SYSTEMS & MANAGEMENT
Electronic magazine

BIG DATA, DATA SCIENCE AND THEIR CONTRIBUTIONS TO THE DEVELOPMENT OF THE USE OF OPEN SOURCE INTELLIGENCE

Danielle Sandler dos Passos[1]

1 NOVA Information Management School, University Institute of Lisbon (ISCTE), New University of Lisbon

ABSTRACT

According to the present technological development and the wide availability of data through the open means of communication, the necessity of new mechanisms that help to correctly absorb and analyze such information is seem needed, which will take advantage of all *Open Source Intelligence (OSINT)* can offer. The goal of the article is to find the advantages of the use of *OSINT* by intelligence agencies, organizations, and enterprises, and how Big Data and Data Science mechanisms can help to spread information, and make it more effective.

Keywords: Intelligence*;* Open Source Intelligence (OSINT); Big Data; Data Science.

1. INTRODUCTION

With the arrival and the broad diffusion of the Internet, the communication vehicles took new shape and dimension. Today, there is a large set of information available, thus leading to what is called Open Source Intelligence (OSINT), which is referred to intelligence, in the sense of information acquired through data available for the general public through means of communication, such as newspapers, sites, blogs, scientific journals, TV, and others.

However, it does not matter to have access to a large load of information if no one knows how to make a good use of it. Because of this necessity to value and treat information, today there are two important tools: *Big Data* and *Data Science*.

Big Data can be simply defined as a large, complex set of information, which traditional methods of processing would be insufficient for its treatment — including processes of analysis, capture, research, sharing, storage, visualization, and safety of information.

Data Science is described as the science responsible for the analysis and use of data, which incorporates techniques and theories from many different areas, such as logic, mathematics, statistics, computing, engineering, and economics.

Thus, based on the demonstrated elements, this article aims to show the benefits of the incorporation of Open Source Intelligence on the daily life, and how Big Data and Data Science can help in this process, making OSINT can be more useful and effective in decision making processes.

2. WHAT IS INTELLIGENCE, AND WHAT IS THE DIFFERENCE BETWEEN DATA, INFORMATION, AND KNOWLEDGE

Among the many definitions of intelligence, the first idea (risen from the corporate and espionage fields) describe it as a product from the collection, analysis, evaluation, and interpretation of all available information, which can affect or not the survival and the success of the organization (Eells *et* Nehemkis, 1984). In a broader definition, in which intelligence is similar to knowledge and information, it is described as all collected, organized, and/or analyzed information aimed to supply the demand of a decision-making agent (Cepik, 2002). And, in a narrower view, it is the collection of information without consent, cooperation, or even acknowledgement from the part of the ones being investigated (Cepik, 2002).

However, it is important to see that, differently from the definition chosen, all activities linked to intelligence aim to produce knowledge based on previously selected, evalua-

DOI: 10.20985/1980-5160.2016.v11n4.1026

Electronic Journal of Management & System
Volume 11, Number 4, 2016, pp. 392-396
DOI: 10.20985/1980-5160.2016.v11n4.1026

ted, interpreted, and in the end, exposed data in a useful format for the decision-making process.

Differently from many people think, intelligence is no the synonym for knowledge or information. Information is a contextualized data; knowledge is a result from the analysis of information based on the learning and in the experience of the individual; and intelligence is the practical use of the knowledge, when triggered by a decision-making process. Therefore, all intelligence is information, but not all information is intelligence (Lowenthal, 2012).

Besides that, the process of creating of intelligence can be categorized according to the source of data collection — OSINT (open source intelligence), HUMINT (human source intelligence), SIGINT (signal intelligence), and IMINT (imagery intelligence) — and, independently from the chosen source, after collection, it is necessary to verify, analyze, and treat data to make them useable in the process of decision making[1].

In synthesis, the process is always the same: collection of information according to necessity and analysis, and the report to the decision-making agent. However, it was possible to observe that open sources became more important for the process. With the arrival of the Internet and other technological developments, the world started to access and share millions of bytes of information in real time — and then, it was possible to see how wrong is the idea that only sensitive information was valuable. Thus, Open Source Intelligence becomes considerably important.

