LA RÉVOLUTION DES DONNÉES EST-ELLE EN MARCHE ?

Implications pour la statistique publique et la démocratie

Thomas Roca et Emmanuel Letouzé

De Boeck Supérieur | « Afrique contemporaine »

2016/2 n° 258 | pages 95 à 111

ISSN 0002-0478
ISBN 9782807390072

La révolution des données est-elle en marche ?
Implications pour la statistique publique et la démocratie

Thomas Roca
Emmanuel Letouzé

« Bien informés, les hommes sont des citoyens ; mal informés, ils deviennent des sujets », écrivait Alfred Sauvy. Avec seulement 0,5 % de l'aide publique au développement (APD) affectée au soutien à la production de statistiques, la statistique officielle fait figure de parent pauvre de l'aide internationale. L'Afrique souffre en particulier d'un déficit de données socio-économiques et démographiques qualifié de « tragédie statistique » par certains experts. Les attentes et les possibilités sont grandes à l'ère de la *dataification* – ou mise en données – de nos vies et de nos sociétés, fruit et facteur de la transition digitale.

Mots clés : Statistique – Aide publique au développement – Dataification – *Big data* – *Open data*

Statisticien, le job le plus sexy de la décennie[1] ? Il n'est pas sûr que cet avis soit partagé par les instituts nationaux de la statistique (INS)[2] de par le monde, particulièrement dans les pays les moins développés, en Afrique notamment, qui affrontent une pénurie de moyens, au moment même où ils font l'objet de demandes et d'attentes croissantes. Le nombre de cibles (169) retenues pour le suivi des Objectifs du développement durable (ODD), par exemple, paraît hors de proportion au regard de leurs capacités. Lors des débats de la commission « Statistiques » des Nations unies, certains représentants des INS, du Sud comme du Nord, ont souligné que seules 29 % de ces cibles seraient mesurables dans le contexte actuel (Stevance, 2015), alors que moins de 0,5 % de l'aide publique au développement est affectée au soutien à la statistique publique (Paris21, 2015). L'Afrique en particulier souffre d'un déficit de données socio-économiques et démographiques qualifié de « tragédie

Thomas Roca est économiste et statisticien à l'Agence française de développement (AFD). Il développe un programme de recherche couvrant la mesure du bien-être et du développement humain et les indicateurs alternatifs de développement et *big data* (rocat@afd.fr). **Emmanuel Letouzé** est directeur de Data-Pop Alliance et chercheur associé au MIT Media Lab et au Harvard Humanitarian Initiative, où il travaille sur les applications et implications du *big data* pour le développement humain, notamment les crises et conflits.

statistique » par certains experts. Pourtant, les attentes et les besoins sont immenses à l'ère de la *dataification* – ou mise en données – de nos vies et de nos sociétés, fruit et facteur de la transition digitale[3].

En effet, le nouveau paysage des données a été révolutionné par l'explosion, depuis moins d'une dizaine d'années, de la quantité et de la diversité des données générées par l'utilisation de services et appareils digitaux – les « données massives » –, ainsi que par les améliorations concomitantes de leurs capacités de stockage et de traitement. Le phénomène *big data*, et plus largement la révolution des données en cours et appelée de leurs vœux par les Nations unies, est amené à transformer la statistique publique au XXIe siècle. Loin de la rendre obsolète, la révolution des données, si elle veut contribuer à l'accélération du développement et au raffermissement de la démocratie en Afrique, requiert de renforcer et réinventer le rôle central de la statistique publique dans les futures sociétés « de la connaissance et de la transparence ».

Si les promesses sont nombreuses, il ne faut pas imaginer que les indicateurs produits à partir des données massives viendront combler l'ensemble des déficits de données[4] ou que la production de statistiques plus fiables et plus fréquentes suffira à améliorer les phénomènes mesurés : pauvreté, inégalité, mortalité infantile et maternelle, illettrisme, dégradation environnementale, etc. Tout d'abord, les données massives (voir encadré 2) permettront de compléter, à la marge, l'offre du système statistique[5] en même temps qu'elles poseront des questions et des défis en termes d'équilibre des pouvoirs et de respect de la vie privée. Mais surtout, ou justement, notre conviction est qu'elles doivent constituer le point d'entrée et d'accroche d'un dialogue et d'un contrat social renouvelé, avec pour outils et objectifs une participation et une évaluation citoyennes renforcées. Pour cela, au XXIe siècle, le citoyen devra « être familier » des données.

De la stratégie statistique à la révolution des données

Pour mieux éclairer ces évolutions et perspectives, un rappel historique est utile. En Afrique, la fin des années 1980 a été marquée par les politiques libérales d'ajustement structurel et, avec elles, le déclin de la planification et des politiques industrielles. Les contraintes budgétaires liées aux politiques de désendettement et l'abandon des politiques économiques volontaristes ont fait

1. « I keep saying the sexy job in the next ten years will be statisticians » (Hal Varian, Google's Chief Economist, *New York Times*, 2009).
2. La dénomination et le statut, notamment son degré d'indépendance, de l'agence ou service en charge de la coordination du système statistique public – tel que l'Insee en France – varient selon les pays ; un acronyme largement utilisé est aussi Office national de la ou des statistiques (ONS). Nous utilisons INS de manière systématique.
3. Selon certaines sources, difficiles à vérifier, la production de données digitales de ces deux dernières années avoisine le stock accumulé depuis l'essor de l'humanité : « 90 % of the data currently in the world was created in the last two years » (House of Commons, 2016). Pour une analyse plus complète du volume de données produites et stockées dans le monde, voir Hilbert (2015).
4. En anglais, on parle de *data gaps*.
5. Nous considérons ici seulement les statistiques officielles. Dans les faits, les données massives, ou *big data*, produisent une quantité phénoménale d'information sur les individus et leurs comportements, en dehors du champ de la statistique officielle.

sombrer la demande pour les statistiques officielles et les budgets alloués à l'administration en général (Razafindrakoto, Roubaud, 2003).

