**INSTITUT POUR UN DÉVELOPPEMENT DURABLE**

**Siège social** : 190 Fond de Bondry à1342 Limelette
**Adresse de contact** : 40 rue Henri Lemaître 5000 Namur
**Tél** : 0476.906.365 **E-mail** : idd@iddweb.be

# *Les données : un défi majeur de la révolution numérique*

# Les données au service du bien commun

## Version courte

*par Philippe DEFEYT*[1] *– avril 2018*

**NB : Les références des citations sont données dans la note complète !**

## 0. Introduction

La collecte, le traitement et l'utilisation de données à des fins commerciales ou autres (recherche, action publique...) ne datent pas d'hier.

Pour ce qui est du commercial, on pense par exemple à a comptabilité des marchands assyriens de Kaniš (XIXe siècle av. J.-C.) mais aussi à l'observation des marchés et des prix pratiquée par les marchands d'Assur du temps de la civilisation mésopotamienne.

Pour d'autres préoccupations, on pense évidemment aux recensements de la population.

Depuis lors, la collecte et l'usage de données se sont considérablement développés. C'est l'action d'Adolphe Quetelet qui, en tout cas en Belgique, vient immédiatement à l'esprit ; il est notamment à l'initiative du premier recensement belge à visée scientifique (1846). Mais, d'une manière générale, la production de données s'est, au cours des deux derniers siècles, élargie à de nouveaux secteurs, s'est intensifiée, a multiplié les techniques de production (dont les sondages et autres enquêtes, avec aujourd'hui le développement de sondages rémunérés).

Une illustration parmi beaucoup d'autres de l'extension statistique d'avant la révolution numérique est l'action de Nielsen, qui a longtemps eu une place prépondérante dans la collecte et de la consolidation des données de ventes dans les "supermarchés", détenant donc des informations stratégiques (par exemple le suivi des ventes suite à l'introduction d'un nouveau produit ou le lancement d'une campagne de promotion) intéressant au plus haut point les entreprises concernées. Au travers de multiples contrats avec les acteurs concernés, Nielsen agit comme un intermédiaire. Il joue, pour partie, le rôle que doit/devrait jouer un office statistique public.

Nielsen, comme beaucoup d'autres acteurs économiques (fédérations professionnelles, entreprises d'assurances, etc., etc.), a pris l'habitude de rendre publiques certaines informations dont il dispose. C'est ainsi, par exemple, que le site de Nielsen propose une analyse du secteur du commerce en Belgique, données qu'on ne trouve pas sur le site de StatBel (nouveau nom de l'Office belge de

1 Cette note est une version quelque peu étendue et remaniée d'une note initialement rédigée dans le cadre des travaux de l'Economic Prospective Club, qui ont porté début 2018 sur l'économie numérique (voir : http://moneystore.be/wp-content/uploads/doc/manifeste_transition_technologique_2018.pdf). Je remercie les économistes qui en font partie, de même que Paul-Marie Boulanger et Thomas Tombal, pour leurs apports directs et indirects à cette note ; j'assume bien sûr seul la responsabilité de son contenu final.

statistique).

L'accès restreint aux données collectées par Nielsen a posé problème au moment (début des années 90) des négociations entre le politique et les acteurs commerciaux concernés par le projet d'une éco-taxe sur les emballages jetables. C'est ainsi, par exemple, que la répartition entre les emballages consignés et les emballages jetables et l'évolution des ventes d'emballages consignés étaient des informations stratégiques que les acteurs/négociateurs publics ne possédaient pas.

Tout cela pour dire que les enjeux des données ne sont pas neufs :

1. Qui produit de l'information et dans quels buts ?
2. Quelle place des acteurs privés et quelle place des acteurs publics ?
3. Où et comment les données sont-elles stockées ?
4. Qui y a accès ? A quelles conditions et par quelles techniques ?

Est-ce à dire que la révolution numérique ne change pas grand-chose à ces questionnements ? Non, bien sûr, il y a trois évolutions majeures :

1. Le traitement d'informations par des dispositifs d'IA (intelligence artificielle), qui génère des questions spécifiques.
2. Les quantités énormes de données produites, qu'il faut contrôler, stocker, gérer et exploiter.
3. L'ampleur et le détail de données « personnelles ».

