LE BIG DATA : UNE NOUVELLE FORME D'INTELLIGENCE COLLECTIVE (2)

Données : des tablettes mésopotamiennes à nos tablettes numériques

Le terme Big Data a été utilisé pour la première fois dans un article présenté en 2000 au congrès de l'*Econometric Society*. Il sera popularisé en 2008 en faisant la couverture des très respectées revues *Nature* et *Science*. Objet d'une page Wikipédia[i] dès 2010, il se répandra ensuite dans les revues informatiques, pour finalement être adopté par le grand public[ii].

Dans le langage courant, l'expression Big Data se réfère à l'explosion des données numériques à laquelle nous assistons actuellement. Cette définition est toutefois un peu courte.

From Small to Big Data : une exploration en quatre temps

Dans ce volet *From Small to Big Data*, les billets porteront successivement sur les data (données), les algorithmes & moteurs de recherche, et big (grand) au sens de Big Data. Des notions fréquemment associées au Big Data – le data mining et l'intelligence artificielle - seront examinées dans un quatrième temps.

Déferlante de données : comment en sommes-nous arrivés là ?

Data, Données… Vient du latin Datum, dati, n. signifiant don, présent[iii].

Selon l'encyclopédie en ligne Wikipédia « Une donnée est une description élémentaire d'une réalité. C'est par exemple une observation ou une mesure »[iv].

Quand bien même ce n'est que récemment qu'elles ont envahi nos vies, l'existence des données ne date pas d'hier, loin de là.

Nous disposons de preuves laissées par nos lointains ancêtres attestant que depuis l'apparition de l'écriture en Mésopotamie peu après le milieu du quatrième millénaire av. J.-C, ils consignaient par écrit certaines informations liées majoritairement à des préoccupations

gestionnaires ou économiques. La documentation comptable constitue de loin la masse la plus importante de ces textes archaïques[v]. La mémoire humaine n'était plus suffisante. L'écriture a permis son externalisation, la conservation de faits importants sur des supports externes nécessaires à transmettre, attester diverses transactions, faits, évènements, propos. Bien plus tard, au 15e siècle, l'invention de l'imprimerie par Johannes Gutenberg fut également un vecteur essentiel dans la transmission et la consignation écrite de données[vi].

La tenue de livres de comptes, de registres d'état civil et fonciers, du commerce sous une forme plus ou moins évoluées ne datent pas d'hier, non plus. Mais les données ne se limitent pas à l'activité commerciale et au secteur administratif. Des données scientifiques étaient également consignées par écrit. Et la conservation des sons ? « Les connaissances scientifiques et techniques ont permis, dans la deuxième moitié du XIX^{e} siècle, l'utilisation de moyens mécaniques pour le réaliser. Jusqu'à ce moment, on ne savait enregistrer, par l'écriture, que l'interprétation qu'une personne faisait du son »[vii]. Et la conservation des images ? Il y a eu les fresques, la peinture, le dessin. La photographie verra le jour dans la première moitié du 19^{e} siècle[viii]. On constate donc que depuis longtemps déjà l'être humain cherche à conserver des traces de ses activités, de ce qui l'entoure, à chaque époque dans les limites des moyens techniques à sa disposition.

Au 20^{e} siècle l'informatique, internet, le web, le mail voient le jour, se développent, se démocratisent, se répandent, explosent avec l'apparition des ordinateurs personnels, portables, puis des tablettes, des smartphones au 21^{e} siècle. L'internet des objets fait une entrée remarquée sur la scène de nos vies quotidiennes. En Anglais, *Internet of Things* (IOT) se rapporte par exemple à la surveillance de l'état des structures de génie civil, à la surveillance environnementale, à la surveillance et à la gestion du trafic routier, aux maisons intelligentes, à la eSanté[ix].

A l'origine de l'écriture, les données étaient rares, très rares. Et sûrement pas numériques ! Elles se sont multipliées au fur et à mesure de l'invention de nouveaux supports et de modes de transmission.

Que s'est-il passé pour déclencher un tel mouvement de production de données, devenues numériques ? Le développement des ordinateurs et de l'informatique est un élément de ce phénomène, l'invention et l'essor d'Internet un deuxième.