Les pays de l'OCDE (Organisation pour la coopération et le développement économiques) ont aussi connu des crises, qui ont eu un double effet. D'une part, une contraction de l'APD pour nombre d'entre eux, et, d'autre part, un renforcement de la demande de redevabilité et d'indicateurs chiffrés, d'évaluations des politiques publiques, etc., qui ont mobilisé les statisticiens des pays en développement, au détriment de la production des statistiques officielles. En revanche, d'autres initiatives soutenues par des pays de l'OCDE, tels que le Demographic and Health Survey Program, ont eu des conséquences bénéfiques sur l'offre de statistiques sociodémographiques.

Les Objectifs du millénaire pour le développement (OMD) annoncés au tournant du siècle ont eu un effet mitigé. Ils ont contribué à mettre le rôle de la mesure – notamment de la pauvreté – au centre de l'agenda politique international. Mais les efforts mis en œuvre autour de la mesure des OMD n'ont pas permis un suivi efficace de l'ensemble des cibles retenues, notamment dans les pays d'Afrique subsaharienne, qu'il s'agisse de la fréquence des enquêtes sur le niveau de vie des ménages ou encore de la fiabilité des données de scolarisation. D'importants écarts ont été relevés entre les données issues des enquêtes administratives et des enquêtes ménages. Par ailleurs, l'ampleur des demandes soumises à des systèmes statistiques publiques affaiblis en Afrique par l'austérité des années 1990 a créé un effet d'éviction en défaveur de la production des comptes nationaux, servant de base au calcul du PIB.

D'autres facteurs, tels que la fuite de cerveaux, voire la fin de la guerre froide et le démantèlement de l'Union soviétique qui fournissait un appui à ces pays alliés, ont également contribué à l'affaiblissement de la statistique publique en Afrique. Tels sont les éléments et ingrédients ayant mené à la « tragédie statistique africaine » décrite par des experts comme Shanta Devarajan (2011) ou Morten Jerven (2013), en référence à la « tragédie de la croissance » que connut le continent dans les années 1990.

Le phénomène prend plusieurs formes. Récemment, on découvrit que les PIB – qui est pourtant l'indicateur le plus observé et commenté – du Ghana, du Nigeria et, dans une moindre mesure, du Kenya, avaient été largement sous-estimés sur plusieurs années. À la suite de l'actualisation de la richesse produite pour mieux rendre compte de la montée en puissance du secteur technologique, le PIB du Ghana gagnait 60 % et près de 90 % pour celui du Nigeria.

Lorsqu'il s'agit de la mesure de la pauvreté en Afrique, premier Objectif du millénaire pour le développement, on se heurte à une pénurie ou « sécheresse » de données : pour un tiers des pays d'Afrique subsaharienne, les chiffres les plus récents, issus d'enquêtes socio-économiques, datent d'au moins sept ans. Bien souvent, ces données manquent de granularité – temporelle, géographique et sociale – par âge, sexe, etc.

Ce constat a joué un rôle essentiel dans l'appel de l'ONU pour un nouveau pacte mondial, une révolution des données, au moment où étaient discutés

et définis les Objectifs du développement durable (ODD). L'expression « révolution des données » fut mentionnée pour la première fois dans un rapport rédigé par un panel de vingt experts internationaux en 2013[6] ; puis, ses termes et contours précisés dans un deuxième rapport publié en novembre 2014[7]. Avec des informations plus fiables, plus fréquentes, plus granulaires, il promet des politiques publiques plus efficaces, plus ciblées, plus agiles, et, *in fine*, plus à même de répondre aux besoins des populations. Selon le résumé du rapport : « Alors que le monde s'engage dans un projet ambitieux pour atteindre les nouveaux Objectifs de développement durable (ODD), il est urgent de mobiliser la révolution des données pour tous et pour la planète entière afin de poursuivre les progrès réalisés, de responsabiliser les gouvernements et de favoriser le développement durable. Une information plus diversifiée, intégrée, opportune et digne de confiance peut mener à une meilleure prise de décision et à une implication en temps réel des citoyens. Cela permet aux individus, aux institutions publiques et privées et aux entreprises de faire des choix qui sont bons pour eux et pour le monde dans lequel ils vivent[8]. » Mais qu'entend-on par « révolution des données » et surtout qu'attend-on d'elle ?

La révolution des données : piliers, promesses et problèmes

La révolution des données est nourrie de deux phénomènes ou mouvements principaux, distincts mais liés de façon croissante : l'*open data*, ou « données ouvertes », et le *big data*, imparfaitement traduit par « données massives ».

Des données ouvertes pour le développement. La montée en puissance du mouvement des données ouvertes est en partie liée aux crises économiques et aux réductions budgétaires, qui ont encouragé une surveillance renforcée des représentations nationales sur les montants et l'efficacité de la dépense publique et plus largement alimenté une demande de transparence et de redevabilité, nécessitant la production et la mise à disposition de données relatives à l'action publique. L'émergence des réseaux sociaux et les changements technologiques ont également contribué à l'essor du mouvement pour les données ouvertes, requérant et alimentant une culture ou une « familiarité » accrue des données, et une amélioration des systèmes de création et de diffusion de l'information.

Les données ouvertes sont alors devenues davantage qu'un outil de communication ; elles reflètent des normes et pratiques qui font de l'ouverture des données un instrument au service de l'efficacité et de la transparence

6. Il convient de noter que l'expression « révolution des données » ou « révolution industrielle des données » est apparu plus de cinq ans auparavant, notamment dans des articles de Joseph M. Hellerstien et de Chris Anderson en 2008.
7. Intitulé « A World That Counts. Mobilising the Data Revolution for Sustainable Development » et rédigé par un groupe d'une vingtaine d'experts indépendants nommés par le secrétaire général des Nations unies.
8. Traduction de la rédaction d'*Afrique contemporaine*.

des politiques publiques et de l'engagement citoyen. Le mouvement est notamment animé par une communauté d'acteurs de la société civile, d'organisations internationales et de gouvernements membres du Partenariat pour un gouvernement ouvert, dont dix pays africains : Tunisie, Sierra Leone, Liberia, Côte d'Ivoire, Nigeria, Burkina Faso, Afrique du Sud, Malawi, Tanzanie et Kenya[9].

