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

par Philippe DEFEYT[1] *– avril 2018*

« Je suis loin sans doute de prétendre que quelques tableaux numériques isolés peuvent suffire pour déterminer complètement tous les éléments si compliqués de nos sociétés modernes. Il faudrait, pour remonter des effets aux causes, ou pour conclure de ce qui est à ce qui sera, avoir égard à un ensemble de circonstances qu'il n'est point donné à l'homme de pouvoir embrasser : de là, la nécessité de négliger toujours, dans toute espèce d'appréciation, un certain nombre de circonstances dont il aurait fallu tenir compte. De là aussi l'absurdité des résultats auxquels conduit souvent cette énumération incomplète, ou le trop d'importance qu'on attache à un élément qui ne devrait être considéré que comme secondaire. La mauvaise foi pourra même porter à ne choisir dans une série de résultats, que ceux qui sont favorables au principe qu'on voudrait faire prévaloir, en passant sous silence ceux qui lui seraient contraires : et c'est ainsi, comme on l'a fort bien observé, que tout pourrait se prouver par les nombres de la statistique. Mais de ce que l'observation est difficile et de ce qu'il existe des ignorants ou des hommes de mauvaise foi, faut-il en conclure qu'on doit rejeter la statistiques ? Non sans doute (...) »

Adolphe QUETELET[2]

« La puissance publique doit amorcer de nouveaux modes de production, de collaboration et de gouvernance sur les données, par la constitution de "communs de la donnée". »[3]

Saint Isidore, priez pour nous[4]

---

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.

2 A. QUETELET, « Recherches statistiques sur le Royaume des Pays-Bas », Mémoire lu à la séance de l'Académie du 6 décembre 1828, Bruxelles, M. Hayez, Imprimeur de l'Académie royale, 1829, p. III

3 Cédric Villani, « Donner un sens à l'intelligence artificielle – Pour une stratégie nationale et européenne », Rapport de la mission parlementaire confiée par le Premier Ministre Edouard Philippe (Voir : https://www.aiforhumanity.fr/)

4 « Le Service d'Observation d'Internet, animé par le Conseil Pontifical pour les Communications Sociales, a réalisé une enquête dans différents milieux du monde de l'ordinateur et de l'espace cyber pour tenter de découvrir quel saint reflète le mieux les inquiétudes et les idéaux des informaticiens. Le saint patron le plus apprécié parmi les professionnels des nouvelles frontières de la technologie est saint Isidore, né en 556 à Séville (Espagne). "Le saint qui a écrit les étymologies" (une sorte de base de données)", conclut l'étude du Service d'Observation d'Internet, a donné à son œuvre un structure qui ressemble énormément au concept de la base de données. Il a conçu un système de pensée que l'on qualifierait aujourd'hui de système par flash, très moderne. Il s'agit d'une méthode particulièrement efficace grâce à laquelle on se sent immédiatement identifié. Saint Isidore a réalisé son œuvre en faisant un effort de cohérence énorme, pour qu'elle soit complète et que ses éléments soient complémentaires les uns des autres. Mais ce n'est pas le seul élément qui rapproche saint Isidore des informaticiens. Le saint de Séville était en avance sur son temps. Il fut un pont culturel entre l'Antiquité et le Moyen Âge. C'est aussi pour cela que nous nous sentons proches de lui car nous nous trouvons au tournant d'une nouvelle étape de l'Histoire, a répondu l'un des informaticiens interrogé par le Service d'Observation promu par le Vatican. (...) Isidore de Séville a été proposé, en 2001, comme saint patron des informaticiens, des utilisateurs de l'informatique, de l'Internet et des internautes La fête de Saint Isidore est le 4 avril » Source : http://www.self-reliance.be/Saint-ISIDORE-Patron-des

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.)[5] mais aussi à l'observation des marchés et des prix pratiquée par les marchands d'Assur du temps de la civilisation mésopotamienne[6].

Pour d'autres préoccupations, on pense évidemment aux recensements de la population. « Le recensement (du latin *recensere*, passer en revue) est une opération statistique de dénombrement d'une population. Les recensements démographiques existent depuis l'Antiquité (Chine, Égypte, Hébreux que la Bible mentionne à plusieurs reprises ; Rome). Ils ne sont mis en œuvre de façon systématique qu'à partir du XVIe siècle et plus encore avec l'avènement de l'État-nation dont ils servent divers objectifs : notamment la conscription militaire, la répartition de l'impôt, la connaissance du nombre et des richesses de la population. »[7]

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[8]).

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. Voici la manière dont il présente son action : « For 93 years Nielsen services have been making the FMCG (Fast Moving Consumer Goods) market visible and comprehensible for manufacturers and retailers. Nielsen is the world's leading provider of information and market analysis in the consumer and service sectors thanks to the quality of its data, the accessibility of its analyses, the expertise it has built up over more than 93 years of practical experience and the rapidity with which it communicates its findings. (...) Active in Belgium since 1954 - Over 200 regular clients - Studies more than 400 product categories. »[9]

Nielsen, comme beaucoup d'autres acteurs économiques (fédérations professionnelles, entreprises d'assurances, etc., etc.), a pris l'habitude de rendre publiques certaines informations. C'est ainsi, par exemple, que l'on trouve les informations suivantes (voir encadré en haut de la page suivante) sur le site de Nielsen Belgique, données qu'on ne trouve pas sur le site de StatBel ! (un rapport complet[10] est aussi accessible, après avoir, bien sûr, rempli un formulaire...).