3. OPEN SOURCE INTELLIGENCE (OSINT)

This is a vast and expanding concept among the intelligence agencies, corporations, and governmental agencies in general. It relates to the idea of the use of open source structures to gather information. OSINT is defined as the analysis based on "legal acquirement of official documents without security restrictions, of direct and non-clandestine observation of political, military, and economical aspects of the internal life of countries or targets, media monitoring, legal acquisition of specialized technical-scientific books and magazines, or in other words, a more-or-less large range of available sources which access is allowed without special security restriction measures" (Cepik, 2003).

It is understood here as open source the means of communication, such as media (newspapers, magazines, radio, TV), public information (governmental reports, public burgets), and professional and academic productions (articles, papers, symposiums, conferences). And also: grey literature (scientific and technological researches and other limited-distribution material), third-part observation and Web content (anyone can become a source of information), and geospatial information (satellite imagery, field mapping) (Brito, 2006).

In OSINT, the processes involved aim to collect open source information and to threat it. In the end, the result will be the product of a rationale based and contextualized to a fact, or movement.

As a pioneer in the usage of OSINT, it is possible to cite the Foreign Broadcast Information Service (FBIS), a North--American organization located in the University of Princeton which, during the Second World War, would gather information from international news headlines as source of intelligence, and during the Cold War, monitored official publications from the Union of Soviet Socialist Republics (USSR). After the end of the Cold War, FBIS lost some of its importance, as theoretically there was not a real threat or an enemy to the USA. However, with the attack in September 11th (2001)[2], the use of open source information was again seen as important. Eventually, after the event it was possible to observe many sources of information could have helped to prevent the attack (and, who knows, it could even have avoided the happenings) were all available to the general public.

Since 2001, NATO[3] defends the wide use of Open Source Intelligence, and within this spectrum, it developed the terminologies Open Source Data (OSD) and Open Source Information (OSI). Both are related to the information before its analysis, as soon it is captured. OSD is used to

[1] The process known as Intelligence Cycle, described by Johnson, R. in Analytic Culture in the US Intelligence Community – an Ethnographic Study, 2005. It consists of the following stages: 1) Planning and direction: management of all effort in the process and setup of requisites for data selection, based on the presented demand; 2) Collection: capture of brute data (not yet analyzed and treated), according to demand; 3) Processing: analysis and treatment of brute data, as they can be used for decision making processing; 4) Analysis and production: testing reliability, validity, and relevance of the information collected; 5) Dissemination: sharing the produced knowledge with the target audience.

[2] Series of terrorist attacks against the United States, coordinated by the Islamic terrorist group Al-Qaeda, which resulted in the collision of two airplanes against the Twin Towers (buildings of the business complex of the World Trade Center, in New York), leading to the death of hundreds of people.

[3] North Atlantic Treaty Organization, also known as OTAN (Organisation du Traité de l'Atlantique Nord), created in 1949, with the objective to guarantee the collective defense of its country members (today, 28) in response to any attack against any member.

design elements, such as photos and commercial satellite imagery, and OSI refers to information that comes from social communication means, reports, books, and similar publications. For the organization, OSINT is "the information that was deliberately found, discriminated, distilled, and disseminated into a selected audience, in order to answer a specific question".

If intelligence comes from found, discriminated, distilled, and disseminated information for the decision-making agent (Steele, 2006) (indifferently if they are from an open or a sensitive source), and if they are accepted into the definition of intelligence as knowledge or analyzed information such as from a secret of a sensitive information, why then secret services and their spies are losing grounds for OSINT?

As an answer, there are many reasons, in which among those it is possible to confirm that there are millions of bytes of relevant information in open sources – such as the fact that, from 1998 to 2008, there were 15 terrorist links in websites to more than 4,500 ones – and the considerable decrease of costs to obtain information, as they are available to any person.

However, the information collected are useless if they are not properly filtered, analyzed, and validated. Thus, it is considerably important the processes used and the analysis involved in processing the data. OSINT will only be beneficial to the process it there is a correct implementation and investment in systems, structure, and technology, with qualified agents, previously trained to find adequate sources of information, who will define the relevance of data to supply the required demand, and to analyze it. After all, today the biggest issue is not the lack of data, but the correct analysis of it.

Based on that, there are new softwares, technologies, concepts, and cultures involved in the process of intelligence. Among those, the ones that are growing are Big Data and Data Science.