Encadré 1 – Qu'est-ce que l'*open data* ?
Selon la définition donnée par la Commission générale de terminologie et de néologie, les données ouvertes sont des « données qu'un organisme met à la disposition de tous sous forme de fichiers numériques afin de permettre leur réutilisation ». Comme le précise Légifrance (2014), elles sont « accessibles dans un format favorisant leur réutilisation » et « n'ont généralement pas de caractère personnel ». L'ouverture des données est une « politique par laquelle un organisme met à la disposition de tous des données numériques, dans un objectif de transparence ou afin de permettre leur réutilisation, notamment à des fins économiques ».

Un bon exemple des données ouvertes au service de l'efficacité de l'action humanitaire est la plateforme HDX[10] développée par le bureau de la coordination des affaires humanitaires des Nations unies, Unocha[11], qui permet aux populations, aux ONG, aux collectivités locales et aux bailleurs de partager de l'information en temps réel et de faciliter la coordination des efforts et de l'allocation de l'aide. Cette plateforme s'est révélée particulièrement utile après le tremblement de terre qui a eu lieu au Népal en 2015[12]. Elle a également montré son utilité au plus fort de la crise du virus Ebola[13].

À la source du *big data* : mise en données du monde et irruption du secteur privé. Les données massives produites par les nouvelles technologies de l'information irriguent d'ores et déjà nos économies et sociétés. À mesure que les services et appareils digitaux se sont immiscés dans nos vies, ils en sont aussi devenus les capteurs et les prescripteurs. Si la première bulle d'Internet a éclaté suite aux promesses non tenues des gains publicitaires engendrés par la Toile, les nouvelles données et capacités d'analyse permettent d'acquérir une connaissance toujours plus fine des individus et des groupes *via* les traces numériques, les *digital breadcrumbs*[14], qu'ils sèment un peu partout – qu'elles soient structurées (données téléphoniques ou bancaires) ou non structurées (photos, vidéos et textes en ligne). En retour, cela permet d'agir sur ces mêmes comportements. Les gains et enjeux commerciaux sont gigantesques ; les algorithmes d'Amazon et de Facebook anticipent et façonnent nos actions dans des mesures impensables il y a dix ans de cela.

Encadré 2 – *Big data* : une nouvelle définition

Les *big data* ou « données massives » recouvrent un ensemble de données hétérogènes – pour ne pas dire hétéroclites. Il est d'usage de les décrire par les « 3V » de « vélocité » (fréquence d'actualisation élevée), « variété » (images, données de téléphonie mobile, données issues de capteurs, textes, etc.) et « volume », la masse d'informations qui en résulte étant considérable. Cependant, cette description laisse de côté le rôle des capacités, notamment technologiques et humaines, sans lesquelles ces données demeureraient inertes, ainsi que le rôle des acteurs et les enjeux d'économie politique. On préfère alors parler des « 3C », de « crumbs » (miettes en anglais – pour évoquer l'idée de traces digitales laissées derrières nous), « capacités » (humaines, technologiques, techniques, institutionnelles) et « communauté » pour évoquer l'émergence et interactions d'acteurs divers, producteurs, collecteurs, analystes, au sein de ce nouvel écosystème. Les données massives ne sont pas des données ouvertes.

Les applications et implications pour les politiques et programmes de développement sont également conséquentes – en particulier celles liées à la collecte et analyse des données de téléphones portables, dénommés Call Detail Records (CDR), comme cela a été démontré et discuté dans une vaste et riche littérature.

De la pauvreté à la mobilité, *via* l'analphabétisme et la criminalité, en passant par l'épidémiologie, la cohésion sociale et la diversité ethnique, rares sont les domaines qui n'ont pas été étudiés par le prisme des CDR. Il est aujourd'hui possible de cartographier de façon précise les déplacements des populations, notamment dans le cas de catastrophes humanitaires[15]. On peut également détecter les lieux de passage les plus favorables aux commerces, étudier les migrations internes – voire internationales – mais aussi optimiser le trajet des transports publics. C'est l'objet du projet AllAboard, vainqueur du concours D4D Côte d'Ivoire. L'équipe du laboratoire IBM de Dublin a ainsi été en mesure d'observer les déplacements des habitants d'Abidjan – ceux munis d'un téléphone portable – détectant les points de départ et d'arrivée des populations et retraçant les trajets effectués. L'optimisation proposée des parcours empruntés par les transports publics permettrait alors aux voyageurs d'économiser 10 % de leur temps de transport.

Le suivi des schémas de propagation des épidémies, observables *via* le même type de données de mobilité, est également en jeu. En effet, en

9. Voir : www.opengovpartnership.org/countries.
10. Humanitarian Data Exchange.
11. United Nations Office for the Coordination of Humanitarian Affairs.
12. 84 jeux de données ont été partagés pour suivre les mouvements de populations à la suite du tremblement de terre, l'évolution des prix, les activités des ONG, etc. Voir : https://data.humdata.org/group/nepal-earthquake.
13. Voir : https://data.humdata.org/ebola.
14. Littéralement, les « miettes numériques ».
15. La première application à grande échelle a été développée par l'ONG suédoise Flowminder en Haïti en 2010. Voir : http://www.flowminder.org/case-studies/haiti-earthquake-2010.

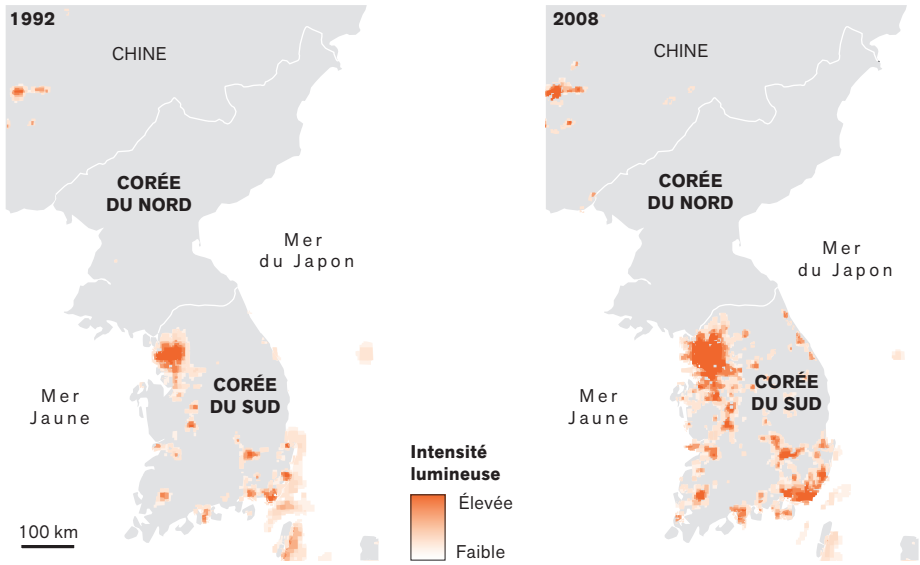
cartographiant les déplacements des populations, il serait possible d'anticiper les contaminations.