---

5 http://journals.openedition.org/comptabilites/1437 ; voir aussi : https://www.numilog.com/package/extraits_pdf/e272118.pdf

6 http://www.salama-mag.com/salamamag/irak-syrie-tradition-commerciale-ancestrale-heritee-de-mesopotamie/

7 https://fr.wikipedia.org/wiki/Recensement_de_la_population

8 « Un sondage rémunéré sur internet c'est quoi ? C'est une enquête que vous remplissez depuis votre ordinateur, vous permettant ainsi de gagner jusqu'à 5€ par sondage en donnant votre avis. Un vrai travail à domicile, permettant de gagner de l'argent facile ! *Facealacrise* a sélectionné pour vous les sites d'enquêtes pour la Belgique les plus rémunérateurs. » (https://www.facealacrise.be/sondage-remunere/) Voir aussi, par exemple : http://www.sondagesremuneres.fr/

9 Nielsen, « GROCERY UNIVERSE 2017 - RESULTS OF THE 55TH INVENTORY OF RETAIL GROCERY IN BELGIUM, DRAWN UP BY NIELSEN » http://www.nielsen.com/be/en/insights/reports/2017/nielsen-grocery-universe-2017.html)

10 http://www.nielsen.com/content/dam/nielsenglobal/eu/docs/reports/nielsen-grocery-universe-2017.pdf

Nielsen Grocery Universe 2017 – Belgium[11]

FMCG and Retail | 10-10-2017

Every year Nielsen creates an overview of the changes in the grocery world. In our Nielsen Grocery Universe 2017 report, you'll find the evolution of a number of retailers, the share of store types, as well as regional and international comparisions (comparing the results till the end of 2016).

In Belgium, the most important conclusions are:

•There were 7163 grocery stores in Belgium at the end of 2016, an increase of 2 stores than in 2015.

•The turnover of the Grocery Universe rose to EUR 25.8 billion in 2016. This means a +0.6% increase compared to the year before. In 2016, the inflation rate was 2.0%, which means that the turnover in constant terms has declined by -1.4%.

•Medium-sized supermarkets (F2) increased their market share to 30.4% which represents an increase of +0.8pt versus 2015.

•Hard Discounters saw a decrease in market share by -0.3 points--now showing a 15.4% market share in Belgium.

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 ?

Rien de neuf sous le soleil

Dans son livre consacré aux gènes[12], le scientifique Adam Rutherford évoque la constitution d'une banque de données anthropométriques par Francis Galton[13], polymathe par excellence et souvent présenté comme le père de l'eugénisme.

Francis Galton avait installé dans le cadre de l'Exposition internationale d'hygiène et d'éducation (Londres, 1884) « un laboratoire d'anthropométrie pour lequel les visiteurs payaient 3 anciens pence (environ 0,90 euro actuels) l'entrée. Ces derniers remplissaient anonymement une carte (comprenant un carbone pour que Galton archive les données) renfermant leurs informations personnelles. "Le but de ce laboratoire, a écrit le scientifique dans le *Journal of the Anthropological Institute* à la fin de l'exposition, était de montrer au public la simplicité des instruments permettant de mesurer et de consigner les principales caractéristiques physiques de l'homme." (...)

Quel programme intelligent ! Aujourd'hui les gens déboursent de l'argent pour obtenir des informations sur leur ADN, tout comme dans le laboratoire d'anthropométrie de Galton. Notre nombrilisme fait que nous livrons volontiers nos paramètres biométriques et que nous sommes prêts à payer pour ce privilège. En tout, lors de cette exposition, 9.337 personnes se sont acquittées de 3 anciens pence pour être mesurées par Galton, qui a ainsi amassé des données, accumulant les similitudes et les différences entre individus sur la base de ces caractéristiques.» (pp.212-213)

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.

11 http://www.nielsen.com/be/en/insights/reports/2017/nielsen-grocery-universe-2017.html

12 Adam Rutherford, "ADN - Quand les gènes racontent l'histoire de notre espèce", Larousse, 2018

13 Voir : https://fr.wikipedia.org/wiki/Francis_Galton

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.

Une définition parmi d'autres du big data[14]

« Although no agreed-on definition of big data currently exists, the term is often characterized by the 3Vs—high-volume, high-velocity, and high-variety. High-volume refers to increasing exabyte data generated by machines, networks, and human interaction; high-velocity refers to the speed at which data are created, processed, and stored; and high-variety relates to the range and complexity of data types and sources. Data sets are so large and complex that traditional data-processing applications become insufficient to capture, store, and analyze the data. Instead, a network of human skills, advanced technologies, and data access infrastructure are essential to handle big data. This is a key challenge for statisticians and policymaking organizations seeking to incorporate big data in their toolkits."