4. BIG DATA

The terminology Big Data arises in the beginning of the 1990s, in NASA[4], with the objective to describe the concept of the set of large and complex data, where computing structures and systems used so far were not enough to properly capture, process, analyze, and store the information. Thus, it can be described as the use of effective systems and technologies to valorize large sets of data, making them more precise, and assisting in the mitigation of the risks involved in the process of decision making.

Big Data is supported by many technologies and algorithms that are implemented in large data banks (structured or not), with the objective to capture, analyze, process, and disseminate correctly the information, according to the demands and the objective set in the beginning of the process, frequently reanalyzed. Its main goal is to provide useful information to the decision-making process.

A study performed by the *OBS* (Online Business School) showed that, from 2004 to 2014, more data was created than any other previous period of history. This finding confirms the words of Peter Norvig, director of researches from Google: *"We do not have better algorithms. We only have more data "*.

Under this circumstance, it is possible to see the primordial role of information. However, despite what Norvig mentions, the reading of the passage does not tell the complete story. The receiver of the information must understand it favorable to him. And it is under this environment that Big Data is so important. Based on volume, variety, velocity, veracity, and value[5] , Big Data is capable to store a large number of varied information, quickly analyze it and observe its veracity, which enables value to be aggregated to the process of decision making, ultimately making it more effective and efficient.

Associated to the OSINT, Big Data is being able to map standards of behavior and tendencies. The project *Google Flu Trends* is a good example of it. Through the project, it was possible to identify an epidemic scenario of flu using as source of information the data users plotted in the search system of Google. Mapping the geographical areas where people searched on the web words related to flu, it was considered that a flu epidemic was really happening in the area. Big Data also has helped to identify the behavior standards of terrorists in social media, which has a considerable value to prevent their development and attacks.

Besides that, it was possible to observe that, with Big Data, there were significant changes in the way data analysis is performed and thought. The first change is, when dealing with large volume of data, the perception of how to look at data changes. In other words, changing the scale, our perception also changes. The second modification is that, because there is a large quantity of information, the *N* of the sample is gigantic, thus the maximum precision is not the focus, and the observation is over the tendencies. This happens because, when dealing with small numbers the goal is

[4] National Aeronautics and Space Administration, North-American agency responsible for research and development of space programs.

[5] The concept of the 5 Vs was created in 2001, by the information analyst Doug Laney to describe the Big Data.

to find the precision to the extent, once there is the confirmation that the result is found in a single number. However, with a very large set of data, the tendency directs to the result. And, in the end, the third change is the release from the effect of causality to dedicate attention to the correlations. As seen, in a large set of data the causalities are considered as in small amount, and if they are not, they are read as correlations, indicating a tendency.

Then, it can be seen that the associations and analysis presente throughout of the processes involving Big Data would be impossible to be performed if the present technology and systems were present, being the statistical methods ineffective for such interpretations. On the other hand, Big Data will only be fully successful without the analytical deficits caused by the lack of information or by the low quality of the data, if the parameters and objectives were well established, and if there are prepared and specialized analysts in their areas of expertise. And here comes Data Science.

5. DATA SCIENCE

It can be defined as a set of techniques used in the process and analysis of data, with the goal to provide information for intelligent decisions. For such, many areas of knowledge must be merged, from simple statistical concepts to complex algorithms.

Its analysts are known as data scientists and it is desirable that they have a background in the area of information technology (IT) to capture effectively and in a timely manner the data requested; mathematical and statistical knowledge to define the models and algorithms to be used, understanding their implications and outcomes; and in the end, business knowledge, to be able to translate the results in information that result in a support for the decision-making agent.

The Data Science process is similar to the one used in Big Data – it starts with the collection of data through the correct dimensioning of the problem/objective. It is followed by the analysis of data, with the visualization and application of techniques and algorithms, and it ends with the communication of the results.

Yet, throughout the process, there will be the necessity of new data – some of which will be discharged and errors in the analysis will appear. That is why the analysis require a vast know- how in many areas of the sciences as they need to make the correct answers, capture the correct data, and they need to have the correct perception in how to proceed throughout the process, so in the end, data becomes intelligence.

6. CONCLUSION

With the great availability of data generated by the "democratization of the information" and by the technological development and its popularization, the Open Source Intelligence came as one the main sources used to acquire data.