Les CDR permettent aussi d'estimer le niveau de pauvreté des populations en temps réel en analysant la relation entre les caractéristiques des appels émis dans une aire géographique – durée, volume, etc. – et le niveau de revenus tiré d'enquêtes officielles ; ainsi que nombre d'indicateurs sociodémographiques classiques, comme l'ont montré les lauréats du D4D[16] 2014 sur le Sénégal[17]. En étudiant les habitudes téléphoniques, les distances d'appels, la durée des communications, leur fréquence avec différents destinataires, la diversité des zones vers lesquelles sont émis les appels, il est possible, depuis plusieurs années, de caractériser et comprendre la structure et nature des réseaux sociaux individuels (Eagle *et al.*, 2009), et on peut imaginer un indicateur de richesse ou de résilience des relations sociales.

D'autres méthodes, qui utilisent des sources de données différentes, ont été développées, seule ou en combinaison avec des CDR. En prenant un peu de hauteur, Henderson *et al.* (2012) ont tenté de mesurer la croissance économique depuis l'espace. Dans leur article, ils observent, de nuit, l'intensité lumineuse pour en déduire l'activité économique et son évolution. La figure 1 ci-dessous

Évolution de la croissance économique sur la péninsule coréenne

Intensité lumineuse vue de l'espace, 1992-2008



Source : Henderson, J.V., Storeygard, A., Weil, D.N. (2012), "Measuring Economic Growth from Outer Space", https://www.aeaweb.org/articles?id=10.1257/aer.102.2.994.

Henderson *et al.* (2012) ont tenté de mesurer la croissance économique depuis l'espace. Dans leur article, « Measuring Economic Growth from Outer Space », ils observent, de nuit, l'intensité lumineuse pour en déduire l'activité économique et son évolution. Ces cartes montrent l'évolution de l'intensité lumineuse sur la péninsule coréenne entre 1992 et 2008.

EdiCarto, 02/2017.

montre l'évolution de l'intensité lumineuse sur la péninsule coréenne entre 1992 et 2008.

Plus généralement, l'utilisation de données issues de capteurs – téléphones mobiles, satellites, etc. – pourrait fournir de nouveaux types de mesures, plus granulaires, plus fréquentes, pour un coût de collecte réduit.

Le secteur privé, engagé dans ce qui a été dénommé « data-philanthropie », n'est pas en reste. En 2013, Orange et d'autres partenaires lançaient la première version de son challenge D4D, mettant à disposition de chercheurs des données issues de Côte d'Ivoire, avec pour objectif d'identifier et de tester, grandeur nature, les usages possibles des *big data* produites *via* son réseau de téléphonie mobile, pour la formulation de politiques publiques – avec un succès qui surprit ses organisateurs. D'autres opérateurs, comme Telefónica, ont également emprunté cette voie.

De nouvelles données et de nouveaux biais. Jusqu'à présent, les données utilisées en sciences sociales étaient « construites » à la suite d'un processus actif et pensé de collecte issue d'observations, de questionnaires. Dans l'ère du *big data*, les données sont principalement « émises » de façon passive et collectées à des fins différentes[18].

L'utilisation de ces données en sciences sociales n'est pas si simple, ni toujours judicieuse. Elle soulève un certain nombre de questions – notamment méthodologiques, mais aussi éthiques. Premièrement, celle de leur validité. Les données traditionnellement utilisées résultent d'une construction théorique et de processus contrôlés : que souhaitons-nous mesurer ? Comment capturer l'information ?

Bien souvent, avec les *big data*, le problème se pose aux statisticiens et aux chercheurs en sens inverse : quelles données existent ? Que peut-on en faire ? Comment y accéder ? Dans la pratique, la distinction est floue. En réalité, seule une minorité de chercheurs en sciences sociales peut se permettre de constituer une base de données spécifique. La majeure partie des chercheurs se pose les mêmes questions : de quelles données disposons-nous ? Comment les traiter de manière adéquate ? Le *big data* ne signe pas la fin de la théorie ni l'obsolescence de la méthode scientifique[19].

La plupart des données massives souffrent de problèmes spécifiques : elles peuvent être en réalité partielles – car tout n'est pas quantifiable – et parfois partiales, et donc trompeuses. D'autant plus qu'elles peuvent donner, par leur taille, richesse et diversité, un sentiment d'exhaustivité. D'une part, les

16. Data for Development.
17. Voir : www.d4d.orange.com/fr/content/download/43452/406501/version/1/file/D4DChallengeSenegal_Book_of_Abstracts_Posters.pdf. Pour une vue d'ensemble des travaux fondateurs dans ce domaine, voir http://www.developmentprogress.org/blog/2013/06/11/could-big-data-provide-alternative-measures-poverty-and-welfare.
18. À cet égard, un responsable de l'US Bureau of Labor Statistics décrit les *big data* comme « nonsampled data, characterized by the creation of databases from electronic sources whose primary purpose is something other than statistical inference ». Voir : http://magazine.amstat.org/blog/2013/01/01/sci-policy-jan2013/.
19. Voir : http://www.wired.com/2008/06/pb-theory/.

données non structurées et « subjectives » – ce que nous choisissons de partager en ligne – doivent être traitées avec prudence, et requièrent un sérieux travail sociologique et anthropologique[20].