The list of Vs has grown over time, emphasizing both opportunities and challenges that companies and organizations face when incorporating big data into their existing business operations. Veracity refers to the noise and bias in the data as one of the biggest challenges to bringing value and validity to big data. Volatility refers to changing technology or business environments in which big data are produced, which could lead to invalid analyses and results, as well as to fragility in big data as a data source."

Classement des sources de big-data proposé par le FMI[15]

1. Social Networks (human-sourced information)
- 1100. Social Networks: Facebook, Twitter, *LinkedIn*
- 1200. Blogs and comments
- 1600. Internet searches *on search engines (Google)*
- 1700. Mobile data content: text messages, *Call Detail Record, Data Detail Record, Location update, Radio coverage updates*
- *Online news*

2. Traditional Business systems (process-mediated data)

21. Data produced by public agencies
- *Administrative data*

22. Data produced by businesses
- 2210. Commercial transactions
- 2220. Banking/stock records
- 2230. E-commerce
- 2240. Credit cards
- *Business websites*
- *Scanner data*

3. Internet of Things (machine-generated data)

31. Data from sensors
- 311. Fixed sensors
  - 3111. Home automation
  - 3112. Weather/pollution sensors
  - 3113. Traffic sensors/webcam
  - 3114. Scientific sensors
- 312. Mobile sensors (tracking)
  - 3121. Mobile phone location *(GPS)*
  - 3122. Cars
  - 3123. Satellite images

14 Cornelia L. Hammer, Diane C. Kostroch, Gabriel Quirós, and STA Internal Group, "Big Data: Potential, Challenges, and Statistical Implications", IMF Staff Discussion Note, IMF, September 2017 (voir : http://www.imf.org/~/media/Files/Publications/SDN/2017/sdn1706-bigdata.ashx)

15 IMF Paper, op.cit.

1. Le contrôle (de la qualité et de la pertinence) des données, notamment de celles injectées dans des dispositifs d'IA basés sur le *deep-learning*, et des résultats qui en découlent

Par définition, ce qui sort comme analyses ou impulsions d'un dispositif d'IA basé sur le *deep-learning* (l'apprentissage profond) 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 (voir encadré ci-après).

Face-recognition technology

Computer programs recognise white men better than black women

Biased training is probably to blame

SOFTWARE that recognises faces has bounded ahead in recent years, propelled by a boom in a form of artificial intelligence called deep learning (see article). Several firms now offer face recognition as a commercial service, via their respective clouds. The ability to recognise in faces such things as an individual's sex has improved too, and this is also commercially available.

The algorithms involved have, however, long been suspected of bias. Specifically, they are alleged to be better at processing white faces than those of other people. Until now, that suspicion has been unsupported by evidence. But next week, at Fairness, Accountability and Transparency, a conference in New York, Joy Buolamwini of the Massachusetts Institute of Technology will present work which suggests it is true.

Ms Buolamwini and her colleague Timnit Gebru looked at three sex-recognition systems, those of IBM, Microsoft and Face++. They tested these on a set of 1,270 photographs of parliamentarians from around the world and found that all three classified lighter faces more accurately than darker ones. All also classified males more accurately than females. IBM's algorithm, for example, got light male faces wrong just 0.3% of the time. That compared with 34.7% of the time for dark female faces. The other two systems had similar gulfs in their performances. Probably, this bias arises from the sets of data the firms concerned used to train theirsoftware. Ms Buolamwini and Ms Gebru could not, however, test this because those data sets are closely guarded.

IBM has responded quickly. It said it had retrained its system on a new data set for the past year, and that this had greatly improved its accuracy. When testing the new system on an updated version of the set of politicians Ms Buolamwini and Ms Gebru had used, the firm said it now achieved an error rate of 3.46% on dark-skinned female faces—a tenth of that the two researchers had found using the existing system. For light-skinned males the error rate also fell, to 0.25%.[16]

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 »[17] [18]. 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

16 Cet article a été publié par The Economist du 17.02.18 sous le titre"Garbage in. Garbage out"

17 Voir : https://framablog.org/2017/10/25/comment-les-entreprises-surveillent-notre-quotidien/

18 Voir aussi : « Getting to know you - Everything people do online is avidly followed by advertisers and third-party trackers », The Economist, 11.09. 14

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[19] parce qu'il est plus contrôlé que d'autres[20].

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.

Le japonais Toray touché à son tour par un scandale de données falsifiées, l'action chute[21]

Le groupe japonais de textiles techniques et fibres de carbone Toray a déclaré (le mardi 28 novembre 2017) être confronté à son tour à une affaire de données falsifiées de certains de ses produits, comme ses compatriotes Mitsubishi Materials et Kobe Steel avant lui. "Il s'est avéré que des données concernant le contrôle qualité de produits ont été incorrectement réécrites" dans la filiale Toray Hybrid Cord, a-t-il déclaré dans un communiqué, alors que le PDG du groupe, Akihiro Nikkaku, tenait une conférence de presse à Tokyo organisée en dernière minute.

