Originally published in French (21/09/2017), this post is a translation (mainly) made by: www.DeepL.com/Translator

BIG DATA: A NEW FORM OF COLLECTIVE INTELLIGENCE (2)

Data : from Mesopotamian tablets to our digital tablets

The term Big Data was first used in an article presented in 2000 at the Congress of the *Econometric Society*. It will be popularized in 2008 by being on the cover page the highly respected journals *Nature* and *Science*. Object of a Wikipedia page[i] from 2010, it will then be spread in computer magazines and finally adopted by the general public[ii].

In everyday language, Big Data refers to the explosion of digital data that we are currently witnessing. However, this definition is somewhat short.

From Small to Big Data : an exploration in four steps

In this section *From Small to Big Data*, the Posts will successively deal with data, algorithms & search engines, and big in the sense of Big Data. Concepts frequently associated with Big Data - data mining and artificial intelligence - will be examined at a later fourth stage.

Data explosion: how did we get there?

Data, ... Comes from the Latin Datum, dati, n. signifying gift, present [iii].

According to the online encyclopedia Wikipedia*"A data is an elementary description of a reality. For example, it's an observation or a measurement"*[iv].

Even though it is only recently that they have invaded our lives, the existence of data is not new, far from it.

We have evidence left by our distant ancestors, that since the emergence of writing in Mesopotamia shortly after the middle of the fourth millennium B. C., some information related mainly to managerial or economic concerns were recorded in writing. Accounting documentation is by far the most important mass of these archaic texts[v]. Human memory was no longer sufficient. Writing has allowed its outsourcing, the retention of important facts on external media necessary to transmit and certify various transactions, facts, events and statements. Much later, in the 15th century, Johannes Gutenberg's invention of printing was also a key factor in the transmission and written recording of data[vi].

The keeping of books of account, civil status and land registers, trade registers in a more or less advanced form is not new either. But the data are not limited to business and administrative activities. Scientific data were also documented. What about sound conservation? *"Scientific and technical knowledge made it possible, in the second half of the 19th century, to use mechanical means to carry it out. Until then, one only knew how to record, through writing, a person's interpretation of the sound"*[vii]. What about image preservation? There were frescoes, painting, drawing. Photography appeared in the first half of the 19th century[viii]. It can therefore be seen that for a long time now, human beings have been trying to preserve traces of their activities, of what surrounds them, in each epoch within the limits of the technical means at their disposal.

In the 20th century computer science, the Internet, the web, e-mail were born, develop, democratize, spread, explode with the emergence of personal computers, laptops, then tablets, smartphones in the 21st century. The Internet of Things (IOT) makes a remarkable entrance on the stage of our daily lives. It refers, for example, to civil engineering condition monitoring, environmental monitoring, road traffic monitoring and management, intelligent houses, eHealth, etc. [ix].

At the origin of writing, data were rare, very rare. And certainly not digital! They have multiplied as new media and modes of transmission have been invented.

What has happened to trigger such a movement to produce data, which became digital? The development of computers and computer science is one aspect of this phenomenon, the invention and expansion of the Internet a second one.

From the very beginning of writing, more than five thousand years ago, algorithms were used; this means that for a long time now, human beings have been resorting to *"a process that makes it possible to solve a problem, without having to invent a solution each time"* according to the definition of the algorithm given by Serge Abiteboul and Gilles Dowek in *Le temps de algorithmes*[x].

Or according to Wikipedia: *"An algorithm is a finite and unambiguous sequence of operations or instructions enabling to solve a problem or obtain a result"*[xi]. To dare the analogy with cooking, an algorithm can be compared to a recipe. Nothing new from this point of view. On the other hand, the invention of machines capable of executing these algorithms, and the development of the science that accompanies them – computer science - have brought about the incredible and radical transformation of the world we live in. Computers, however, only perform algorithms on symbolic data, i. e., numerical data. As a result, some data such as images, sounds and videos, which naturally do not come in the form of directly computer processable symbols, had to be transformed - for example, into pixels for images - in order to represent a series of symbols, often limited to 0 and 1. This transformation is called digitization and has given our time its name: the digital age[xii].

And then there was the Internet.

Internet, a term of American origin, derives from the concept of *internetting* (in French : *" interconnecter des réseaux* “). The first documented use dates back to October 1972 in the mouth of Robert E. Kahn at the initial edition of the *International Conference on Computer Communications* (ICCC) in Washington, D. C.[xiii].