Dès le début de l'écriture, il y a plus de cinq mille ans, des algorithmes étaient utilisés ; cela signifie que depuis fort longtemps déjà, les êtres humains recourent à « un procédé qui permet de résoudre un problème, sans avoir besoin d'inventer une solution chaque fois » selon la définition de l'algorithme donnée par Serge Abiteboul et Gilles Dowek dans *Le temps de algorithmes*[x]. Ou selon Wikipédia : « Un algorithme est une suite finie et non ambiguë d'opérations ou d'instructions permettant de résoudre un problème ou d'obtenir un résultat »[xi]. Pour oser l'analogie avec la cuisine, un algorithme peut-être comparé à une recette. Rien de nouveau de ce point de vue. En revanche, l'invention de machines capables d'exécuter ces algorithmes, et le développement de la science qui l'accompagne - l'informatique – ont engendré l'incroyable et radicale transformation du monde dans lequel nous vivons. Les ordinateurs n'exécutent toutefois des algorithmes que sur des données

symboliques, soit des données portant sur des chiffres. Par conséquent, certaines données telles que les images, les sons, les vidéos, qui naturellement ne se présentent pas sous forme de symboles traitables directement par un ordinateur ont dû subir une transformation – par exemple en pixels pour les images – afin de représenter une suite de symboles, souvent limitées à des 0 et des 1. Cette transformation s'appelle la *numérisation* et a donné son nom à notre temps : l'ère numérique[xii].

Et puis il y a eu Internet

Internet, terme d'origine américaine, découle du concept d'*internetting* (en français : « interconnecter des réseaux »). La première utilisation documentée remonte à octobre 1972 dans la bouche de Robert E. Kahn lors de l'édition initiale de l'*International Conference on Computer Communications* (ICCC) à Washington[xiii].

Internet, réseau informatique mondial accessible au public, réseaux de réseaux, qui n'a pas été conçu en un jour mais dont l'histoire débute à la fin des années 1950[xiv]. « Les bases d'Internet datent de mai 1974, lorsque Vinton Cerf (professeur assistant à l'Université de Stanford) et Robert Kahn (qui œuvre au Dapra, l'agence américaine pour les projets de recherche avancée de défense) publient leurs recherches sur un protocole de réseau d'échange de paquets. Le protocole TCP/IP voit le jour, base de l'architecture mondiale que deviendra Internet »[xv]. Les deux hommes ont imaginés « un moyen de relier différents réseaux baptisés « internetwork », qui est formalisé en 1974 dans l'article *A Protocol for Packet Network Intercommunication*. C'est l'acte de naissance du protocole TCP/IP »[xvi].

Par la suite des applications d'internet ont été inventées, notamment le *World Wide Web* (WWW) qui a bouleversé notre monde. Si le Web constitue une composante d'Internet, il n'est pas internet mais une de ses applications. Le *mail* en est une autre. Conçu entre 1989 et 1990 par un employé du CERN, le physicien britannique Tim Berners-Lee, la finalité initiale du Web était de répondre au besoin de partage d'informations entre scientifiques partout dans le monde. Le 30 avril 1991, Tim Berners-Lee dévoile son invention, laquelle sera officiellement versée par le CERN dans le domaine public le 30 avril 1993 « sans versement de redevances et sans aucune restriction »[xvii] autorisant ainsi l'utilisation gratuite de cette nouvelle technologie. Quelques mois plus tard, en novembre 1993, le *National Center for Supercomputing Applications* (NCSA) rend le Web accessible au grand public grâce à son navigateur Mosaic[xviii].

Internet et *la Toile* allait rapidement entrer dans les usages et dans foyers ; cela a nécessité l'invention et le déploiement d'ingéniosités techniques pour répondre de manière performante aux requêtes toujours croissantes des utilisateurs. La performance des moteurs de recherche devait être améliorée pour les rendre capables d'analyser efficacement, rapidement et de manière automatisée des millions de pages web. A ce stade, nulle préoccupation Big Data, dans son sens actuel d'analyse de données massives. En revanche, la genèse du Big Data se trouve dans cette nécessité initiale d'améliorer l'efficacité des moteurs de recherche, ce qui sera l'objet du Billet #3.

Une déferlante de données

Avant de nous pencher sur ces défis, voici une illustration chiffrée comparative de la déferlante vertigineuse de données pour *60 secondes en ligne entre 2014 et 2016* représentation par l'exemple du chemin parcouru depuis les débuts d'Internet :

Source : *What happens online in 60 seconds*, publié le 6/2/2017
par Robert Allen sur Smart Insights[xix]

Pour l'année écoulée, on peut encore ajouter ces chiffres pour *60 secondes sur Internet*[xx]:

- 69 444 heures de films regardées sur *Netflix*
- 1 389 demandes de courses traitées par *Uber*
- 527 760 photos partagées sur *Snapchat*
- 51 000 applications téléchargées *sur Apple App Store*
- 203 596 dollars de chiffres d'affaire réalisés par *Amazon*
- 120 nouveaux comptes créés sur *Linkedin*
- 38 194 *posts* partagés sur *Instagram*
- 1,04 million de boucles de vidéos de *Vine* regardées
- 38 052 heures de musique écoutées sur *Spotify*
- 972 222 *swipes* (réponses négatives et positives) réalisées sur *Tinder*