Les produits en question sont des cordages pour pneumatiques et autres fils tramés spéciaux à usage industriel, selon le communiqué.

Des falsifications de contrôles de qualité dans cette filiale ont été identifiées entre avril 2008 et juillet 2016, affectant 13 sociétés clientes, a précisé le groupe.

Ces falsifications ne remettent pas en cause la qualité et la sécurité des produits, a toutefois assuré Toray, affirmant également n'avoir pas détecté d'autres mauvaises pratiques dans les autres activités du groupe.

Le titre du groupe à la Bourse de Tokyo perdait plus de 3,5% vers 11H21 (02H21 GMT), à 1.065 yens, après avoir chuté de plus de 8% juste après la révélation de cette affaire en milieu de matinée, tandis que l'indice Nikkei gagnait 0,16% à la même heure.

Plusieurs groupes nippons ont déjà été éclaboussés par des scandales similaires ces derniers mois, notamment le sidérurgiste Kobe Steel et tout récemment Mitsubishi Materials via plusieurs de ses filiales.

Les constructeurs automobiles japonais Nissan et Subaru ont aussi été touchés ces derniers mois par des irrégularités de procédures dans l'inspection finale de leurs véhicules produits au Japon et destinés à ce seul marché.

Too little information

GE's flow of financial information has become fantastically muddled[22]

If they are to save the firm, General Electric's bosses and board need far better information

(...) The curse of rotten information can strike companies, too. That seems to be the case with General Electric (GE), which has had a vertiginous fall. Its shares, cashflow and forecast profits have dropped by about 50% since 2015. On January 16th it disclosed a huge, $15bn capital shortfall at its financial arm due to a revision in insurance reserves. And on January 24th it revealed a $10bn loss for the fourth quarter. In its core industrial arm, returns on capital have sunk from 20% in 2007 to a puny 5% in 2017.

(...) Schumpeter's theory is that GE's flow of financial information has become fantastically muddled. There is lots of it about (some 200 pages are released each quarter) and it is audited by KPMG. But it offers volume and ambiguity instead of brevity and clarity. It is impossible—certainly for outsiders, probably for the board, and

19 Voir, par exemple : http://www.revue-banque.fr/management-fonctions-supports/article/qualite-des-donnees-comme-vecteur-competitivite

20 Voir, par exemple, pour la Belgique le rôle de la BNB : https://www.nbb.be/fr/supervision-financiere/controle-prudentiel/domaines-de-controle

21AFP Publié in La Libre le mardi 28 novembre 2017

22 Cet article a été publié par The Economist du 27.01.18 sous le titre "The fog of war"

possibly for Mr Flannery—to answer central questions. How much cashflow does GE sustainably make and where? How much capital does it employ and where? What liabilities must be serviced before shareholders get their profits?

Perhaps GE has a better, parallel accounting system that it keeps under wraps. But the public one reveals eight problems. First, it has no consistent measure of performance. This year it has used 18 definitions of group profits and cashflow. As of September 2017, the highest number was double the average one. There is a large gap between most measures of profits and free cashflow.

Second, GE's seven operating divisions (power, for example, or aviation) are allowed to use a flattering definition of profit that excludes billions of dollars of supposedly one-off costs. Their total profits are almost twice as big as the firm's. It is the corporate equivalent of China's GDP accounting, where the claimed outputs of each province add up to more than the national figure.

Third, GE does not assess itself on a geographical basis. Does China yield solid returns on capital? Has Saudi Arabia been a good bet? No one seems to know. This is unhelpful, given that the firm does half its business abroad and that the long-term decline in returns has taken place as the firm has become more global.

Fourth, GE pays little attention to the total capital it employs, which has ballooned by about 50% over the past decade (excluding its financial arm). Its managers rarely talk about it and have set no targets. It is unclear which parts of the firm soak up disproportionate resources relative to profits, diluting returns.

Fifth, it is hard to know if GE's leverage is sustainable. Its net debts are 2.6 times its gross operating profits, again excluding its financial arm. That is high relative to its peers—for Siemens and Honeywell the ratio isabout one. Some of those profits are paper gains. And the average level of debt during the year is much higher than the figures reported at the end of each quarter.

Sixth, the strength of GE's financial arm is unclear. The new insurance loss will lower its tangible equity to 8% of assets. This is well below the comfort level, although regulators seem to have granted it forbearance in order gradually to rebuild its capital.

Seventh, it is hard to calibrate the risk this poses to GE shareholders. GE likes to hint that its industrial and financial arms are run separately. But they are umbilically connected by a mesh of cross-guarantees, factoring arrangements and other transactions.

Eighth, is GE sure that its industrial balance-sheet accurately measures its capital employed and its liabilities? Some 46% of assets are intangible, which are hard to pin down financially: for example, goodwill and "contract" assets where GE has booked profits but not been paid yet. Hefty liabilities, including pensions and tax, are also tricky to calculate. Based on GE's poor record of forecasting, it seems that large write-downs are possible. On January 24th GE said that regulators were looking into its accounting. (...)