Les données plus objectives, comme les données de téléphones portables ou de transactions bancaires, ne sont pas sans poser de problèmes méthodologiques. Elles tendent à sous-représenter les activités de la part de la population la moins connectée, notamment les plus pauvres, les personnes à handicap, etc. Dans le cas de catastrophes naturelles, ces « signaux » numériques, provenant de volontaires ou passivement collectés, peuvent provenir de zones relativement épargnées – invalidant tout diagnostic hâtif[21]. La question de la (non)-représentativité de l'échantillon et de la correction des biais statistiques garde son sens pour établir des inférences valides. Les deux règles d'or de la recherche classique s'appliquent : que disent vraiment ces données sur l'échantillon considéré ? Il s'agit d'une question de validité interne. D'autre part, les conclusions (valides) établies pour cet échantillon sont-elles généralisables dans le temps et dans l'espace ?

Quelle articulation avec la statistique publique et la méthode scientifique ?

Il est difficile d'imaginer que le mouvement de production et de collecte massive de données s'arrête dans un futur proche. La mise en données du monde est bel et bien en marche et avec elle des enjeux qui incluent les aspects méthodologiques évoqués, mais aussi éthiques, institutionnels et politiques. Sous la pression des données massives, le monopole du chiffre officiel continue de se fissurer[22] et l'on assiste à la montée en puissance d'acteurs privés, producteurs de données[23], aux côtés des INS.

Néanmoins, un certain nombre de questions se posent quant aux incitations et à la capacité du secteur privé à partager ces informations, ainsi que sur la réticence des citoyens à la réutilisation de données à caractère personnel, ou encore la propension des pouvoirs publiques à contrôler ces usages. Enfin, des interrogations subsistent sur la volonté et la capacité des INS à tirer profit de ces nouvelles données. Il s'agit, dans le premier cas, de surmonter les craintes liées au partage de données pouvant mettre en danger une position concurrentielle ou la vie privée, et, dans le second cas, d'une question d'économie politique qui interroge le caractère monopolistique de l'information statistique qui rend compte des performances et des priorités des politiques publiques et en permet le débat. Ainsi formulé, on comprend l'enjeu démocratique sous-jacent.

Ne pas confondre *big data* et *open data*. Si leur émergence est concomitante et qu'elle participe de la « révolution des données », il ne faut pas confondre *big data* et *open data*. Une grande partie des *big data* revêt un caractère personnel[24]. Elles sont émises par les individus[25], traitées, puis stockées[26], par des

entreprises privées. Les *open data* sont, quant à elles, des données administratives ou issues du secteur public et mises à disposition des citoyens et des entreprises, généralement sous une forme structurée. En tant que mouvement ou écosystème, le *big data* et l'*open data* sont également distincts, le second, plus ancien, étant logiquement plus mature et structuré que le premier ; même si au fil des années, des liens et entrelacements ont progressivement vu le jour pour former, au côté de la statistique publique, deux des trois principaux ensembles constituant l'univers actuel des données.

L'analyse démographique en temps réel offre un exemple pertinent aussi bien sur le plan scientifique que politique. Connaître la distribution et la densité – voire la composition – d'une population en temps réel peut sauver des vies dans le cas d'une catastrophe naturelle. Les recensements en fournissent une cartographie fine, mais tous les dix ans au mieux, certains pays n'en ayant pas organisé depuis des décennies. L'activité téléphonique ne serait-elle pas une meilleure source de données ? Des travaux récents ont montré que l'hypothèse était réaliste (Wesolowski *et al.*, 2013 ; Deville *et al.*, 2014). Cela suppose de comprendre et de pouvoir corriger le biais de sélection inhérent au fait que la possession et l'utilisation d'un téléphone portable ou d'une messagerie électronique varient en fonction de l'âge, du niveau d'éducation, etc., des individus et pays concernés. Ce qui vaut pour Manhattan ne vaut pas pour Nouakchott. L'objectif est de mieux calibrer les modèles utilisés, ce qu'un ensemble d'études tente de faire depuis plusieurs années (Zagheni, Weber, 2015 ; Pestre *et al.*, 2016). Ce travail de calibration requiert que soient disponibles des données « réelles » – dites *ground truth* – qu'elles proviennent de recensement, de bases de données administratives ou d'enquêtes idoines.

Opposer les sources et méthodes du *big data* à celles des sciences sociales et de statistique publique est donc une erreur. Il y a là aussi une dialectique et des complémentarités à renforcer[27]. Il convient d'approcher les données comme un écosystème, avec différents types de données pour des usages différenciés mais complémentaires, et surtout de nouveaux acteurs.

20. Voir : https://www.timeshighereducation.com/books/its-complicated-the-social-lives-of-networked-teens-by-danah-boyd/2013266.article.
21. Voir : http://benetech.blogspot.com.br/2011/03/crowdsourced-data-is-not-substitute-for.html et https://hbr.org/2013/04/the-hidden-biases-in-big-data.
22. Voir : http://en.istat.it/istat/eventi/2010/10_conferenza_statistica/Relazione_pres_10conf.pdf.
23. Bien entendu, il existe déjà des entreprises privées productrices de données, des instituts de sondages, etc. La différence vient du fait que les données massives sont émises et sont souvent le « sous-produit d'une autre activité ». Elles ne sont pas produites à dessein ou commandités.
24. C'est le cas des informations issues de capteurs humains (téléphonies mobiles, consommations électriques, utilisation de réseaux sociaux, etc.). C'est un peu moins le cas d'imageries satellites observant des phénomènes naturels.
25. De façon active (publication sur les réseaux sociaux, appel téléphonique, paiement par carte bancaire ou mobile) ou passive (données collectées lors de la navigation sur Internet, *via* les téléphones mobiles lors des déplacements, etc.).
26. Elles ne sont pas forcément conservées, il peut s'agir de flux.
27. Voir : https://web.stanford.edu/class/ee380/Abstracts/140129-slides-Machine-Learning-and-Econometrics.pdf.

La difficulté pour le secteur privé de partager des données. Faisons l'hypothèse que les entreprises privées souhaitent mettre à disposition des informations pour améliorer leur image, ou encore l'efficacité des politiques publiques – ce qui en retour pourrait améliorer leurs performances si en découlent des services publics plus efficaces. Elles font alors face à un défi non négligeable, car ce partage d'information n'est ni simple, ni anodin. Il induit certains coûts[28], et des risques. Quelles données partager ? Sous quelle forme ? Avec qui ? Comment éviter de fournir à ses concurrents une information stratégique ? Qui prendra la décision au sein de l'entreprise ? Si les données concernent leurs clients, comment ces derniers vont-ils réagir ? Ces données peuvent-elles leur porter préjudice ?