NB : Il sera certes beaucoup question de big data dans la suite de cette note, mais pas seulement ; de nombreuses sources de données, qui ne peuvent être qualifiées de big data, sont tout aussi utiles, en ce y compris en les mettant en perspective avec du big data.

Dans ce contexte, en pleine évolution, je vois cinq catégories d'enjeux :

1. le contrôle (de la qualité et de la pertinence) des données
2. la qualité/justesse des informations contenues dans les banques de données et de celles qui en découlent
3. l'accès à l'information
4. la numérisation de données qui ne le sont pas
5. quelques autres enjeux.

## 1. Le contrôle (de la qualité et de la pertinence) des données, notamment de celles injectées dans des dispositifs d'IA basés sur l'apprentissage, et des résultats qui en découlent

Par définition, ce qui sort comme analyses ou impulsions d'un dispositif d'IA basé sur l'apprentissage dépend étroitement des données qui sont injectées pour activer les réseaux de neurones. Tout utilisateur d'un tel dispositif devrait donc être en droit de connaître les données qui ont servi à alimenter le dispositif. L'exemple de la technologie de la reconnaissance faciale illustre bien cet enjeu.

Mais, d'une manière générale, il faut un réel contrôle de qualité sur les données mises en circulation, quelle qu'en soit l'origine, mais qui pourrait être plus rigoureux quand ces données sont produites par des intérêts privés.

## 2. D'une manière générale se pose la question de la qualité/justesse des informations contenues dans les banques de données et de celles qui en découlent

C'est en particulier le cas des données personnelles ; certes le droit de vérification et de correction existe dans notre législation et est renforcé par la mise en œuvre du Règlement européen sur la protection des données. Mais il n'est pas sûr que les utilisateurs de données personnelles qui leur ont

été transmises par l'opérateur initial contactent effectivement les « personnes concernées ». L'immense majorité d'entre nous ne connaît pas les très nombreuses banques de données commerciales évoquées par exemple dans l'article « Comment les entreprises surveillent notre quotidien ». Il faut donc renforcer les contrôles sur le respect de la lettre et l'esprit du règlement général sur la protection des données (RGPD) et, en même temps, rendre les démarches de consentement plus claires, plus faciles à décrypter , plus « user-friendly ».

Le contrôle de la qualité/justesse des données se pose d'une manière globale. La quantité d'informations rend difficile le contrôle de chaque donnée. Des altérations volontaires (ex : les logiciels truqués concernant les émissions de voitures) ou involontaires (erreurs de codage, différences de classements ou définitions...), sont possibles, compromettant dès lors les résultats issus du traitement de ces données. Le secteur financier est particulièrement concerné par cet enjeu parce qu'il est plus contrôlé que d'autres.

Cette problématique n'est pas spécifique aux big data. Mais la grande taille des banque de données comme leur utilisation diffuse (informations reprises par d'autres acteurs et/ou pour d'autres objectifs) nécessitent de mettre en place des garde-fous spécifiques comme des techniques de repérage d'erreurs.

Cet enjeu concerne à la fois le secteur des entreprises dans leurs choix stratégiques (par exemple les stratégies d'investissements, financiers et réels) et la puissance publique dans ses actions de contrôle et d'impulsion de politiques.

## 3. L'accès à l'information : pour une libre circulation des données, dans le respect de la vie privée

C'est pour moi le défi essentiel, parce que c'est la libre circulation des données qui permettra de contrôler leur qualité, de lutter contre la constitution de monopoles ou oligopoles, de susciter et d'alimenter de nouvelles activités, de booster la recherche scientifique et industrielle, d'éclairer et d'orienter l'action publique. Les difficultés d'accès à l'information prennent de nombreuses formes :