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. »[23] Autre illustration : deux recherches wallonnes relatives respectivement aux expulsions domiciliaires et aux garanties locatives n'ont pu se faire qu'en consultant des

23 Jean-Pierre Buyle et Adrien Van Den Branden, « Pa d'intelligence artificielle en droit sans l'open data », Opinion, L'Echo, 25.11.17

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![24]

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.[25]
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 »[26] (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. »[27] 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

---

24 Voir :
- Anne Deprez et Vincent Gérard avec la collaboration de Mathieu Mosty, « Les expulsions domiciliaires en Wallonie : Premier état des lieux », Rapport final, IWEPS, janvier 2015 (https://www.iweps.be/wp-content/uploads/2017/01/exp_rapportversion28janv15.pdf)
- Robin Lebrun, « Une appréhension de la sinistralité dans le cadre de l'activation d'un régime de garantie locative », Rapport du CEHD, Charleroi, 2017, 78 pages (http://www.cehd.be/media/1109/lebrun_2017_gl_rapport_final.pdf)

25 The Economist, « Data markets – Exchange value", March 31th 2018, April 7th 2018

26 Renaud Maes, « L'impalpable sugar baby », Revue Nouvelle, 8/2017, pp.2-6

27 Adam Rutherford, op.cit.,p. 211

privées (a fortiori quand leur constitution a été soutenue par les pouvoirs publics.[28]

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.[29] 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.[30] 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.[31] 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.[32]
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.

---

28 "There was strong support (71% of 197 respondents) for making available for re-use data generated in the context of a predominantly publicly funded public task, irrespective of the private or public nature of the entity providing the service".
"88% of the 205 respondents to the question considered that access to data from private sector entities and its use by public authorities for reasons of public interest should be allowed" (= "Reverse-PSI").
Source : https://ec.europa.eu/digital-single-market/en/news/summary-report-consultation-review-directive-re-use-public-sector-information

29 Voir, par exemple : Laurent Alexandre, « Nos mandarins ringardisés par les géants du numérique », Le Monde, 13.09.17

30 Un très intéressant exemple, parmi beaucoup d'autres, d'exploitation de données : « The pumpkin index », The Economist, 25.11.17

31 Voir http://www.independent.co.uk/life-style/health-and-families/hair-dye-use-breast-cancer-risk-increase-frequent-five-times-per-year-a8002641.html

32 "More than 90% of the 159 respondents to the question considered that scientific research results (publications and research data) resulting from public funding should in principle be available under open access".
Source : https://ec.europa.eu/digital-single-market/en/news/summary-report-consultation-review-directive-re-use-public-sector-information

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 : [33]
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.[34] Autre illustration : les nombreuses applications qui sont mises en œuvre en matière de santé génèrent elles-mêmes des flux d'informations potentiellement intéressantes.[35]
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. Voici, par exemple, les tarifs demandés par l'Agence internationale de l'énergie : http://www.iea.org/statistics/onlinedataservice/ Ce n'est pas donné, en tout cas pour des personnes ou équipes de recherche aux moyens limités.

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 (voir encadré ci-après) 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. »[36] « 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. »[37]

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. »[38]

---

33 « The subscription boom will doubtless continue. So much so that antitrust regulators may eventually become nervous if too many consumers are unable to switch from their providers, either because they are contractually bound in or because the cost of doing so is prohibitively high (for example, if they lose their historical data). », The Economist, « The suscription addiction », April 7th 2018

34 « Flight response », The Economist, 17.09.17

35 Voir, par exemple : « Pill crushers », The Economist, 03.02.18

36 Mehdi Ouraoui et Pierre Singaravélou, Le Monde, 21.04.16

37 Cédric Villanni, op.cit., p.14

38 Economic Prospective Club, « Pour une transition technologique cohérente et équitable », mars 2018 (voir : http://moneystore.be/wp-content/uploads/doc/manifeste_transition_technologique_2018.pdf

Genomics

Sequencing the world[39]

How to map the DNA of all known plants and animal species on Earth

IN NOVEMBER 2015, 23 of biology's bigwigs met up at the Smithsonian Institution, in Washington, DC, to plot a grandiose scheme. It had been 12 years since the publication of the complete genetic sequence of Homo sapiens. Other organisms' genomes had been deciphered in the intervening period but the projects doing so had a piecemeal feel to them. Some were predictable one-offs, such as chickens, honey bees and rice. Some weremore ambitious, such as attempts to sample vertebrate, insect and arachnid biodiversity by looking at representatives of several thousand genera within these groups, but were advancing only slowly. What was needed, the committee concluded, was a project with the scale and sweep of the original Human Genome Project. Its goal, they decided, should be to gather DNA sequences from specimens of all complex life on Earth. They decided to call it the Earth BioGenome Project (EBP).

