Exploitation of "Big Data": the experience feedback of the french military health service on sanitary data

M. Tanti
Service de veille sanitaire, Centre d'épidémiologie et de santé publique des armées
Camp militaire de Ste-Marthe, UMR912-SESSTIM
Marseille, France
e-mail address: mtanti@gmx.fr

***Abstract*— In recent years, the number of data circulating on the Internet has exploded causing the phenomenon of "Big Data". In the health field, the mass of data circulating on the Internet has also become bloated. From a study conducted within the french medical health service, this article aims to explore the construction of knowledge generated by the "Sanitary Big Data". It analyzes the creation of value made and the limitations of this " Big Data". The result is a an exploitation of polymorphic data mainly from global programs for infectious disease surveillance including PromedMail and Sentiweb. It shows the use of data from clinical trials of the Cochrane Library and the use of data of social medias and social networks. This real time data are exhautive of the population's health status. It also shows the construction of knowledge, which quickly reveals the emergence of diseases that may affect the forces in operation, as in 2009, during the H1N1 pandemic. We may also mention the anticipation of new pharmaceutical innovations, for example to detect the effectiveness of new treatments, such as in the context of the current epidemic of Ebola. As main limitations, it shows a data mining that does not offer interpretation of results and potential ethical drifts on handling health data.**

Keywords—Sanitary Big Data, French military health service, Knowledge, Value

I. INTRODUCTION

In recent years, the number of data circulating on the Internet has exploded and continues to grow exponentially. Today, there are almost as electronic data as there are stars in the universe. This phenomenon is known as "Big Data". We can give a definition of this phenomenon from three components :

- Its volume, generally we defines « Big Data » from five terabytes of data to be processed,

- Variety, the acquired data are rough or structured, text or image format, with owners and rights of use as different as their sources,

- Its velocity, it must be able to integrate in real-time the latest available data and link them to other data sets, without starting a full analysis at each cycle.

These three components therefore require new forms of information processing (Pouyllau, 2013).

In the medical field, the mass of data circulating on the Internet has also become bloated. The sanitary sector is particularly affected by this glut of information, especially with the increasing emergence of new epidemics, the frantic rush to new treatments linked with the economic interests of pharmaceutical industries and the running of publication of public laboratories.

The exploitation of this huge mass of data has opened new avenues in terms of exploration of information and production of new knowledge. This article aims to present the experience feedback of the French military health service in the use of sanitary data of the « Big Data ». The first object is to propose a definition of « Sanitary Big Data ». The second object is to explore the construction of knowledge generated. The third object is to analyze what is the creation of value provided. The last object is to determine what are the limits of this phenomenom?.

II. CONSTRUCTION OF KNOWLEDGE GENERATED BY THE EXPLOITATION OF « SANITARY BIG DATA »

A. *What is "Sanitary Big Data"?*

In the health sector, the global programs of emerging disease surveillance including PromedMail (http://www.promedmail.org) are source of bloated data, and source of very high strategic value. These programs broadcast on the democratized Internet, in near real time, data and trends data on outbreaks affecting humans and animals, on any geographic area (number of cases, deaths, geographical location), from ground relay (WHO experts, MSF, Institut Pasteur ...).

Works from scientific research and global clinical trials are also an inexhaustible source of sanitary data and high value.

For example, the Cochrane library (http://www.thecochranelibrary.com/) lists the global clinical trials and their results. PubMed references over 20 millions of international scientific works on domains as vast as epidemiology, public health, infectious diseases and health economics. The Protein Data Bank (PDB) (http://www.rcsb.org/pdb/home/home.do) broadcasts a worldwide collection of millions of data concerning biological macromolecules (DNA, proteins...) (Berman, 2000) from the genome sequencing works (human, viral, bacterial...) and the proteomics.

978-1-4799-8934-8/15/$31.00 ©2015 IEEE

In France, there is a monitoring population system called "Sentiweb" (https://websenti.u707.jussieu.fr/sentiweb/?page=presentation), which is based on private and hospital sentinel physicians who transmit, every week, by secure Internet, data from the monitoring of eight health indicators of the population (measles, viral hepatitis, influenza-like illness...), which is an enormous mass of data.

On the same principle, GrippeNet.fr (https://www.grippenet.fr/) collects anonymously via the Internet, but this time, directly by the population, data on influenza. Each week, participants report symptoms that they had since their last connection and provide a wealth of information, including the behavior of populations.