1. Les difficultés d'accès liées à la non-numérisation, non seulement d'archives passées mais aussi d'une partie de l'information produite aujourd'hui (voir aussi la section 4). Comment, par exemple, évaluer les politiques des CPAS en matière de soins de santé et de revenus d'intégration étudiants en l'absence d'informations numérisées.
2. Les difficultés d'accès liées aux manques de moyens des fournisseurs publics d'informations. Deux illustrations concernent les jugements des tribunaux, dont une loi de 2001 prévoit pourtant un accès intégral et gratuit. C'est ainsi que la banque de données appelée BelgiqueLex « n'est pas téléchargeable en une fois ce qui implique que les sources ne peuvent être traitées à grande échelle de manière efficace ». « Plus inquiétant encore, (la banque de données) Juridat ne contient qu'une partie infime de la jurisprudence prononcée à ce jour. Environ 160.000 décisions seulement, toutes années confondues, sont aujourd'hui (fin 2017) disponibles sur Juridat, ce qui correspond à 0,47% des jugements prononcés depuis la Seconde Guerre mondiale. » Autre illustration : deux recherches wallonnes relatives respectivement aux expulsions domiciliaires et aux garanties locatives n'ont pu se faire qu'en consultant des jugements sur papier de plusieurs cantons de justices de paix et en les numérisant. L'accès aux données d'un canton n'a pas été possible pour cause de désorganisation du « classement » (sic), un autre parce que pas de réponse du juge de paix à la sollicitation de consultation des archives !
3. Les difficultés d'accès liées à une interprétation trop stricte de la protection de la vie privée. Il s'agit bien d'interprétations abusives parce que, en soi, le RGPD n'interdit pas d'ouvrir l'accès à des données à caractère personnel. Certes, on sait que le *data-mining* peut permettre, dans certaines conditions, de relier des données anonymisées à des personnes ou ménages "connus". Mais il doit être possible de développer des techniques protectrices, par exemple en

intervenant à distance sur des données sans y avoir un accès direct. Notons à cet égard que des technologies se développent pour pouvoir utiliser les données directement dans les serveurs de l'organisation qui les cède ou les vend ce qui permet d'éviter les risque de rupture de confidentialité liés au transfert "physique" de données.

4. Les difficultés d'accès liées aux réticences des "propriétaires" de données, propriétaires commerciaux certes, ceux auxquels on pense le plus souvent, mais aussi les propriétaires agissant dans le secteur non marchand. C'est ainsi, par exemple, que Renaud Maes estime que « les statistiques sur l'origine sociale des étudiants universitaires font l'objet d'un véritable hold-up rectoral » (au travers du CREF – Conseil des recteurs francophones). Notons que, dans les faits, les grands accumulateurs de données peuvent avoir un intérêt stratégique (par exemple dans le cadre d'un échange de données) à diffuser une partie de leurs données. La liberté de circulation des données concernant des secteurs sensibles (contrôle aérien, pilotage des centrales nucléaires y compris les résultats d'exercices de simulation d'accidents, résultats et interprétations des examens des scanners, les données produites par les – futurs – centre de protonthérapie…) doit être totale et automatique.
5. La lenteur d'arrivée de certaines données les rend moins intéressantes une fois disponibles. Alors que la Flandre produit rapidement (en début d'année académique) des statistiques d'inscription dans l'enseignement supérieur, informations intéressantes pour le secteur mais aussi pour la compréhension de ce qui se passe sur le marché du travail des jeunes, les données concernant la Fédération Wallonie-Bruxelles sont publiées plusieurs années après.
6. Le non accès à des données privées mais d'intérêt général. Un exemple très parlant est celui de la non publication par des laboratoires de recherche de résultats négatifs ce qui entraîne, par exemple, le financement par d'autres acteurs privés et/ou publics de recherches dont on aurait pu savoir dès le début qu'il s'agissait d'impasses. Autre exemple : dans son livre consacré aux gènes, Adam Rutherford évoque la « nouvelle activité commerciale qu'est la généalogie génétique, qui vous permet, moyennant quelques centaines d'euros et des crachats dans une éprouvette, de disposer d'un aperçu de votre ADN. Les résultats sont à (son) sens d'un intérêt individuel accessoire, mais en recueillant ces échantillons, les entreprises en question (…) amassent un ensemble colossal de données sur le génome humain, qui dépasse largement les éléments à disposition pour la recherche scientifique universitaire. » Signalons à cet égard que lors de la consultation publique relative à la Directive PSI – Public sector information (Directive 2003/98 concernant la réutilisation des informations du secteur public, telle que modifiée par la Directive 2013/37), une large majorité de répondants a souhaité l'accès à des données privées (a fortiori quand leur constitution a été soutenue par les pouvoirs publics.
7. D'une manière générale de nombreux domaines de recherche dépendent aujourd'hui de l'accès à des banques de données spécifiques (celles relatives à des recherches menées par d'autres – ce qui est aussi nécessaire pour « contrôler » les résultats de recherche – ou des banques de données « généralistes » détenues par les GAFA, les banques, les détaillants, les assureurs, les mutuelles et beaucoup d'autres acteurs) mais plus encore de l'accès à des banques de données couplées. C'est notamment le cas en matière de santé. C'est probablement dans l'exploitation simultanée de banques de données de nature différente (ex : données médicales et données relatives à la consommation alimentaire et/ou des données relatives aux lieux de vie et/ou etc. ou données comportementales dans divers domaines et données relatives aux revenus et à la composition des ménages) que se trouvent les champs d'application les plus intéressants de l'IA et les possibilités les plus grandes de découvrir des liens insoupçonnés ou en tout cas de donner de la substance à de nombreuses hypothèses ou connaissances fragmentaires. Les études épidémiologiques, par exemple, gagneraient beaucoup à pouvoir utiliser de manière simultanée des banques de données variées, visant un ensemble de facteurs explicatifs potentiels (exposition à des pollutions locales, exercice de métiers dangereux, alimentation...). On peut imaginer, par exemple, qu'on aurait découvert plus tôt le lien entre coloration des cheveux et cancer du sein ou en tout cas qu'il serait plus

