DATA-POP
ALLIANCE

Mobiliser et humaniser la Révolution des Données pour la statistique publique, le développement et la démocratie

Note au Bureau Pays du PNUD Togo

Emmanuel Letouzé[1]

28 novembre 2018

Les pays les moins développés, en Afrique notamment, souffrent d'un déficit de données socio-économiques et démographiques qualifié de « tragédie statistique » par certains experts, alors même que le renforcement des institutions et pratiques démocratiques s'y déroule de façon inégale. 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. Le paysage des données a été en effet révolutionné par le phénomène des données ouvertes ou Open Data au tournant du siècle, puis l'explosion depuis 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. Les phénomènes Open Data, Big Data, et plus largement la révolution des données en cours et appelée de leurs vœux par les Nations unies, et l'émerge de l'Intelligence Artificielle (IA), sont amenés à transformer les sociétés au 21ème siècle, et notamment la prise de décision et le suivi des politiques et programmes de développement. Loin de la rendre obsolète, cette révolution des données, si elle veut contribuer à l'accélération du développement et au raffermissement de la démocratie en Afrique, requiert de renforcer le rôle central de la statistique publique dans les futures sociétés de la connaissance.

Si les promesses sont nombreuses, les indicateurs produits à partir des données massives ne pourront pas combler l'ensemble des déficits de données ; pas plus que la seule production et diffusion de statistiques plus fiables et plus fréquentes ne suffiront à améliorer les phénomènes mesurés, tels que pauvreté, inégalité, mortalité infantile et maternelle, illettrisme, dégradation environnementale, etc. Dans un premier temps, les nouvelles données permettront de compléter l'offre du système statistique en même temps qu'elles poseront des questions et des défis pour l'équilibre des pouvoirs et de respect de la vie privée. Mais surtout, les données doivent constituer le point d'entrée et d'accroche d'un dialogue et d'un contrat social renouvelés, avec pour outils et objectifs une participation et une évaluation citoyennes renforcées et une meilleure efficacité, redevabilité, et réactivité des décisions publiques. En filigrane, c'est l'application maitrisée des principes et outils de l'intelligence artificielle qui déterminera pour partie les trajectoires de développement économique et politique des pays et régions du monde— avec notamment le rôle essentiel de la mesure et de l'apprentissage.

[1] Mise à jour d'un article de l'auteur et de Thomas Roca, publié en 2016: "La révolution des données est-elle en marche ? Implications pour la statistique publique et la démocratie", Afrique contemporaine. http://www.afrique-contemporaine.info/articles/2016-2-la-revolution-des-donnees-est-elle-en-marche/

Contexte : Des OMDs aux ODDs ; de la stratégie statistique à la révolution des données

Les Objectifs du millénaire pour le développement (OMD) annoncés au tournant du siècle ont contribué à mettre le rôle de la mesure – notamment celle 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. 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. D'autres facteurs, tels que la fuite de cerveaux ou le démantèlement de l'URSS qui fournissait un appui à des 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. On découvrit il y a quelques années que les PIB du Ghana, du Nigeria et du Kenya, avaient été largement sous-estimés sur plusieurs années, du fait de la mauvaise prise en rendre compte de la montée en puissance du secteur technologique. 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 de plusieurs années. Bien souvent, ces données manquent de « granularité » – temporelle, géographique et sociale – par âge, sexe, etc.

C'est dans ce contexte que l'ONU a lancé en 2013 et 2014, au moment où étaient définis les Objectifs du développement durable (ODD), un appel pour « révolution des données », 5 ans après la 1ère mention de la notion de « révolution industrielle des données » en 2008. 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 second rapport de 2014 « A World That Counts » :

> *« 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. »*

Mais qu'entend-on par « révolution des données » et qu'attend-t-on d'elle ?

La révolution des données : piliers, promesses et problèmes

La révolution des données est nourrie de deux phénomènes (ou mouvements) distincts mais liés de façon croissante : l'Open Data, ou « données ouvertes », et le Big Data, imparfaitement traduit par « données massives ». 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 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 douze pays africains : Tunisie, Maroc, Sénégal, Sierra Leone, Liberia, Côte d'Ivoire, Nigeria, Burkina Faso, Afrique du Sud, Malawi, Tanzanie et Kenya.

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 ».

La source du Big Data est la « mise en données « du monde et l'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. Les nouvelles données et capacités d'analyse permettent d'acquérir une connaissance toujours plus fine des individus et des groupes via les traces ou miettes numériques, les « *digital breadcrumbs* » 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. 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.