Et l'explosion des données continue exponentiellement, de plus en plus rapidement. D'autres données viennent enfler ce flot gigantesque. Les données liées à Internet et aux réseaux sociaux ne constituent pas l'ensemble des données que nous générons continuellement. S'ajoutent à ces chiffres, les données des entreprises, de l'état (données publiques), nos

données médicales ou de santé et autres données personnelles, qui viennent grossir d'autant la déferlante des données produites. Tout devient donnée, les sources sont multiples et se multiplient, données générées par des êtres humains mais aussi de manière automatisées, et significative, par des *bots*[xxi].

Bienvenue au royaume des puissances

De nos jours, l'ordre de grandeur du trafic annuel sur internet est le Zettaoctet ou *Zettabyte* en anglais, soit 10^{21} octets, c'est-à-dire10x10x10…. vingt-et-une une fois.

Mais au fait, c'est quoi un octet ou un *byte* en anglais ? Selon Wikipédia:« Un octet est un multiple de 8 bits codant une information. Dans ce système de codage, s'appuyant sur le système binaire, un octet permet de représenter 2^8 nombres, soit 256 valeurs différentes. Un octet permet de coder des valeurs numériques ou jusqu'à 256 caractères différents. Le terme est couramment utilisé comme unité de mesure en informatique (symbole : o) pour indiquer la capacité de mémorisation des mémoires (mémoire vive ou morte, capacité des clés USB ou des disques, etc.). À cette fin, on utilise couramment des multiples de l'octet, comme le kilooctet (Ko) ou le mégaoctet (Mo). Cette unité permet aussi de quantifier la rapidité de transfert d'informations en octets par seconde »[xxii].

Etant ajouté que :

- *Bit* vient de l'anglais vient de l'anglais binary digit.
- Le système binaire signifie 0 ou 1.
- Pourquoi 256 valeurs différentes ? De manière imagée, 1 bit = 1 wagon avec un 1 ou un 0 dedans, 1 octet (*byte*) est 1 train de 8 wagons (sans locomotive), et il existe 2^8, 2x2x2…. huit fois, soit 256 combinaisons possibles de trains différents.

Cette même page Wikipédia précise que : Kilooctet = 10^3 octets, mégaoctet = 10^6 octets, gigyoctet = 10^9 octets, téraoctet = 10^{12} octets, pétaoctet = 10^{15} octets, exaoctet = 10^{18} octets, zettaoctet = 10^{21} octets.

Ces chiffres excèdent notre capacité de représentation humaine.

Concrètement que représentent ces chiffres ?

Une page de texte contient quelques kilooctets (10^3) d'informations, un livre quelques mégaoctets (10^6), une bibliothèque de mille volumes renferme quelques gigaoctets (10^9) d'informations, la quantité d'informations dans l'ensemble des textes conservés à la Bibliothèque Nationale de France avoisinent les quelques téraoctets (10^{12}), l'information produite annuellement par l'Organisation européenne pour la recherche nucléaire (CERN) s'élève à quelques pétaoctets(10^{15})[xxiii]. C'est déjà plus parlant.

La croissance exponentielle du nombre de données va-t-elle continuer ? De passage récemment à Genève, Robert Kahn, un des pères d'Internet, affirmait dans une interview donnée au journal *Le Temps* : « Il n'y a pas de limite à l'expansion d'Internet »[xxiv]. A suivre.

Chère Lectrice, cher Lecteur, avant de conclure cette deuxième étape, un élément me taraude. Data, Données… Nous avons vu que ce mot vient du latin Datum, dati, n. signifiant don, présent. Donnons-nous nos données ? A qui ? Pour quel usage ? Dans l'affirmative, pourquoi les donner, s'en dépouiller, même si l'étymologie semble nous y pousser ? A moins qu'elle ne nous renvoie à une qualité intrinsèque des données qui agrégées et croisées nous permettent d'accéder à une nouvelle forme d'intelligence collective ? Tel serait le don des données ? Nous aborderons tout au long de cette série de billets différents éléments qui vous permettrons de vous forger vos propres réponses à ces questions. Nous retrouverons, finalement, ces questions dans le billet de conclusion.

A bientôt !

Anne-Sylvie Weinmann

Nota bene :

Tous les billets de cette série seront publiés sur LinkedIn mais également disponibles en format pdf avec les références détaillées sur le blog de mon site (www.medialien.ch).

An English version of this series of posts - Big Data: a new form of collective intelligence - will follow in a while.