Les entreprises doivent se poser toutes ces questions avant de partager une information vers l'extérieur. Différentes directions au sein des entreprises doivent être mobilisées : la communication, le service juridique, le marketing, la production, etc. Ainsi, c'est souvent au niveau du comité exécutif, plutôt réticent à la prise de risques, que ce type de décisions est pris. Avant même la transformation de ces données non structurées en indicateurs robustes de suivi des politiques publiques, une difficulté majeure se pose : comment partager des données ? Comment collaborer avec les pouvoirs publics, avec les chercheurs, pour tirer profit du potentiel qu'offrent ces nouvelles données et ces nouveaux outils en minimisant les risques ? Comment soutenir le secteur privé et l'inciter à partager une information potentiellement utile à la formulation et au suivi des politiques publiques ? Comment organiser concrètement ce partage ?

Ces discussions sont arrivées en force sur l'agenda des organisations internationales, à l'ONU, la Banque mondiale, au Forum économique mondial[29], notamment à l'occasion de l'épidémie d'Ebola en Afrique de l'Ouest. Fallait-il, comme certains l'ont demandé, « ouvrir » – mettre à disposition à des équipes universitaires ou agences onusiennes – les données de téléphones portables collectées en Sierra Leone, en Guinée et au Liberia ? Divers facteurs et obstacles institutionnels et légaux, mais aussi éthiques, l'ont empêché[30]. Mais la question de fond et de long terme demeure : faudra-t-il le faire à l'avenir et si oui comment ?

Nouveaux espaces, nouveaux systèmes de dialogues et nouveaux partenariats « public-privé-personnes »

Les défis soulevés par la mise en données du monde ne doivent pas être considérés comme des questions uniquement techno-scientifiques. Ils sont autant – comme pour toutes innovations majeures – politiques et éthiques. Comment protéger la vie privée des citoyens ? Comment réguler l'utilisation des données ? Le fait que nous ne puissions pas connaître à l'avance l'usage qui sera fait de nos données privées pose un problème éthique. Pouvons-nous refuser certains usages de nos données *a posteriori* ? Aujourd'hui, la réponse est non car pour utiliser les réseaux sociaux ou avoir accès à une ligne téléphonique,

chaque utilisateur donne son « consentement » à la réutilisation de ses données personnelles, sans savoir quel usage en sera fait dans le futur.

De fait, ce consentement n'est ni vraiment libre – car de ce choix dépend l'accès ou non à un service perçu comme essentiel – ni pleinement éclairé, car personne ne prend le temps de lire les termes et conditions afférents, ni ne peut anticiper les utilisations futures. En effet, souvent l'opérateur ne le sait pas lui-même ! Les fondateurs de Facebook auraient sans doute eut de la peine à imaginer la richesse et la valeur des données qu'ils allaient engendrer et collecter. Facebook en est aujourd'hui pleinement conscient, et pourtant l'entreprise semble en quête de sens, multipliant les consultations, contacts, expérimentations, et commettant parfois des impairs surprenants.

Facebook s'est ainsi confronté à une importante levée de boucliers suite à la parution d'une étude interne qui visait à analyser les réactions de ses utilisateurs auxquels étaient proposés, au gré de changements de ses algorithmes, des fils de nouvelles positives ou négatives. Celle-ci respectait les termes et conditions acceptés par toute personne ayant un compte. Sa légalité n'était nullement en question. Et pourtant l'argument légal est apparu non pertinent car chacun sait que personne ne lit ces conditions. C'est sur le terrain de l'éthique que les critiques et attaques se sont déployées. « On ne peut pas jouer avec nos sentiments », « Nous ne sommes pas des rats de laboratoire » ont dit en en substance ses opposants (certains n'étant d'ailleurs pas tous des utilisateurs). Au fond, c'est un lien de confiance implicite qui a semblé rompu, une ligne imaginée comme allant de soi qui s'est trouvée franchie. Facebook a rapidement pris la mesure du faux pas et présenté ses excuses.

La question du consentement, du contrôle, mais aussi de la confiance, et ainsi des modalités d'établissement de ce qui constitue des pratiques jugées éthiques, ou simplement socialement acceptables, est au cœur du futur du *big data*, la condition de sa survie (Pentland, 2014). Jusqu'ici, les régulations nationales – ou européennes – définissent certaines règles qui protègent les citoyens-utilisateurs, plus ou moins efficacement. Mais des règles trop restrictives signeraient la fermeture de nombreuses pistes de recherche permettant de mieux comprendre le fonctionnement des sociétés humaines comme systèmes complexes. Cela doit se faire au travers de débats démocratiques, informés, où les considérations et contraintes de divers acteurs pourront être exposées et sous-pesées.

À l'image de ce qui est advenu dans le champ du vivant avec la bioéthique, il faut définir une éthique des données, une « data-éthique ». De par leur diversité, le quasi-monopole du secteur privé dans leur collecte mais aussi de par leur importance stratégique dans une économie dématérialisée, la mise en place d'un « consensus global des données » est complexe, mais le jeu

28. Qui peuvent être considérés comme des investissements.
29. World Economic Forum.
30. Voir : http://cis-india.org/papers/ebola-a-big-data-disaster et http://datapopalliance.org/wp-content/uploads/2015/04/WPS_LawPoliticsEthicsCellPhoneDataAnalytics.pdf.

démocratique et les intérêts des citoyens pourraient en sortir renforcés dans des pays où l'information est contrôlée de près par le pouvoir.

Quels que soient les contours exacts de cette « nouvelle donne » sur les données[31], une chose semble certaine : les citoyens doivent pouvoir exercer un contrôle plus grand sur l'utilisation de leurs données. Dans cette optique, le *big data* n'est plus la seule affaire des « geeks » ; il doit servir de cadre d'un dialogue renouvelé entre acteurs sociaux portant sur l'utilisation de la matière première de l'économie dématérialisée.