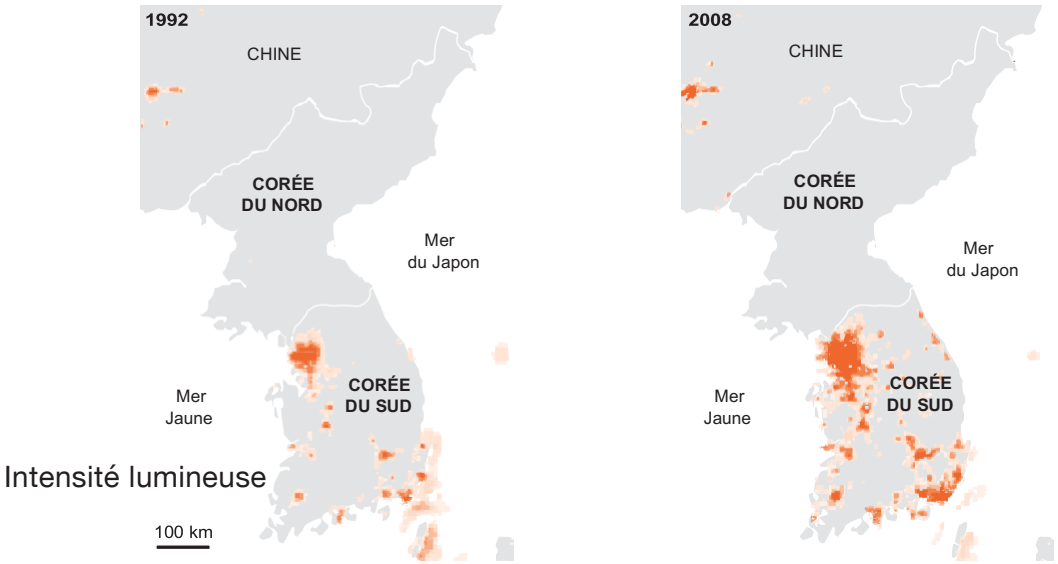
Les applications et implications pour les politiques et programmes de développement sont également conséquentes – en particulier celles liées à la collecte et analyse des données de téléphones portables, dénommés Call Detail Records (CDRs). De la pauvreté à la mobilité, via l'analphabétisme et la criminalité, 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 CDRs. Il est aujourd'hui possible de cartographier de façon précise les déplacements des populations, notamment dans le cas de catastrophes humanitaires.

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'était l'objet du projet AllAboard, vainqueur du concours « Data for Development » (D4D) d'Orange Côte d'Ivoire en 2012-13. L'équipe du laboratoire IBM de Dublin a ainsi été en mesure d'observer les déplacements des habitants d'Abidjan 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 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 2014 sur le Sénégal. 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 vue de l'espace : Intensité lumineuse, 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 montre l'évolution de l'intensité lumineuse sur la péninsule coréenne entre 1992 et 2008. Plus généralement, l'utilisation de données issues de capteurs – télé phones 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, susmentionné, 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.

Ceci pose certains défis. 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. L'utilisation de ces données en sciences sociales n'est pas si simple. 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 ?

Avec les Big Data, le problème semblent se poser 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, mais il les secoue.

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 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.

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. 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, la domination (déjà relative) du chiffre officiel continue de se fissurer et l'on assiste à la montée en puissance d'acteurs privés, producteurs de données, aux côtés des INS—et du phénomène des *fake news*, notamment.

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.

Il ne faut pas confondre Big Data et Open Data, même si leur émergence est largement concomitante et qu'elle participe de la « révolution des données ». Une grande partie des Big Data revêt un caractère personnel. Elles sont émises par les individus, traitées, puis stockées, 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 par exemple 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 (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. 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. Il reste difficile pour le secteur privé de partager des données. Les entreprises privées qui souhaitent mettre à disposition ou « monétiser » leurs données font face à un défi non négligeable, car ce partage d'information n'est ni simple, ni anodin. Il induit certains coûts, 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, 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. 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 eu 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, voire des écarts inquiétants.

Facebook s'est ainsi d'abord 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. Pourtant, depuis a éclaté le scandale dit « Cambridge Analytica », qui a encore plus abimé son image, et avivé les craintes et les questions face au Big Data. Les questions 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 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, 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 chaînes et zones de pouvoir actuelles (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. Moins que ne l'était l'avènement du suffrage universel il y a 5 siècles. Qui sait ce qu'il adviendra dans 5 décennies ?

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. 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.