In the field of worlwide detection of influenza, we can cite Google Flu Trends (http://www.google.org/flutrends/). This site developed by Google in collaboration with the US Department of Health detects the spread of influenza. It allows, by sophisticated correlations based on the search terms in the Google search engine (such as "flu", "fever"), to predict the outbreak of influenza in any part of the territory in real time.

In the health sector, social media are a source of huge data. For example, messages and conversations exchanged on social networks (Tweeter...), medical discussion in forums (Doctissimo...) posts on personal and professional blogs are a real-time reflection of the health of the population and concerns in the field of public health (Raghupathi, 2014).

We can define the « Sanitary Big Data » as the set of large medical data, extremely varied, produced in near real time, by the population and the scientific communities. These data are raw or structured. They are in any format and they provide information concerning population health status, behaviors and health concerns.

B. Construction of Knowledge

B.1. By what method ?

The mass of information from« Sanitary Big Data » is collected and analyzed by the medical intelligence service of the Military Centre of Epidemiology and Public Health, part of the French military health service, dedicated for the exploitation of the data from the « Sanitary Big Data » (Boutin, 2004).

To do this, it uses data mining tools, methods of collection and analysis that are automatic or semi-automatic. It uses monitoring software and specialized documentary platforms. It also uses cloud computing tools, and software of clustering and data classification.

B.1. Which knowledge is build?

A number of examples of construction of knowledge from the « Sanitary Big Data » can be presented.

For example, in 2009, in full H1N1 pandemic, the use of Google Flu trends by the Medical Intelligence Service allowed to monitor the dynamics of the epidemic in real time and allowed to anticipate the evolution of the epidemic. This tool revealed soon the emergence and spread of the disease as the traditional methods of field collections.

Indeed, statistics from field physicians take several days to be analyzed (eg Sentiweb in France, and the Center for Disease Control and Prevention in the USA- CDC). This tool is very fast, see almost instantaneous. However, a study from 2014 tempers the advantages and demonstrates that the figures developed by the tool are overestimated and exceed 50% of those from the field (CDC). It would requires a recalibration of data from field data (Lazer, 2014).

In addition, the Medical intelligence service, object of our study, uses data from genomics and proteomics, from the Protein Data Bank (PDB). Data exploitation of this « Big Data » opens the way for the construction of new medical knowledge and explore new pharmaceutical innovations, including vaccine and therapy. It can be used to detect the effectiveness of influenza vaccine.

The exploitation of the data from Sentiweb allows to track the detection of influenza epidemics or acute diarrhea in France. It allows modeling with a view to decision support. It allows estimation of the basic parameters of the transmission, it assesses the impact of control and intervention strategies It allows the integration of medical and economic aspects.

In addition, new knowledge are constructed by crossing different data of the « Sanitary Big Data », therapeutic, vaccine, diagnostic, sanitary, medical, economic and strategic data.

III. CREATION OF VALUE

A. Definition

Value creation is a theme that actually raises increasing interest in different areas of management science: strategic management, corporate finance, accounting, management control, organization, marketing. Bourguignon A. (Bourguignon, 1998) distinguishes three meanings of value: the value in the sense of measure (especially in science such as mathematics and physics), the value in the economic sense and the value in the philosophical sense.

The term of value is synonymous with the term of wealth. The theme of the value is the subject of multiple looks or paradigms, eg common visions to members of a particular group (Kuhn, 1983). The issue of value therefore refers to the question of the recipients of the value created: for whom we create value? As part of the corporate finance, value is often a financial value for the shareholder.

In our study, we define value as the strategic value for military decision-maker in health, so this is a larger value than the economic or financial value. Since it also includes medical, regulatory and adjudicative aspects. Creating economic value is to vary in the direction of increasing. Conversely, destroy value, it is lower over time. The creation of economic value is in the heart of organizational activities and at the center of their vocation and strategy (Savall, 1998).

The creation of the value is also at the center of the concerns of military organizations. In our study, it is defined as the increase of the strategic value for the military decision-maker in the domain of health. The creation of the value will provide benefits to anticipate the military decisions, to prevent

risks to the soldier and the people under its responsibility. Particularly in our framework, health risks, risk of epidemics which by definition have an impact on military operations, peacekeeping and homeland security.

B. What value created?Equations

New knowledge constructed by crossing different data of the « Sanitary Big Data » are made available to decision-makers of the French military health service, in the form of a dedicated intranet platform, which allows the creation of value, that we describe in this section.