facile et plus rapide de confirmer la robustesse statistique de ce lien. Il semble que l'exploitation de nombreuses données par des dispositifs d'IA peuvent aussi aider à la détection de de signaux avant-coureurs (par exemple en matière de santé : détérioration de la vue, des capacités cognitives, etc.). Autre exemple : de nombreuses recherches sémantiques, sociologiques, etc., bénéficieraient grandement de l'accès aux contenus numérisés des médias et de l'édition de livres. L'accès libre à des données sensibles est aussi une condition indispensable pour alimenter et donc renforcer des experts indépendants, qui manquent cruellement dans de nombreux dossiers où les intérêts privés disposent de moyens très importants, dont l'accès à des informations connues d'eux seuls. Ici aussi, une large majorité de répondants à la consultation publique relative à la Directive PSI (voir point 6) soutient un accès ouvert aux données scientifiques.

8. L'accès différencié de candidats ou partis à des données stratégiques (pour des raisons financières et/ou partisanes) risque de poser autant de problèmes démocratiques que l'inégal accès à des ressources financières.
9. De même l'accès difficile à des documents internes à l'administration ou aux cabinets ministériels est un frein à la vigueur démocratique. L'existence même de certains rapports n'est pas connue. Difficile donc d'activer son droit à la communication. Ici aussi une communication pro-active et dès la commande des études vers les assemblées concernées devrait être imposée.
10. L'action publique aussi a besoin d'informations détenues dans des banques de données privées. Les données collectées par UBER (ou entreprises équivalentes, y compris les sociétés de taxis) ou les opérateurs de téléphonie mobile peuvent être d'une grande utilité dans la définition et le suivi d'une politique de mobilité, au niveau local comme à des niveaux supérieurs. Notons à cet égard qu'UBER a déjà des accords avec de nombreuses villes de par le monde pour partager ses données sur les flux de mobilité.
11. La création de monopoles liée au non-accès aux données pour des « entrants » (potentiels) sur le marché. C'est ainsi que des start-up peuvent être limitées dans leur développement parce qu'elles n'ont pas accès aux données nécessaires pour "entraîner" leurs dispositifs d'IA. Autre exemple, un recours croissant à des contrats d'assurances personnalisés, tenant compte de l'historique des risques et comportements des clients, rend plus difficile la concurrence par de nouveaux acteurs dès lors que ceux-ci ne disposent pas du même niveau d'information. C'est aussi le cas des services avec souscription (Netflix, HelloFrsch...) qui semblent en hausse.Un exemple donné par The Economist :
12. L'accès à des résultats de la mise en œuvre de dispositifs d'IA ou d'applications spécifiques est aussi un enjeu en matière de diffusion de l'information. C'est ainsi, par exemple, que les résultats de démarches d'IA effectuées sur des simulateurs de vol peuvent intéresser tous les acteurs de la sécurité aérienne. Autre illustration : les nombreuses applications qui sont mises en œuvre en matière de santé génèrent elles-mêmes des flux d'informations potentiellement intéressantes.
13. Le coût trop important d'accès à des revues scientifiques, parfois même à des articles basés sur des financements publics, est un frein pour la recherche, en tout cas pour les chercheurs/pays les plus pauvres. Il est par ailleurs anormal, me semble-t-il, de devoir payer (une partie) des informations récoltées par, par exemple, l'OCDE ou l'Agence internationale de l'énergie, alors que ces agences sont financées par de l'argent public.