At around the same time as this meeting, a Peruvian entrepreneur living in São Paulo, Brazil, was formulating an audacious plan of his own. Juan Carlos Castilla Rubio wanted to shift the economy of the Amazon basin away from industries such as mining, logging and ranching, and towards one based on exploiting the region's living organisms and the biological information they embody. At least twice in the past—with the businesses of rubber-tree plantations, and of blood-pressure drugs called ACE inhibitors, which are derived from snake venom—Amazonian organisms have helped create industries worth billions of dollars. Today's explosion of biological knowledge, Mr Castilla felt, portended many more such opportunities.

For the shift he had in mind to happen, though, he reasoned that both those who live in the Amazon basin and those who govern it would have to share in the profits of this putative new economy. And one part of ensuring this happened would be to devise a way to stop a repetition of what occurred with rubber and ACE inhibitors—namely, their appropriation by foreign firms, without royalties or tax revenues accruing to the locals.

Such thinking is not unique to Mr Castilla. An international agreement called the Nagoya protocol already gives legal rights to the country of origin of exploited biological material. What is unique, or at least unusual, about Mr Castilla's approach, though, is that he also understands how regulations intended to enforce such rights can get in the way of the research needed to turn knowledge into profit. To that end he has been putting his mind to the question of how to create an open library of the Amazon's biological data (particularly DNA sequences) in a way that can also track who does what with those data, and automatically distribute part of any commercial value that results from such activities to the country of origin. He calls his idea the Amazon Bank of Codes.

Now, under the auspices of the World Economic Forum's annual meeting at Davos, a Swiss ski resort, these two ideas have come together. On January 23rd it was announced that the EBP will help collect the data to be stored in the code bank. The forum, for its part, will drum up support for the venture among the world's panjandrums—and with luck some dosh as well.

Branching out

The EBP's stated goal is to sequence, within a decade, the genomes of all 1.5m known species of eukaryotes. These are organisms that have proper nuclei in their cells—namely plants, animals, fungi and a range of single-celled organisms called protists. (It will leave it to others to sequence bacteria and archaea, the groups of organisms without proper nuclei.) The plan is to use the first three years to decipher, in detail, the DNA of a member of each eukaryotic family. Families are the taxonomic group above the genus level (foxes, for example, belong to the genus Vulpes in the family Canidae) and the eukaryotes comprise roughly 9,300 of them. The subsequent three years would be devoted to creating rougher sequences of one species from each of the 150,000 or so eukaryotic genera. The remaining species would be sequenced, in less detail still, over the final four years of the project.

(...)

Banking on it

The idea of the code bank is to build a database of biological information using a blockchain. Though blockchains are best known as the technology that underpins bitcoin and other crypto-currencies, they have other uses. In particular, they can be employed to create "smart contracts" that monitor and execute themselves. To obtain access to Mr Castilla's code bank would mean entering into such a contract, which would track how the knowledge thus tapped was subsequently used. If such use was commercial, a payment would be transferred automatically to the designated owners of the downloaded data. Mr Castilla hopes for a proof-of-principle demonstration of his platform to be ready within a few months.

In theory, smart contracts of this sort would give governments wary of biopiracy peace of mind, while also encouraging people to experiment with the data. And genomic data are, in Mr Castilla's vision, just the start. He sees the Amazon Bank of Codes eventually encompassing all manner of biological compounds—snake venoms of the sort used to create ACE inhibitors, for example—or even behavioural characteristics like the congestion-free movement of army-ant colonies, which has inspired algorithms for co-ordinating fleets of self-driving cars. His eventual goal is to venture beyond the Amazon itself, and combine his planned repository with

39 The Economist Jan 23rd 2018 | Print edition | Science and technology

similar ones in other parts of the world, creating an Earth Bank of Codes. (...)

Open data : A new goldmine[40]

Making official data public could spur lots of innovation.

AFTER a Soviet missile shot down a South Korean airliner that strayed into Russian airspace in 1983, President Ronald Reagan made America's military satellite-navigation system, GPS, available to the world. Entrepreneurs pounced. Car-navigation, precision farming and 3m American jobs now depend on GPS. Official weather data are also public and avidly used by everyone from insurers to ice-cream sellers.

But this is not enough. On May 9th Barack Obama ordered that all data created or collected by America's federal government must be made available free to the public, unless this would violate privacy, confidentiality or security. "Open and machine-readable", the president said, is "the new default for government information."

This is a big bang for big data, and will spur a frenzy of activity. Pollution numbers will affect property prices. Restaurant reviews will mention official sanitation ratings. Data from tollbooths could be used to determine prices for nearby billboards. Combining data from multiple sources will yield fresh insights. For example, correlating school data with transport information and tax returns may show that academic performance depends less on income than the amount of time parents spend with their brats.

Over the next few months federal agencies must make an inventory of their data and prioritise their release. They must also take steps not to release information that, though innocuous on its own, could be joined with other data to undermine privacy—a difficult hurdle.