Internet, a publicly accessible global computer network, network of networks, which was not developed overnight but whose history begins in the late 1950s[xiv]. *" The bases of the Internet date back to May 1974, when Vinton Cerf (Assistant Professor at Stanford University) and Robert Kahn (who works at Dapra, the US Agency for Advanced Defence Research Projects) published their research on a packet exchange network protocol. The TCP/IP protocol is born, the basis of the global architecture that will become the Internet“*[xv]. The two men imagined *"a way to connect different networks called " internetwork ", which is formalized in 1974 in the article A Protocol for Packet Network Intercommunication. This is the birth certificate of the TCP/IP protocol "*[xvi].

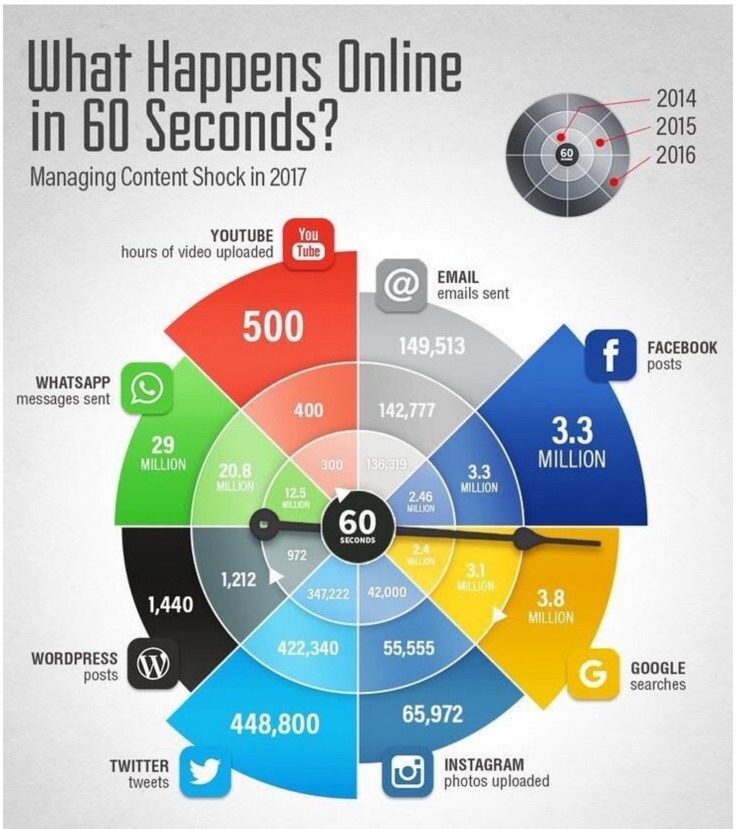
Subsequently, Internet applications were invented, namely the World Wide Web (WWW) which changed our world. If the Web is a component of the Internet, it is not the Internet but one of its applications. The mail being another one. Designed between 1989 and 1990 by a CERN staff member, the British physicist Tim Berners-Lee, the initial purpose of the Web was to address the need for information sharing among scientists around the world.

On 30 April 1991, Tim Berners-Lee unveiled his invention 991, which was officially transferred by CERN to the public domain on 30 April 1993 *"without royalty or restriction"*[xvii], thus authorizing the free use of this new technology. A few months later, in November 1993, the National Center for Supercomputing Applications (NCSA) made the Web accessible to the general public through its Mosaic browser[xviii].

The Internet and the World Wide Web were soon to enter into habits and homes, requiring the invention and deployment of technical ingenuity to respond effectively to the ever-increasing demands of users. Search engines performance needed to be improved to enable them to analyze millions of web pages efficiently, quickly and automatically. At this stage, no Big Data concerns, in its current sense of massive data analysis. However, the genesis of Big Data lies in this initial need to improve search engines efficiency, which will be the subject of Post #3.

A data flood

Before looking at these challenges, here is a comparative numerical illustration of the staggering data explosion for *60 seconds online* between 2014 and 2016, an example of how far we have come since the Internet began :



Source : *What happens online in 60 seconds* published on February 6, 2017 by Robert Allen on Smart Insights[xix]

For 2016, these figures can still be added *for 60 seconds on the Internet*[xx] :

- 69 444 hours of movies watched on *Netflix*
- 1389 requests for rides processed by *Uber*
- 527 760 photos shared on *Snapchat*
- 51 000 apps downloaded on *Apple App Store*
- 203 596 dollars of turnover generated by *Amazon*
- 120 new accounts created on *Linkedin*
- 38 194 posts shared on *Instagram*
- 1,04 million loops of *Vine* videos watched
- 38 052 hours of music listened to on *Spotify*
- 972 222 *swipes* (negative and positive responses) on *Tinder*

And the data explosion continues exponentially, faster and faster. More data comes to swell this huge flood. Data related to the Internet and social networks are not all of the data that we continuously generate. Added to these figures are company data, state data (public data), our medical or health data and other personal data, which increase the volume of data produced. Everything becomes data, the sources are multiple and multiply, data produced by human beings but also in an automated and significant fashion by bots.[xxi].