This platform leads to consider unpublished data correlations. It offers two challenges. The first challenge will be to reuse, share and verify the knowledge resulting from the exploitation of the « Big Data ». Another challenge will be to reuse data to extract innovations and trends. Main value of the creations are:

-The Early detection of epidemic events on current, possible or probable military operations theatres,

-The Monitoring of known epidemic events (epidemics, pandemics), to anticipate their occurrence and their potential impact on the soldier,

-The Monitoring of the social impact of a health event, of an epidemic risk,

-The monitoring of an action or health policy, such as a vaccine or therapeutic policy,

-The Identification of preventive or therapeutic innovations, to anticipate and prevent health risks and their economic and human impacts.

The exploitation of « Big Data » finds particular value creation, in the context of emerging epidemics, like in the current Ebola outbreak raging in Africa, at the time of writing this article. It allows the monitoring of the event, in real time, for example through the use of the data from PromedMail.

The exploitation of the « Sanitary Big Data » also allows the identification of therapeutic and vaccine innovations, notably through the analysis of data from the Protein Data Bank (Reynard, 2014). The identification of these innovations will create strategic value for military decision makers. They will be able to anticipate the risks in operation and preserve the health of the population from which they are responsible.

IV. LIMITS OF BIG DATA

The limits of the « Sanitary Big Data » are, first, related to the data mining tools, that do not offer interpretation of the results, because an expert analyst of data mining and a person familiar with the trade which the data are extracted (one epidemiologist in this context) are needed to analyze the software deliverables.

In addition, data quality, relevance and completeness of data, is a necessity for data mining, but that is not enough. The input errors, duplicates, not filled data or data indicated without reference to time, also affect the quality of the results.

A limit also comes from the interoperability of the different systems and their ability to work together. Which is currently not the case in the field of « Sanitary Big Data ».

Another limitation, under ethics, is possible diversion of medical or personal data for any purpose other than that assigned initially, or for any other purpose. For example, possible threats to the privacy of individuals are possible, especially in search of personal data collected on the Internet or on social networks where people reveal themselves and voluntarily their health. In this context, should be questioned the feasibility of an exploration of the « Big Data » that would preserve the privacy of individuals.

The storage of the data necessary for their exploitation also poses another technical problem. Digital data can be hacked. In this case, the encryption is an existing technical solution. Finally, one of the limitations of the « Sanitary Big Data » is on logistics information : how to ensure that the relevant information reaches the right place at the right time to the right person ? Especially for decision making in real time or near real time? This is a micro-economic approach that is being evaluated by the medical intelligence service. However, its effectiveness also depends on the combination between micro and macro approaches of the problems.

V. CONCLUSION

The « Big Data » brings advances in the field of health. Particularly in the sanitary field, exploitation of the « Big Data » allows to the french medical forces, value creation, particularly for the decision-making. The « Big Data » is also an innovation in terms of economic and social models. But is it just an evolution of the performance of existing tools or a simple fad? This question remains open. Finally, the "Big Data" involves the use of ultra-sensitive data, that even if they are used anonymously, must be handled with care. The debate is more relevant than ever, and the time will tell how to reconcile these two aspects.

REFERENCES

[1] Berman HM & al (2000). The Protein Data Bank. Nucleic Acids Research; 28 (1): 235-242.
[2] Boutin JP & al (2004). Pour une veille sanitaire de défense. Medecine et Armées ; 32 (4):366-372.
[3] Bourguignon A (1998). Management Accounting and Value Creation : Value Yes but what Value ? Working Paper, ESSEC, November 1998, 19 p.
[4] Kuhn T (1983). La structure des révolutions scientifiques, Flammarion
[5] Lazer D & al (2014). The Parable of Google Flu: Traps in Big Data Analysis. Science ; 343 (6176): 1203–1205.
[6] Pouyllau S (2013). Web de données, big data, open data, quels rôles pour les documentalistes. Documentaliste - Sciences de l'Information ; 50 : 32-33
[7] Raghupathi W & Raghupathi V (2014). Big data analytics in healthcare: promise and potential. Health Information Science and Systems; 2: 3

[8] Reynard O, Volchkov V & Peyrefitte C (2014). Une première épidémie de fièvre à virus Ebola en Afrique de l'Ouest. Medecine Sciences ; 30 : 671–673
[9] Savall H, Zardet V, Cappelletti L, Beck E, Noguera, Ocler R (1999-200I). Rapport de recherche, bilan de réalisation d'une recherche-intervention conduite sur 72 entreprises du secteur de la gestion de patrimoine, ISEOR.