l'émergence de nombreuses infrastructures de données dans le monde académique. L'ampleur de ce changement d'échelle, peu étudié, est illustrée par le nombre croissants de « courtiers » (*brokers*) en données et l'explosion du secteur de l'analyse des données (*analytics*).

Avec l'émergence de ces infrastructures vient un mouvement qui lui est plus ou moins relié : l'*open data* (les données ouvertes). Ce mouvement souhaite encourager la ré-utilisation des données par leur ouverture, leur mise en réseau (*linked data*, qui a pour but de transformer le web en données reliées entre elles) et le développement d'outils pour les traiter, à des fins à première vue peu contestables : transparence, innovation, productivité, etc. Or, l'auteur relève que le coût économique des données ouvertes est élevé et leur modèle économique fondé pour le moment presque uniquement sur fonds publics. De plus, peu de recherche portent sur les projets de données ouvertes entendus comme systèmes socio-techniques complexes où interviennent des acteurs dont les buts sont contradictoires. Ainsi, notre compréhension des avantages et inconvénients de ces projets est faible.

L'auteur s'attache ensuite, logiquement, à définir le *big data.* Récente, la notion de « données massives » n'est pas encore clairement définie, faisant l'objet de débats intenses. Outre le volume, d'autres caractéristiques en font des données différentes des autres : vélocité, variété, exhaustivité, résolution, relations, flexibilité notamment. Pour mieux étudier la nature des données massives, une taxonomie fiable devrait être établie, qui pourrait aussi insister sur d'autres facteurs (qualité et provenance par exemple). Comment peut-on aujourd'hui accumuler autant de données ? Pour l'auteur, de grandes innovations techniques (la capacité de calcul, le réseau, l'identification des sources de données et la capacité distribuée à stocker les données), ont permis l'émergence de différents systèmes socio-techniques de production des données massives.

Une fois les données stockées, elles doivent être analysées, ce qui est le but de leur collecte : leur donner sens et valeur. En gestation depuis longtemps, des techniques de traitement et d'analyse des données se démocratisent, après une forte évolution – qui se poursuit – liée au *big data* dont la vélocité (leur évolution constante) et le volume ont remis en cause les méthodes traditionnelles. Ces enjeux ne sont pas uniquement humains (l'adaptation des chercheurs aux nouvelles techniques) et techniques : l'analyse des données est le reflet d'une épistémologie particulière.

Au fil des chapitres consacrés au *big data*, Rob Kitchin s'interroge sur le manque de recherche et de compréhension des données massives. Quelles en sont les implications pour les sociétés, les gouvernements, le secteur commercial ? Favoriseront-elles le bien public ou les intérêts privés ? Quel est leur mode de production et leur histoire ? Comment déconstruire le discours qui les entourent ? Les données massives et leur analyse insufflent de nouveaux modes de production de la connaissance du monde. Quel en sera en retour l'impact sur le monde lui-même ?

Ces questions trouvent pour partie réponses dans les chapitre 7 à 10 qui analysent les conséquences de la révolution des données dans les domaines public et privé (chapitre 7), de la recherche (chapitre 8) puis se penchent sur les conséquences organisationnelles (chapitre 9), éthiques, politiques, sociales et légales (chapitre 10).

Analysant le discours en faveur de l'adoption du *big data* par les gouvernements et le secteur privé, Rob Kitchin constate qu'il oblitère les conséquences potentiellement négatives de cette adoption sur les libertés civiles, la sécurité des données, le *social sorting* (l'analyse automatique des données personnelles) ou la gouvernance par anticipation. Ces arguments positifs sont d'abord développés par des entreprises vendant des solutions d'analyse de données et par des gouvernements d'idéologie néo-libérale. L'auteur attire également l'attention sur le fait que les solutions fondées sur les données aboutissent rarement aux résultats escomptées, engendrant de nouvelles problématiques, peu étudiées, liées à une sorte de subversion des métriques, ces dernières poussant les acteurs à modifier leur comportement.

Dans le cas de la recherche, Rob Kitchin estime que la révolution des données engendre une transformation épistémologique des sciences - « dures », sociales ou humaines. Ces transformations encouragent de nouvelles approches pour l'analyse et la création des données, ce qui permet de poser et de répondre à des questions d'une nouvelle manière. Pour autant, mieux vaut parler de recadrage des sciences, plus que de subversion par les données, surtout dans le domaine des SHS. La diversité des fondements philosophiques de ces dernières empêchera la venue d'un nouveau paradigme : les *big data* viendront plutôt élargir leur socle épistémologique que le transformer pleinement. Il n'en reste pas moins qu'une réflexion sur les conséquences de la révolution des données est nécessaire en sciences humaines et sociales comme dans les science « dures ».

Les *big data* n'ont pas que des conséquences épistémologiques. Elles ont aussi des suites techniques et surtout organisationnelles. Parmi les éléments évoqués par l'auteur, nous mentionnerons le fait que les organisations ont du mal à s'adapter, par manque de savoir-faire notamment. Les initiatives prises alors dans le domaine des données manquent souvent de maturité. L'une des grandes difficultés est de former les chercheurs, d'adapter leurs compétences. Dans les domaines éthiques, politiques, sociaux et légaux, si la création, l'intégration et l'utilisation des données a des bénéfices incontestables pour les gouvernements, entreprises et citoyens, elles apportent également des conséquences négatives. Analyser ces dernières relève d'une mission ardue, alors que le besoin est pressant, notamment dans le domaine des libertés publiques et de la vie privée. Le monde de la recherche devrait encourager en conséquence des études sur les questions éthiques, juridiques politiques et sociales et, plus particulièrement, des recherches empiriques sur les effets discursifs et matériels de l'emploi des données.

Pour conclure, faisons nôtre les derniers mots de l'auteur : « For too long data and the constitution and operation of the assemblages surrounding them have been taken for granted, with attention focused on the information and knowledge distilled from them. It is time to rectify this neglect. » (p. 192).