Among the advantages of the use of this source, it is important to highlight the high level of opportunities, with a large amount of information, and the low cost required to access it. In a period of economic crisis and budget adjustments, its use is more attractive, permitting the amplification of the possibilities of intelligence services. On the other hand, the excessive amount of data, the dubious quality of the information, and the lack of confidence in the sources can disable the benefits of the use of open sources. Therefore, the importance to combine Big Data and Data Science practices to the use of OSINT.

The use of Big Data and Data Science in intelligence processes brings more value, as they enable effective profits in costs, innovation, and productivity. This happens because, to perform such processes, analysts with vast experience and knowledge in many areas are used, besides having many technologies at hand, systems and structures that permit the capture and the manipulation of the information needed to the demand, transforming the data into intelligence – useful information to the process of decision making.

In the end, it is evident that the intelligence services, enterprises and organizations benefit considerably with the use of Bid Data and Data Science when manipulating the information provided from open sources. They make Open Source Intelligence a wide, safe, cheap, and effective source of information, which contributes to the result of their activities and provide the institutions with a competitive advantage during the decision-making processes.

REFERENCES

Afonso, L. (2006), "Fontes abertas e Inteligência de Estado", Revista Brasileira de Inteligência, No. 2, disponível em: www.abin.gov.br/modules/mastop_publish/?tac= Fontes_abertas_e_Inteligencia_de_Estado (Acesso em 01 de junho de 2015).

Best, C. (2008), "Open Source Intelligence". Joint Research Centre, disponível em: media.eurekalert.org/aaasnewsroom/2008/FIL_000000000010/071119_MMDSS-chapter_CB.pdf (Acesso em 05 de junho de 2015).

Brito, V. (2006), O Papel Informacional dos Serviços Secretos, Dissertação de Mestrado em Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, MG.

Cepik, M. (2003), Espionagem e democracia, 1 ed., FGV, Rio de Janeiro, RJ.

Cepik, M. Inteligência e Políticas Públicas: dinâmicas operacionais e condições de legitimação, Security and Defense Studies Review, Nº 2, vol. 2. Rio de Janeiro, 2002.

Ghiggi, L. et SEBBEN, S. (2009), "Inteligência", Dossiê Temático Nº06, disponível em www.ufrgs.br/nerint/folder/artigos/artigo76.pdf (Acesso em 02 de junho de 2015).

Eells, R. et NEHEMKIS, P. (1984). Corporate intelligence and espionage: A blueprint for executive decision making, 1 ed., Macmillan, New York, NY.

Gonçalves, J. (2013), Atividade de Inteligência e Legislação Correlata, 3 ed., Impetus, Niterói, RJ.

Johnston, R. (2005), "Analytic Culture in the US Intelligence Community – an Ethnographic Study", disponível em: www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/analytic-culture-in-the-u-s-intelligence-community/ analytic_culture_report.pdf (Acesso em 25 de maio de 2015).

Leite, S. (2014), "O Emprego das Fontes abertas no Âmbito da Atividade de Inteligência Policial", Revista Brasileira de Ciências Sociais, Vol. 1, No. 5, Brasília, DF.

Lowenthal, M. (2012), Intelligence: From Secrets to Policy, 5 ed., CQPress, Washington, DC.

Mendes, G., Moresi, E., Silva, W. (2010), "Estudo sobre Portais Públicos como Fontes Confiáveis para Inteligência de Fontes Abertas", artigo apresentado no COVIBRA 2010: Congresso Virtual Brasileiro – Administração, 19 - 21 de novembro, 2010, disponível em: www.convibra.org/2010.asp?ev=71&p=&lang=en (Acesso em 20 de maio de 2015).

Mendes, G. et MORESI, E. (2012), "Operações de Informação: um estudo sobre o desenvolvimento de doutrina aplicada à prevenção à fraude", Sistemas, Cibernética E Informática, Vol. 9, No. 1, Brasília, DF.

North Atlantic Treaty Organization. (2001), "Open Source Handbook", Vol. 1, disponível em: www.oss.net/dynamaster/file_archive/030201/ca5fb66734f540fbb4f8f6 ef759b258c/NATO%20OSINT%20Handbook%20v1.2%20-%20Jan%202002.pdf (Acesso em 20 de maio de 2015).

Steele, R. (2006), "Open Source Intelligence", Forbes, disponível em: http://www.forbes.com/2006/04/15/open-source-intelligence_cx_rs_06slate_0418steele. html (Acesso em 19 de maio de 2015).