C'est une véritable politique d'open-data proactive qu'il faut mettre en route. Les données ne peuvent plus être considérées comme une propriété strictement privée. On en est très loin... Des voies originales sont pourtant possibles pour concilier les intérêts des uns et des autres en matière d'accès à des données stratégiques. « Le partage doit devenir le ciment idéologique du camp du progrès. » « Les données bénéficient aujourd'hui majoritairement à de très grands acteurs. Ce n'est qu'au prix d'un plus grand accès et d'une meilleure circulation de ces données, pour en faire bénéficier les pouvoirs publics,

mais aussi les acteurs économiques plus petits et la recherche publique, qu'il sera possible de rééquilibrer les rapports de forces. »

Comme le souhaitent les économistes de l'Economic Prospective Club, « il convient de protéger l'usage qui est fait des données plutôt que les données en tant que telles. »

Signalons pour conclure cette section l'idée développée par des économistes de Stanford, de Columbia et de Microsoft de considérer les données personnelles comme la propriété de ceux qui les génèrent.

## 4. La numérisation de données qui ne le sont pas

De nombreuses données, en particulier de la responsabilité du secteur public, ne sont pas numérisées. Parfois parce qu'elles sont déclarées non numérisables (n'étant pas ordonnées, structurées, compilées, même pas sous format papier – voir remarque plus loin) mais le plus souvent par manque d'ambition statistique.

Trois exemples parmi (malheureusement) beaucoup d'autres :

1. Les discussions sur la politique du logement profiteraient grandement d'une numérisation du contenu des baux à loyer dont le dépôt à l'enregistrement est en principe obligatoire. Mais à ce jour ce n'est pas encore le cas ! Difficile d'imaginer une volonté de traiter ces informations alors qu'il n'y a même pas de volonté de faire respecter l'obligation de l'enregistrement. Un pas en plus pourrait être fait en complétant le contenu du bail avec quelques données essentielles (nombre de pièces, surface totale habitable...) ouvrant la voie à des analyses plus subtiles de la formation et de l'évolution des loyers. En France, le gouvernement Macron « envisage la création d'un bail numérique, sans papier, souhaitant que les gérants en transmettent automatiquement les données au fichier national ». « Il n'en est pas question », s'insurge Jean-Marc Torrollion, président de la Fédération nationale (française) de l'immobilier, « nous ne communiquons pas les données personnelles de nos clients propriétaires et locataires, même si nous consentons à alimenter en données non nominatives les observations de loyers que le gouvernement veut généraliser » dans la perspective de renforcer/étendre les Observatoires des loyers. Un observateur (anonyme) du secteur du logement public déplorait qu'il y avait « pas mal de méfiance de la part des SLSP (Sociétés de logement de service public ou Sociétés de logements sociaux dans le langage courant) dans le transfert de données vers la SWL (Société wallonne du logement) », ce qui empêche de faire des analyses fines et politiquement utiles sur l'évolution des loyers dans le logement social.
2. L'AViQ (Agence wallonne pour une vie de qualité) dispose des rapports d'activités remis par les services d'aides familiales. L'absence totale d'exploitation des données qu'ils contiennent est une des nombreuses raisons pour lesquelles l'AViQ ne peut éclairer les débats sur la mise en œuvre d'une assurance-autonomie en Wallonie.
3. Très peu de CPAS (wallons en tout cas) ont une politique de structuration et d'exploitation des très nombreuses informations dont ils disposent (parcours de vie, études, enfants, loyers, autres revenus, taux d'emploi suite à des formatons ou à un Article 60...) au travers des dossiers individuels. Une exploitation intelligente de ces données permettrait un pilotage budgétaire plus rigoureux, de repérer l'émergence de nouveaux publics, d'évaluer ses politiques d'insertion, etc. Mieux encore : un effort concerté de collecte et de traitement de ces données au niveau régional et/ou national fournirait des données intéressant plusieurs acteurs (gouvernements, chercheurs, associations de lutte contre la pauvreté).