Références :

[i] *Big Data*, https://fr.wikipedia.org/wiki/Big_data

[ii] DELORT Pierre, *Le Big Data*, Paris, Presses universitaires de France, 2015, pp. 5 et 11. (https://www.puf.com/content/Le_Big_Data).

[iii] Cette référence vient de mon dictionnaire latin-français qui venait lui-même d'un de mes parents. Et avant, je l'ignore. Il est maintenant si ancien que la page contenant l'éditeur et l'année de parution s'est détachée.

[iv] *Donnée*, https://fr.wikipedia.org/wiki/Donn%C3%A9e

[v] *Début de l'écriture en Mésopotamie*, https://fr.wikipedia.org/wiki/D%C3%A9buts_de_l%27%C3%A9criture_en_M%C3%A9sopotamie, précise notamment que « de manière assez générale, on y traite d'opérations gestionnaires concernant des produits céréaliers, des produits laitiers, des inventaires de troupeaux ou de personnel et même de calculs prévisionnels sur les rendements des champs et des cheptels. Si l'identification des différents acteurs de ces opérations reste difficile, les volumes engagés révèlent qu'elles émanent d'importantes structures officielles ».

[vi] ABITEBOUL Serge, PEUGEOT Valérie, *Terra Data*, Paris, Le Pommier, 2017, pp. 11-16 (https://www.editions-lepommier.fr/terra-data) ; *Ecriture*, https://fr.wikipedia.org/wiki/%C3%89criture ; *Imprimerie*, https://fr.wikipedia.org/wiki/Imprimerie

[vii] *Enregistrement sonore*, https://fr.wikipedia.org/wiki/Enregistrement_sonore

[viii] *Photographie*, https://fr.wikipedia.org/wiki/Photographie

[ix] PIERSON Lillian, *Data Sciences for dummies*, Hoboken (USA), John Wiley & Sons, 2e éd., 2017, p. 110 et pour quelques explications complémentaires p. 114.

[x] ABITEBOUL Serge, DOWEK Gilles*, Le temps des algorithmes,* Paris, Le Pommier, 2017, pp. 11. (https://www.editions-lepommier.fr/le-temps-des-algorithmes).

[xi] *Algorithme*, https://fr.wikipedia.org/wiki/Algorithme

[xii] ABITEBOUL Serge, DOWEK Gilles*, op. cit.*, pp. 11-12, 25, 29-33.

[xiii] *Internet*, https://fr.wikipedia.org/wiki/Internet

[xiv] *Histoire d'Internet*, https://fr.wikipedia.org/wiki/Histoire_d%27Internet; *Internet*, https://fr.wikipedia.org/wiki/Internet

[xv] SEYDTAGHIA Anouch, Interview de Robert Kahn, Le temps du 13/08/2017 (https://www.letemps.ch/economie/2017/08/13/robert-kahn-inventeur-protocole-tcpip-ny-limite-lexpansion-dinternet).

[xvi] *Vint Cerf*, https://fr.wikipedia.org/wiki/Vint_Cerf

[xvii] *Internet : le World Wide Web a 15 ans*, Futura Tech (http://www.futura-sciences.com/tech/actualites/internet-internet-world-wide-web-15-ans-9453/).

[xviii] *Internet*, https://fr.wikipedia.org/wiki/Internet; *World Wide Web*, https://fr.wikipedia.org/wiki/World_Wide_Web; *Internet : le World Wide Web a 15 ans*, Futura Tech (http://www.futura-sciences.com/tech/actualites/internet-internet-world-wide-web-15-ans-9453/).

[xix] ALLEN Robert, *What happens online in 60 seconds*, publié le 6/2/2017 sur Smart Insights (http://www.smartinsights.com/internet-marketing-statistics/happens-online-60-seconds/).

[xx] PERRICHOT Rozenn, *60 secondes sur Internet en 2016 : les chiffres clés*, Le Blog du Modérateur 25/04/2016 (http://www.blogdumoderateur.com/chiffres-internet-2016-une-minute/).

[xxi] TURRETTINI Emily, *La moitié du trafic sur Internet n'est pas humain*, Le Temps du 19/02/2017 (https://www.letemps.ch/opinions/2017/02/19/moitie-trafic-internet-nest-humain).

[xxii] *Octet*, https://fr.wikipedia.org/wiki/Octet

[xxiii] ABITEBOUL Serge, DOWEK Gilles*, op. cit.*, p. 32.

[xxiv] SEYDTAGHIA Anouch, Interview de Robert Kahn, Le temps du 14/08/2017 (https://www.letemps.ch/economie/2017/08/13/robert-kahn-inventeur-protocole-tcpip-ny-limite-lexpansion-dinternet).