Many countries have moved in the same direction. In Europe the information held by governments could be used to generate an estimated €140 billion ($180 billion) a year. Only Britain has gone as far as America in making data available, however. For example, it requires the cost of all government transactions with citizens to be made public. Not all public bodies are keen on transparency. The Royal Mail refuses to publish its database of postal addresses because it makes money licensing it to businesses. On May 15th an independent review decried such practices, arguing that public-sector data belong to the public.

Rufus Pollock of the Open Knowledge Foundation, a think-tank, says most firms will eventually use at least some public-sector information in their business. But no one has a clue what breakthroughs open data will allow, just as Reagan never guessed that future drivers would obey robot voices telling them to turn left.

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 : « Rather than being regarded as capital, data should be treated as labour—and, more specifically, regarded as the property of those who generate such information, unless they agree to provide it to firms in exchange for payment. In such a world, user data might be sold multiple times, to multiple firms, reducing the extent to which data sets serve as barriers to entry. Payments to users for their data would help spread the wealth generated by AI. Firms could also potentially generate better data by paying. Rather than guess what a person is up to as they wander around a shopping centre, for example, firms could ask individuals to share information on which shops were visited and which items were viewed, in exchange for payment. Perhaps most ambitiously, the authors muse that data labour could come to be seen as useful work, conferring the same sort of dignity as paid employment: a desirable side-effect in a possible future of mass automation. »[41]

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. A ce jour

40 The Economist, may 18th 2013

41 « The digital proletariat – Economist propose a radical solution to the problem posed by artificial intelligence », The Economist, 13.01.18

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.[42] 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[43]) 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.[44]
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* socio-culturel lié aux « prescriptions » découlant de l'exploitation

42 Isabelle Rey-Lefebvre, « Logement : une moi selon la méthode Macron », Le Monde du 13 janvier 2018

43 Voir : https://www.aviq.be/

44 Catherine Daloze, « L'homo digitalis risque l'amnésie », En Marche, 04.01.18

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. »[45] 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.[46]
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. »[47]
6. L'intégration de données micro-économiques du big data dans l'analyse macro-économique et le suivi conjoncturel est encore à développer. Comme le note le FMI, « Big data can benefit macroeconomic and financial statistics and ultimately policymaking through at least three features:
   1) By answering new questions and producing new indicators
   2) By bridging time lags in the availability of official statistics and supporting the timelier forecasting of existing indicators
   3) As an innovative data source in the production of official statistics."

   Mais, « the incorporation of big data as new data sources, either supplementing or substituting for traditional data sources, will not be exempt from methodological, organizational, and budgetary challenges."[48]

   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

45 Alexandre de Streel, « Le juge face à l'intelligence universelle », La Libre Entreprise, 24 mars 2018

46 Voir : https://socialsecurity.belgium.be/fr/news/personnes-handicapes-reaction-du-spf-securite-sociale-21-02-2017

47 Cédric Villanni, op.cit. p. 14

48 IMF Paper, op.cit.

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. »[49]

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.[50] 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.[51] 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. »[52]
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

49 Voir : https://statbel.fgov.be/sites/default/files/files/documents/Analyse/FR/Webscraping_fr.pdf

50 The Economist, «The gender pay gap – XY > XX », April 7th 2018

51 The Economist, « Ocean technology – Bleu-sea thinking », Technology Quarterly, March 10th 2018

52 L'ECHO, 8 mars 2018

web scraping[53]. Par exemple, l'exploitation du big data peut permettre de faire du *now-casting*[54], 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[55], 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[56], 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[57].

   – Dans le domaine social, où les données détenues par la Banque Carrefour de la Sécurité

53 « Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites. Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying, in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis. » (voir : https://en.wikipedia.org/wiki/Web_scraping)

54 IMF Paper, op.cit., p.16

55 IMF Paper, op.cit., p.17

56 C'est ainsi, par exemple, qu'un article de Guillaume Chapelle et Jean-Benoît Eyméoud ("Can Big Data increase our knowledge of local rental markets ?, Working Paper, Miméo Science Po, Department of economics, septembre 2017) arrive à la conclusion suivante concernant l'observation des loyers en France : « In this paper, we describe a new data collection technique in order to provide accurate data on local housing markets for researchers and statisticians thanks to online data. As we show, this can provide a relatively cheap and precise way to collect an important amount of micro data in order to answer research questions related to market dynamics. If these online data correspond to posted rents and not to signed contracts, one can think that the relative transparency of online platforms tends to force landlords to reveal the market price. The comparison between our dataset and standard surveys as the French housing survey supports this intuition. Indeed, one can observe that the distribution of rents in both sources is very close. As a conclusion, we would like to emphasize that one shouldn't neglect the opportunity offered by alternate data collection methods as webscraping. In our paper, we provide evidence that online rental data can be used in order to build local rent index following similar goods across space or to study the difference between private and social housing. We hope that this database will be used in the future to follow rental housing market dynamics using econometric methods as hedonic regressions. »

57 « Catalogue is an openly-governed community that relies on the collaborative input of its members. It is our mission to create and foster the largest global source of mobility data, connecting data producers with developers across the world. We provide the tools and the mechanisms for data to be shared, stored, organized, verified and utilized, so that together we can build the mobility solutions of tomorrow. » (voir : http://www.catalogue.global)

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[58], 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, "reverse PSI"[59], 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. »[60] 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.