Rencontrer cet enjeu implique une profonde transformation culturelle dont on peine à voir l'amorce aujourd'hui. En tout état de cause une excuse facile – les données ne sont pas structurées rendant leur numérisation et leur organisation impossibles ou trop coûteuses – ne tient plus la route : une des caractéristiques des outils d'IA est précisément de pouvoir lire et interpréter des données mêmes non structurées.

## 5. D'autres enjeux

J'en vois six principaux :

1. Les problèmes liés au stockage des données. Deux grandes difficultés doivent ici être mises en évidence :
   - La durabilité des techniques de stockage.
   - Les choix inévitables qu'il faudra faire devant l'impossibilité de tout stocker. « Pérenniser l'éphémère » (le thème d'un colloque organisé par l'UCL en mai 2016), voilà le défi des archivistes d'aujourd'hui.
2. La formation d'un nombre suffisant de spécialistes (informaticiens, utilisateurs, statisticiens...) des questions (techniques, éthiques, culturelles...) liées aux big-data.
3. Le risque de *lock-in* (qu'on pourrait traduire par blocage, enfermement) socio-culturel lié aux « prescriptions » découlant de l'exploitation d'informations par des dispositifs d'IA. C'est ainsi que les "suggestions" proposées par des sites d'informations, commerciaux ou de vidéos à la demande risquent d'enfermer leurs utilisateurs et de les conforter dans leurs visions et leurs habitudes. De même l'exploitation de données permettant aux éditeurs, aux producteurs, etc., de « déterminer » ce qui fait le succès de films, de livres ou d'autres produits artistiques risque de déboucher sur des prescriptions rendant plus difficiles (encore ?) l'émergence d'œuvres ou d'approches originales. Les précautions à prendre pour que le recours de techniques d'IA produise des résultats et des aides à la décision pertinents. Le cas de la justice dite prédictive est à cet égard particulièrement parlant. C'est ainsi qu'Alexandre de Streel propose, à juste titre, la mise en place de garde-fous pour "augmenter" les juges sans "diminuer" la justice : « d'abord, un maximum de décisions judiciaires passées doivent être disponibles en accès ouvert (tout en respectant la vie privée des parties) pour pouvoir éduquer au mieux les algorithmes ; ensuite, l'expertise en intelligence artificielle doit être renforcée dans les prétoires pour permettre aux avocats de contester et aux juges d'évaluer la qualité des prédictions qui leur sont soumises ; ensuite, une transparence doit être assurée sur les inputs (les données ), le processus (les fonctions d'optimisation) et les résultats (les prédictions) de l'algorithme pour pouvoir en contrôler les erreurs et les biais ; en outre, les algorithmes utilisés pour rendre la justice devraient être développés par les pouvoirs publics ou, au moins, certifiés par eux ; enfin, et plus fondamentalement, l'algorithme ne devrait jamais se substituer aux juges mais seulement l'aider dans sa mission éminemment humaine de rendre la justice. » Si cette dernière condition n'était pas remplie, et la tentation de s'appuyer sur une décision « toute faite » est grande, on court le risque majeur en justice d'une jurisprudence bloquée, non évolutive, non adaptative.
4. Les nombreux échecs ou en tout cas retards, explosion des coûts... liés à la mise en place de banques de données. Ex : « La banque de données unique du secteur de l'énergie s'enlise », titrait l'Echo du 28.09.17. On pense aussi aux très nombreux retards liés à la mise en œuvre d'un nouveau logiciel pour traiter les demandes de personnes handicapées, difficultés qui ne sont d'ailleurs pas niées par le SPF Sécurité sociale.
5. La nécessaire réforme du droit d'auteur pour « permettre d'autoriser les pratiques de fouille de texte et de données (*text and datamining*) dans un objectif de compétitivité de la recherche publique. »
6. L'intégration de données micro-économiques du big data dans l'analyse macro-économique et le suivi conjoncturel est encore à développer. Dans cette perspective signalons, par exemple, une note d'analyse de StatBel qui « explique l'utilisation du webscraping dans l'indice des prix à la consommation. En quoi consiste le webscraping ? À quoi ressemblent ces données? Et comment Statbel (DG Statistique – Statistics Belgium) traite-t-elle ces données ? Cette analyse décrit les études de cas réalisées et les différentes méthodes de calcul de l'indice sur la base du webscraping qui ont été testées. Elle présente également un certain nombre d'algorithmes