Welcome to the Kingdom of powers

Nowadays, the order of magnitude of annual traffic on the Internet is the Zettabyte (*Zettaoctet* in French), i. e. 10^{21} bytes or 10x10x10…. twenty-one times.

But by the way, what is a byte (an *octet* in French)? According to Wikipedia *"A byte is an 8 bits multiple coding an information. In this coding system, based on the binary system, one byte represents* 2^8 *numbers, i. e. 256 different values. One byte can code numerical values or up to 256 different characters. The term is commonly used as a unit of measure in computer science (symbol: o) to indicate the storage capacity of the memories (rear or dead memory, capacity of USB sticks or disks, etc.). For this purpose, byte multiples, such as kilobytes (KB) or megabytes (MB), are commonly used. This unit can also be used to quantify the speed of information transfer in bytes per second "*[xxii].

Moreover:

- Bit comes from the English **bi**nary digi**t**.
- The binary system means 0 or 1.
- Why 256 different values? Using an analogy, 1 bit = 1 car with 1 or 0 in it, 1 byte is 1 train of 8 wagons (without locomotive), and there are 2^8, 2x2x2x2.... eight times, i. e. 256 possible combinations of different trains.

This same Wikipedia page states that : Kilobyte = 10^3 bytes, megabyte = 10^6 bytes, gigabyet = 10^9 bytes, terabyte = 10^{12} bytes, petabyte = 10^{15} bytes, exabyte = 10^{18} bytes, zettabyte = 10^{21} bytes.

These figures clearly exceed our human capacity for representation.

Concretely, what do these figures represent ?

A text page contains a few kilobytes (10^3) of information, a book holds a few megabytes of information, a library of a thousand volumes equals a few gigabytes (10^9) of information, the amount of information in all the texts kept at the Bibliothèque Nationale de France is close to a few terabytes (10^{12}), the information produced annually by the European Organization for Nuclear Research (CERN) amounts to a few thousand petabytes (10^{15}) [xxiii]. This is more eloquent, isn't it ?!

Will the exponential growth in the number of data continue? Recently present in Geneva, Robert Kahn, one of the fathers of the Internet, said in an interview with the newspaper Le Temps: *"There is no limit to the expansion of the Internet"*[xxiv]. To be continued.

Dear Readers, before concluding this Post #2, I am wondering…. Data, Données... We have seen that this word comes from the Latin word Datum, dati, n. signifying donation, gift. Do we give our data? To whom? For what purpose? If so, why give them, get rid of them, even if etymology sems to spur us to do so? Unless it echoes an intrinsic quality of the data which aggregated and cross-referenced allows us to reach a new form of collective intelligence?

Would it be the meaning of data donation? Throughout this series of Posts, we will discuss various elements that should help you shape your own answer to these questions. And, eventually, we will find them again in the concluding Post.

See you soon!

Anne-Sylvie Weinmann
www.medialien.ch

References & Notes :

[i] *Big Data*, https://fr.wikipedia.org/wiki/Big_data . In this Post, references and quotes from Wikipedia, are usuallly from its version in French, translated in English for this Post.

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

[iii] This reference comes from my Latin-French dictionary, which itself came from one of my parents. And before that, I don't know. It is now so old that the page containing the editor and year of publication has come off.

[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, states that *"in a fairly general manner, it deals with managerial operations involving grain products, dairy products, herd or staff inventories and even forecast calculations of fields and livestock yields. While it remains difficult to identify the various stakeholders in these operations, the volumes involved reveal that they originate from important official structures"*.

[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 and for additional information, see 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 of Robert Kahn, Le Temps, 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*, published on 06/02/2017 on 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, 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 of Robert Kahn, Le Temps, 14/08/2017 (https://www.letemps.ch/economie/2017/08/13/robert-kahn-inventeur-protocole-tcpip-ny-limite-lexpansion-dinternet).