Un exemple (parmi d'autres) de manquement des pouvoirs publics de fournir des données nécessaires à l'évaluation des politiques et législations

Pourquoi il n'y a plus de chiffres sur l'avortement en Belgique

Combien y a-t-il eu d'avortements l'an dernier ? Pas de réponse. Combien d'adolescentes ont-elles demandé d'interrompre leur grossesse ? On ne sait pas. Quel était le statut familial des femmes qui se sont rendues dans un planning familial pour avorter ? Étaient-elles en couple ? Seules ? Avaient-elles des enfants ? Mystère. Utilisaient-elles une méthode contraceptive ? Le point d'interrogation reste de mise…

Depuis six ans, on ne dispose plus d'aucun chiffre et d'aucune statistique concernant la pratique (déclarée) de l'interruption volontaire de grossesse (IVG) avant le délai de 12 semaines fixé par la loi de 1990 dépénalisant l'avortement sous conditions.

Plus de signe de vie depuis 2012

Cette même loi créait la Commission nationale chargée d'évaluer l'application des dispositions légales relatives à l'interruption de grossesse. Elle a notamment pour mission d'établir, tous les deux ans, à l'attention

58 « Les citoyens peu informés peuvent penser que les contrôles et les sanctions sont justes et efficaces. Mais, pour savoir si c'est le cas, il faudrait disposer d'études précises à ce sujet. » dit Jean-Claude Barbier dans un texte d'opinion (« Faut-il traquer les chômeurs ? ») publié dans Le Monde du 6 avril 2018

59 Ce que dans le jargon on appelle "reverse PSI" c'est l'accès à des données privées par les pouvoirs publics, l'inverse de ce que que prévoit la Directive sur l'utilisation des données publiques : « The directive on the re-use of public sector information provides a common legal framework for a European market for government-held data (public sector information). It is built around two key pillars of the internal market: transparency and fair competition. The directive on the re-use of public sector information (Directive 2003/98/EC, known as the 'PSI Directive') entered into force on 31 December 2003. It was revised by Directive 2013/37/EU which entered into force on 17 July 2013. In 2017, the European Commission has launched a public online consultation on the Review of the Directive 2013/37/EU, fulfilling the periodic review obligation prescribed by the Directive. The PSI directive focuses on the economic aspects of re-use of information rather than on the access of citizens to information. It encourages the Member States to make as much information available for re-use as possible. It addresses material held by public sector bodies in the Member States, at national, regional and local levels, such as ministries, state agencies, municipalities, as well as organisations funded for the most part by or under the control of public authorities (e.g. meteorological institutes). Since 2013 content held by museums, libraries and archives falls within the scope of application as well. » (Source : https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information)

60 L'ECHO, 7 avril 2018, p.16

du Parlement, un rapport sur base des documents d'enregistrement complétés par les médecins qui ont pratiqué une interruption de grossesse. Ces informations sont précieuses : elles livrent des données socio-démographiques, psycho-sociales et médicales qui permettent de mieux cerner les circonstances dans lesquelles les femmes recourent à un avortement.

Mais cette Commission d'évaluation n'a plus donné signe de vie depuis décembre 2012… Impossible donc de savoir si le nombre d'avortements augmente ou diminue. En 2011, dernière année pour laquelle des statistiques ont été publiées, 20000 IVG avaient été déclarées. Impossible encore de dire si l'âge moyen (27 ans) de ces femmes reste stable ou si le nombre de très jeunes filles qui avortent a diminué (87 avaient entre 10 et 14 ans)…

Bref, sans ces informations, il est très difficile d'étayer une analyse de la pratique de l'avortement. Et, partant, de construire une politique de prévention efficace.

Faute de candidats…

Les chiffres existent pourtant. Le secrétariat de la Commission nationale d'évaluation encode chaque jour les documents d'enregistrement d'une interruption de grossesse. Ils comprennent des dates (de la demande, de l'entretien, de l'IVG); des informations sur l'âge, l'état civil, le lieu de domicile ainsi que l'état de détresse invoqué par la patiente; ils précisent les moyens contraceptifs utilisés (ou pas) ainsi que les méthodes appliquées pour interrompre la grossesse.

Mais ces données restent confidentielles. Deux rapports statistiques, relatifs aux années 2012-2013 et 2014-2015, sont prêts. Le troisième rapport bisannuel (2016-2017) le sera en septembre 2018. Ils ne peuvent pas être publiés ni transmis au Parlement parce qu'ils doivent d'abord être validés par la Commission d'évaluation. Qui, virtuellement, n'existe plus, faute de combattants. Les appels aux candidats successifs, publiés au Moniteur, n'ont pas donné lieu à un nombre suffisant d'amateurs pour renouveler l'instance. (...) »

La Libre 27.04.18