pour l'apprentissage automatique. L'apprentissage automatique est l'étude dans le cadre de laquelle des algorithmes sont créés pour que les machines/ordinateurs/programmes puissent "apprendre" eux-mêmes. »

## 6. Pistes d'action

Il m'apparaît qu'il faut centrer l'action publique autour de cinq axes (en plus des politiques mises en place ou souhaitables en matière de protection des données personnelles et des politiques visant à diminuer l'empreinte écologique de la filière des données, politiques pas abordées dans cette note) :

1. Développer une culture statistique, portant sur les données en général, big data bien sûr y compris mais pas seulement. Tous les cursus académiques et les formations qui ponctuent la vie professionnelle doivent développer un intérêt pour la statistique et les statistiques comme compétence transversale. Il s'agit de construire à la fois une capacité critique et une vision positive de la collecte et de la structuration de données (sur l'environnement et l'interne de l'organisation, privée comme publique) pour définir des orientations stratégiques et piloter l'action. Cette culture statistique doit, en particulier, viser tous ceux qui seront amenés à utiliser des outils prédictifs.
2. Susciter/encourager la production et la diffusion d'informations d'intérêt général. C'est ainsi, par exemple, que la publication de données concernant les différences salariales hommes-femmes par les organisations britanniques (entreprises, ONG…) de 250 employés ou plus par une loi de 2016, a déjà eu des retombées intéressantes, à la fois sur la connaissance de cette inégalité et sur la prise de conscience des organisations, qui, pour certaines d'entre elles, ont déjà développé des politiques internes pour les corriger. Cette volonté passera par diverses actions : obligations de divulgation (des données et de leur caractéristiques) et d'accès (ex : statistiques d'inscriptions dans l'enseignement supérieur, données des opérateurs de téléphonie mobile, d'opérateurs internet…), soutien à des plates-formes sectorielles visant à la mutualisation de données, évolutions de la loi sur les droits d'auteur, etc.
3. Activer les moyens nécessaires pour que des organisations belges (entreprises, pouvoirs publics, universités) puissent participer, ou continuer à la faire, à des activités, publiques et privées, générant des données dans des secteurs stratégiques (exploration spatiale, exploitation des océans, santé personnalisée, véhicules autonomes…). On notera à cet égard que les partenaires de telles collaborations seront de plus en plus des entreprises privées investissant stratégiquement dans des activités de récolte de données nouvelles dans l'espoir de se rendre incontournable dans le domaine d'activité choisi. C'est par exemple le cas en matière de recherche océanique. Ceci dit, la taille n'est pas, au départ en tout cas, une contrainte pour arriver à développer des activités susceptibles d'intéresser des partenaires économiques, nationaux ou internationaux. L'exemple, parmi beaucoup d'autres, de la start-up Agroptimize le montre. « Cette start-up d'Arlon utilise le big data pour optimiser l'agriculture de demain. Basé sur des recherches universitaires, le modèle proposé vise à prévenir les maladies tout en utilisant moins de pesticides, à rendre la terre plus fertile et à maximiser les rendements agricoles. »
4. Développer l'intégration d'informations (données et schémas comportementaux) issues de l'analyse du big data dans l'analyse macroéconomique et le suivi conjoncturel, notamment via le web scraping. Par exemple, l'exploitation du big data peut permettre de faire du *now-casting*, à savoir de capter et décrire quasiment de manière immédiate des flux financiers ou autres indicateurs conjoncturels. D'une manière générale l'exploitation du big data, en particulier à une époque de déclin dans les taux de réponse aux enquêtes, peut aider à comprendre, mieux (plus vite et avec plus de détails) qu'aujourd'hui, les évolutions des comportements susceptibles d'éclairer les choix politiques dans les divers domaines d'action (logement, santé, mobilité…) des gouvernements.