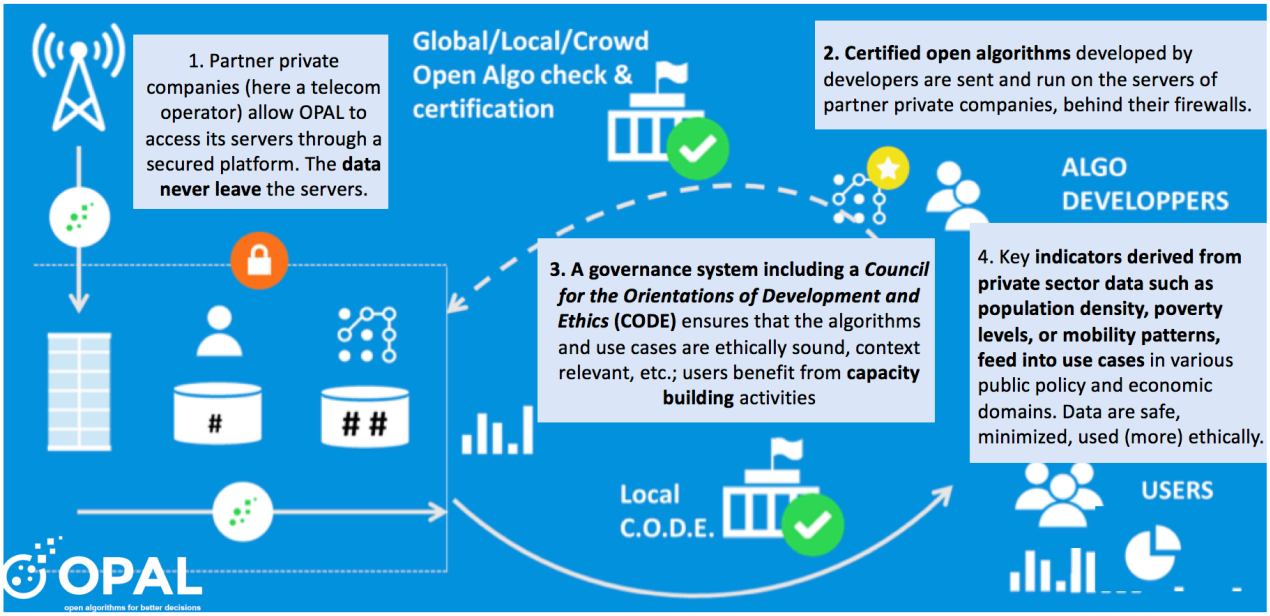
À ce jour, il n'existe pas de système qui permettrait la mise à disposition de données privées, ouvertes mais protégées, pour servir à l'élaboration et au suivi de politiques publiques et au développement d'indicateurs statistiques complémentaires ou plus granulaires. C'est peut-être que la question 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 et expériences 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. OPAL (pour 'Open Algorithms') est un projet d'innovation technologique et sociale visant à déverrouiller le potentiel des données collectées par les compagnies privées pour renforcer la définition et le suivi des politiques et programmes de développement, la redevabilité de l'action publique, et l'engagement des citoyens—et en particulier la mesure et la poursuite des Objectifs de développement durable (ODDs).

Pour ce faire OPAL permettra notamment l'estimation et la diffusion, quasiment en temps réel et à coût marginal nul, d'indicateurs socio-économiques clés tels que des densités de population, des niveaux de pauvreté, ou des mesures de mobilité, par le biais d'une plateforme et d'algorithmes ouverts et certifiés, tournant sur les données d'entreprises privées partenaires—notamment opérateurs de téléphonie mobile—sans jamais les exposer. Ces indicateurs agrégés seront ensuite utilisés dans différents cas d'usages dans divers domaines, telle que la santé publique, l'éducation, le transport, la cohésion sociale, etc.

Une caractéristique principale d'OPAL est d'envoyer les algorithmes aux données—et non l'inverse vers des équipes de recherche spécifiques comme cela a été fait jusqu'alors, avec des résultats prometteurs, mais qui ne permet ni d'assurer le passage à l'échelle, ni la protection des données, ni l'inclusion des populations concernées dans le processus d'analyse et d'utilisation des données. Outre cette composante technique, OPAL vise en effet aussi à développer des normes et capacités de gouvernance locale fortes, via un co-développement, la mise en place d'un Comité d'Orientation pour le Développement et l'Ethique (CODE), et des activités de formation.

OPAL: 1st Generation Data Systems and Standards



OPAL est développé par un consortium autour de Data-Pop Alliance, du MIT Media Lab, d'Orange, d'Imperial College London, et du World Economic Forum, dans le sillage de nombreux travaux et expérimentations de ce groupe dans le domaine des données pour le développement. Le projet a démarré avec 2 pilotes au Sénégal et en Colombie en partenariat avec leurs agences nationales de statistiques et 2 opérateurs téléphoniques majeurs, Sonatel et Telefonica, sur financement de l'Agence française de développent (AFD) à hauteur de 1.5 millions d'Euros, et un appui complémentaire de la Banque mondiale et du Global Partnership for Sustainable Develoment Data (GPSDD). A moyen terme, OPAL a pour ambition de s'ancrer dans ces 2 pays pour y devenir le système socio-technologique de référence pour l'accès et l'utilisation éthiques du 'Big Data' pour le développement, puis de s'étendre à d'autres industries et géographies, en Afrique, Amérique Latine, et Asie.

Le projet vise notamment 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, du chômage – non conformes aux principes fondamentaux de la statistique publique, qui rendraient tout débat démocratique difficile. OPAL entend renforcer l'utilisation maitrisée des principes et outils de l'IA et notamment le rôle de l'apprentissage et l'adaptation sur la base d'observations et de « feedback loops » à l'échelle non plus de simples jeux de données (tels que le fait Google Translate) mais à celle de écosystèmes humains entiers.

Conclusions

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 et la culture 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 développement humain, et la démocratie.