À quelles conditions cette « nouvelle donne » peut-elle voir le jour et être pérennisée ? Tout d'abord, la réalisation de cette vision requiert le développement, à grande échelle, de la *data literacy* – difficilement traduisible par « alphabétisation (ou familiarité) aux données » – à divers niveaux de la société. Qui plus est, il ne doit pas s'agir de produire des *data crunchers* en série, mais de donner aux citoyens les incitations et outils nécessaires au plein exercice de leur rôle et à la poursuite de leurs objectifs dans un monde où les données seront omniprésentes. Les futurs programmes scolaires incluront évidemment l'apprentissage des bases du code ; il conviendra également de l'accompagner de cours sur l'histoire et les enjeux éthiques et politiques des données.

Cette vision de citoyens désireux et à même de se saisir de la donnée comme levier de pouvoir peut sembler utopiste, mais c'est peu ou prou celle de l'*open data*, à une échelle infiniment plus grande. En ce sens, elle pourrait remettre en question les systèmes et les zones de pouvoir actuels (les grandes entreprises, les États-nations). La gouvernance d'un monde de citoyens-données en réseau hors des canaux et schémas contemporains n'est pas chose aisée à imaginer ; peut-être pas moins que ne l'était l'avènement du suffrage universel il y a cinq siècles. Qui sait ce qu'il adviendra dans cinq décennies ?

Des partenariats « publics-privés-personnes » pour la société digitale. Avec l'irruption du secteur privé et la mise en données du monde, la production de données décrivant la sphère publique n'est plus un monopole d'État[32]. Le dialogue doit donc s'élargir, s'ouvrir aux entreprises, à la société civile, aux collectivités locales, etc. Puisque tous ces acteurs sont concernés, il faut créer un nouveau contrat social. Pour que cela puisse se faire au service des citoyens, il faut leur permettre de pouvoir participer aux débats, d'avoir voix au chapitre. Pour ce faire, il faut réduire la fracture numérique et renforcer la *data literacy*, la familiarité avec les données.

À ce jour, il n'existe pas de système qui permettrait la mise à disposition de données privées, ouvertes mais protégées, pour servir à l'élaboration et au suivi de politiques publiques et au développement d'indicateurs statistiques complémentaires ou plus granulaires. C'est peut-être que la question est mal posée ; il s'agit en définitive moins de mettre à disposition de façon épisodique que de rendre accessible, de façon stable et prévisible.

Face à ces défis et fort d'expérimentations[33] et expériences[34] accumulées au fil des années, une coalition d'acteurs autour de Data-Pop Alliance, le

MIT, Orange et le Forum économique mondial, avec le soutien et l'implication de l'AFD et de la Banque mondiale, travaille actuellement au développement d'une plateforme et d'un environnement permettant non pas de sortir les données, mais de leur soumettre des questions.

Ce projet, dénommé OPAL (Open Algorithms), prend acte du fait que les entreprises privées ne peuvent à ce jour partager leurs données massives à grande échelle de façon stable, systématique et sécurisée. Il vise à mettre en place des modalités d'accès aux données massives pour la formulation des politiques publiques et le suivi de certaines cibles des ODD. Concrètement, l'idée consiste à rendre disponible des indicateurs calculés à partir de *big data*, à l'aide d'algorithmes ouverts et vérifiables par des tiers, mais dont les données sources ne seraient pas ouvertes.

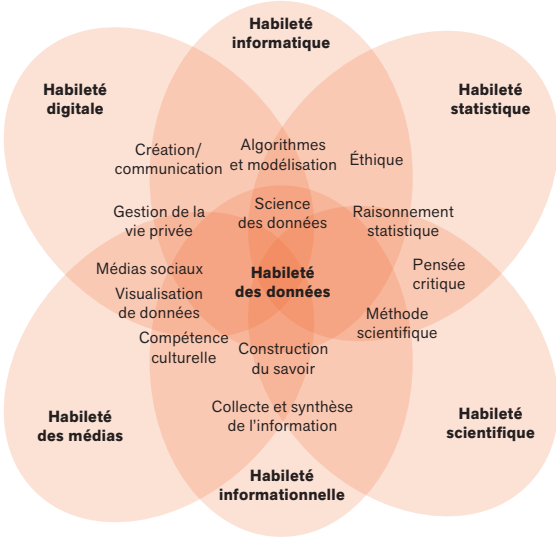
Pour cela, il propose d'inverser le paradigme actuel et de repenser le processus de production d'indicateurs. Actuellement, les données sont acheminées vers les INS et les algorithmes de calcul. Avec OPAL, c'est l'inverse, car il propose aux algorithmes de rejoindre les données : *bring the code to the data, not the data to the code*[35].

Ce compromis permettrait de rassurer à la fois les entreprises et les individus, tout en mettant à disposition l'information réellement utile aux citoyens et aux pouvoirs publics. Cette solution souple, décentralisée et relativement peu onéreuse permettrait d'inciter le secteur privé à collaborer davantage avec les INS et de promouvoir l'idée d'algorithmes ouverts[36]. Cette collaboration n'impliquerait pas une diminution des ressources des INS. Au contraire, elle permettrait un renforcement des capacités par une offre de formation aux nouveaux outils utilisés par la science des données, qui pourraient, à terme, faciliter le travail de production et de diffusion des statistiques officielles (notamment en ce qui concerne les méthodes de collecte et de traitement des données, visualisation de données, etc.).

Le projet vise à donner ou rendre un rôle central aux systèmes statistiques publics dans la production des indicateurs, dans le respect des principes fondamentaux de la statistique publique. Un des risques que pose l'émergence de nouveaux acteurs aptes à « produire du chiffre » est la prolifération de statistiques « officielles » – portant sur le niveau d'inflation, de PIB,

31. New Deal on Data-WEF, Alex Pentland.
32. Rappelons que le mot « statistique » a été forgé au XVIII[e] siècle en Allemagne par l'économiste Gottfried Achenwall pour décrire l'ensemble des connaissances qu'un homme d'État doit posséder.
33. Voir : http://bandicoot.mit.edu et http://openpds.media.mit.edu.
34. Dont les D4D susnommé ; voir : http://www.d4d.orange.com/.
35. Littéralement : « Apporter les code vers les données et non les données vers le code. »
36. Un algorithme est une série d'instructions informatique permettant d'arriver au résultat souhaité. À l'ère des données massives, ils sont partout et permettent, par exemple, à un moteur de recherche d'afficher une centaine de pages Web liées à une requête après avoir analysé l'ensemble d'Internet. On reproche souvent leurs opacités aux algorithmes. Par exemple, les citoyens ne peuvent savoir quelles actions ont participé à l'affiche sur leur navigateur de tel ou telle publicité ou quelles informations sont collectées.