5. Développer la constitution de banques de données et leur exploitation. Voici, à titre exemplatif, quelques axes qui pourraient structurer une politique des données, au niveau régional, au niveau des communautés ou au niveau fédéral :

   - Dans le domaine de la santé, la constitution d'une banque de données portant sur les analyses de sang ; son couplage à d'autres banques de données relatives à la santé (en particulier les informations détenues par les mutuelles) devrait permettre, via des techniques d'IA, de tirer un maximum d'enseignements et d'orientations en matière de recherche, privée et/ou publique.
   - Dans le domaine de l'enseignement, une banque de données, à constituer progressivement, reprenant depuis l'enseignement fondamental les parcours des élèves/étudiants, avec le plus de caractéristiques possibles pour mieux qu'aujourd'hui éclairer les choix politiques sensibles, présents dans les débats sur le Pacte d'excellence.
   - Dans le domaine de la mobilité, il faut forcer le couplage, via des techniques d'IA, de données issues d'opérateurs privés (notablement les informations détenues par les opérateurs de téléphonie mobile), de celles dont disposent les sociétés de transport (qui s'amélioreront encore quand les cartes du type MOBIB seront généralisées), des flux mesurés localement (il faut ici stimuler les autorités locales pour les produire) et à d'autres niveaux (par exemple les données régionales résultant de la mise en place d'une redevance au km pour les camions, étendue peut-être un jour aux véhicules légers) et d'informations obtenues par web scraping. Signalons ici une initiative non-marchande et collaborative en matière de mobilité, la plate-forme Catalogue.
   - Dans le domaine social, où les données détenues par la Banque Carrefour de la Sécurité Sociale constituent un excellent point de départ, la mise en perspective et l'analyse des parcours, grâce à l'exploitation d'autres données (formations initiales et au long de la carrière, interruptions de carrière, secteurs d'activité, mobilité géographique …), pourraient éclairer de nombreux débats concernant les politiques sur le marché du travail (ex : impact des sanctions chômage, valorisation salariale des formations, recours à des formation(s) tout au long de la vie…).
   - Enfin, dans le domaine social toujours, l'analyse fouillée des dossiers et décisions des CPAS dans le domaine des aides sociales devrait permettre de mieux comprendre les "logiques" à l'œuvre, leurs évolutions et leurs différences entre CPAS. Par là même on pourrait, comme en matière de justice prédictive, mettre en place une aide à la rédaction des dossiers sociaux et à la décision (en laissant bien sûr la décision ultime aux conseillers élus) et rapprocher les pratiques entre CPAS, dans le but d'assurer l'équité entre bénéficiaires.

Au-delà de ces cinq axes, il va de soi que les pouvoirs publics doivent s'impliquer dans des efforts nationaux et internationaux pour rencontrer d'autres défis évoqués dans cette note (formations de professionnels, adaptations du droit d'auteur, contrôle de la qualité des données, l'accès à des données privées par les pouvoirs publics, etc.).

J'ai aussi conscience que de nombreuses démarches sont déjà prévues ou en cours. Je pense par exemple à l'appel d'offre lancé récemment (début 2018) par la SOFICO, le gestionnaire du réseau routier wallon, afin de mettre en place « un système de gestion dynamique des données et de ses interfaces pour le centre Perex. » Mais de nombreux domaines de l'action publique wallonne, notamment les pôles de compétitivité, gagneraient à passer à la vitesse supérieure en matière de collecte et d'exploitation des données nécessaires à l'ère numérique, par exemple en systématisant des démarches de web scraping pour alimenter la veille technologique et économique.