Schéma sur l'habileté des données



Source : Bhargava, R., Deahl, E., Letouzé, E., Noonan, A., Sangokoya, D., Shoup, N. (2015), "Beyond Data Literacy. Reinventing Community Engagement and Empowerment in the Age of Data", http://datapopalliance.org.

EdiCarto, 02/2017.

du chômage – non conformes aux principes fondamentaux de la statistique publique, qui rendraient tout débat démocratique difficile.

Conclusion

Chercheurs, praticiens, citoyens réalisent et démontrent par leurs travaux et leurs échanges la centralité des dimensions éthiques, légales, institutionnelles et politiques du *big data* – sous son vernis techno-scientifique. La technologie et la science continuent aussi de modifier les termes mêmes des débats publics, alors que ceux-ci ont à peine commencé ; la notion d'« anonymisation » semble déjà considérée comme obsolète.

Avec en ligne de mire la montée en puissance de l'« Internet des choses » et de l'intelligence artificielle, ce qui est en cours et en jeu est une course entre sociétés et machines, entre acteurs et visions du monde ; ce qui est requis s'apparente à une remise en cause, voire à plat, de structures et relations de pouvoir fondées sur la maîtrise d'une matière principale rare ou limitée du système économique et politique : la terre, le savoir, le capital, les sources d'énergie fossiles. À quoi ressemblera le système économique et politique de l'âge des données ? Qui décidera de ses codes et métriques ? Quels nouveaux rôles et responsabilités incomberont aux anciens protagonistes ? Quels nouveaux acteurs et alliances peuvent et doivent émerger ? Quels principes éthiques et investissements éducatifs sont-ils nécessaires ? Quels systèmes légaux et arrangements

institutionnels sont appelés à être inventés ? Comment « sauver le *big data* de lui-même » et en faire un levier et moteur d'émancipation et de progrès plutôt que d'asservissement et de contrôle est une question essentielle.

Au stade actuel du développement du *big data*, deux priorités émergent. D'une part, il est urgent d'investir massivement dans la familiarité des données, dans les INS et en dehors. D'autre part, il est essentiel de stimuler le dialogue et les collaborations entre acteurs qui pourront tisser des liens de confiance et de redevabilité, promouvoir l'innovation, favoriser le progrès social. Par leur légitimité technique et politique, les INS ont un rôle central à jouer.

Bibliographie

Bhargava, R., **Deahl**, E., **Letouzé**, E., **Noonan**, A., **Sangokoya**, D., **Shoup**, N. (2015), "Beyond Data Literacy. Reinventing Community Engagement and Empowerment in the Age of Data", http://datapopalliance.org.

Razafindrakoto, M., **Roubaud**, F. (2003), « Les dispositifs existants de suivi de la pauvreté : les faiblesses des enquêtes classiques auprès des ménages », *in* J.-P. Cling, M. Razafindrakoto, F. Roubaud, *Les Nouvelles Stratégies internationales de lutte contre la pauvreté*, 2e éd., Paris, Economica/IRD, p. 313-338.

Devarajan, S. (2011), "Africa's Statistical Tragedy", Banque mondiale, http://blogs.worldbank.org.

Deville, P., **Linard**, C., **Martine**, S., **Gilbert**, M., **Forrest**, R.S., **Gaughan**, A.E, **Blondel**, V., **Tatem**, A.J. (2014), "Dynamic Population Mapping Using Mobile Phone Data", www.pnas.org.

Eagle, N., **Pentland**, A., **Lazer**, D. (2009), "Inferring Social Network Structure Using Mobile Phone Data", http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.379.4719&rep=rep1&type=pdf.

Henderson, J.V., **Storeygard**, A., **Weil**, D.N. (2012), "Measuring Economic Growth from Outer Space", www.aeaweb.org.

Hilbert, M. (2015), "Quantifying the Data Deluge and the Data Drought", http://datapopalliance.org.

Jerven, M. (2013), *Poor numbers. How We Are Misled by African Development Statistics and What to Do about It*, New York, Cornell University Press, 2013.

Legifrance (2014), « Vocabulaire de l'informatique et du droit », www.legifrance.gouv.fr.

New York Times (2009), "For Today's Graduate, Just One Word. Statistics", www.nytimes.com.

Paris21 (2015), "A Road Map for a Country-Led Data Revolution", http://datarevolution.paris21.org/sites/default/files/Road_map_for_a_Country_led_Data_Revolution_web.pdf.

Pentland, A. (2014), "Saving Big Data from Itself", *Scientific American*, http://www.nature.com/scientificamerican/journal/v311/n2/full/scientificamerican0814-64.html.

Pestre, G., **Letouzé**, E., **Zagheni**, E. (2016), "The ABCDE of Big Data. Assessing Biases in Call-Detail Records for Development Estimates", http://pubdocs.worldbank.org/en/551311466182785065/Pestre-Letouze-Zagheni-ABCDE-May-2016.pdf.

Stevance, A.S. (2015), "Review of Targets for the Sustainable Development Goals. The Science Perspective", ICSU, ISSC, http://www.icsu.org/publications/reports-and-reviews/review-of-targets-for-the-sustainable-development-goals-the-science-perspective-2015/SDG-Report.pdf.

Wesolowski, A., **Eagle**, N., **Abdisalan**, M.N., **Snow**, R.W., **Buckee**, C.O. (2013), "The Impact of Biases in Mobile Phone Ownership on Estimates of Human Mobility", 10.1098/rsif.2012.0986 http://rsif.royalsocietypublishing.org/content/10/81/20120986.

Zagheni, E., **Weber**, I. (2015), "Demographic Research with Non-Representative Internet Data", http://www.zagheni.net/uploads/3/1/7/9/3179747/zagheni_weber2015.pdf.