Big data and official statistics in Morocco : opportunities and challenges

Lina TAZI

Head of Foreign Trade Unit

INAC- High Commission for Planning, Morocco

linatazi.inac@yahoo.fr **/** l.tazi@hcp.ma

IAOS 2016 CONFERENCE, Abu Dhabi, 6-8 December 2016

Abstract : To continue improving its statistical value proposition, many National Statistical Offices (NSOs) or departments strive to reduce the cost of statistical production, improve the timeliness and frequency of its offerings, and create new or richer statistics that meet emerging data needs. This paper outlines the opportunities and challenges that Big Data presents to NSOs, especially in developing countries such as Morocco. An overview of the current Moroccan situation is provided through a SWOT analysis that highlights the strengths and weaknesses of the Moroccan statistical system, with the purpose of integrating Big Data as additional data sources for Official Statistics.

Keywords : Big Data, official statistics

NB : views expressed in this paper are those of the author and do not necessarily reflect the views of the High Commission for Planning of Morocco

1. Introduction

Every day, more and more data are generated on the web, in an extremely rapid way. A huge volume of data is produced by sensors in the ever growing number of electronic devices surrounding us. According to statistics from International Telecommunication Union (ITU, 2015), around 44% of the world population has an internet connection today up from less than 1% in 1995. The amount of data and the frequency at which they are produced have given birth to the concept of 'Big data'. The term emerged in the 1990s, referring to the growth in volume of data [5].

Big data is characterized as data sets of increasing volume, velocity and variety. Big data is often largely unstructured, meaning that it has no pre-defined data model and/or does not fit well into conventional relational databases. In addition to generating new commercial opportunities in the private sector, Big data can also be used as an input for official statistics ; either for use on its own, or in combination with more traditional data sources such as sample surveys and administrative registers [13].

The advent of Big Data is introducing important innovations. The availability of new data sources, with dimensions greater than previously experienced but, with questionable consistency, poses new challenges to National Statistical Offices (NSOs). It imposes a general rethinking that involves tools, software, methodologies and organizations. Challenges coming from Big Data are not only due to their particular characteristics, but also to the fact that their origin and generation mode are often completely out of NSOs control.

The rest of this paper is organized as follows : section two lays out some Big Data concepts and characteristics before briefly outlining the main sources of Big data. Section three discusses the main opportunities that Big data can offer and some possible uses in official statistics. The next section includes an overview of major challenges faced by NSOs when deciding to adopt Big Data sources in the regular production of official statistics. A synopsis of the current Moroccan situation concerning Big Data and the national statistical system is provided in the fifth section. The concluding section gives some recommendations for NSOs to engage with Big Data.

2. Big data and official statistics : Definitions and sources

2.1 Definitions

There are several definitions of Big data, " which differ on whether you are a computer scientist, a financial analyst, or an entrepreneur pitching an idea to a venture capitalist…"[10].

United Nations defines it as "... an umbrella term referring to the large amounts of digital data continually generated by the global population." [14]

According to Wikipedia, Big data is " … a blanket term for any collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications."

Big Data is a term "that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.” [11]

NIST (National Institute of Standards and Technology) defines Big data as data of which the data volume, acquisition speed, or data representation limits the capacity of using traditional relational methods to conduct effective analysis or the data which may be effectively processed with important horizontal zoom technologies [9].

The Five V’s of Big Data

Big Data is typically defined in terms of 3Vs, a designation originally developed by Gartner [7] : greater **Volume**, more **Variety** and a higher rate of **Velocity**. Volume refers to the exponential increase in the amount of data that are generated and stored. It is estimated that data production will be 44 times greater in 2020 than it was in 2009.Though the word big implies such, big data isn’t simply defined by volume, it’s about complexity. Many small datasets that are considered Big data do not consume much physical space but are particularly complex in nature. At the same time, large datasets that require significant physical space may not be complex enough to be considered Big Data.

In addition to volume, the Big data label also includes data variety and velocity. **Variety** references the different types of structured and unstructured data that organizations can collect, such as transaction-level data, video, and audio, or text and log files. **Velocity** is an indication of how quickly the data can be made available for analysis.

Big Data has also been defined to consist of 5Vs that adds Veracity and Value to the already existing 3Vs [16]. **Veracity** is an indication of data integrity and the ability for an organization to trust the data and be able to confidently use it to make crucial decisions. It refers to the confidence level associated with certain types of data accounts for the correctness of the data, and can include data quality problems such as noise or missing values. **Value** accounts for Big Data in the sense that if particular data does not provide significance or value, it is not relevant for Big Data analysis or for development.

Table 1 illustrates characteristics of Big Data that more completely describe the difference between “Big Data” and the historical perspective of “normal” data.

Table 1 : Characteristics of Big Data

Characteristic	Description	Attribute	Driver
Volume	The sheer amount of data generated or data intensity that must be ingested, analyzed, and managed to make decisions based on complete data analysis	According to IDC's Digital Universe Study, the world's "digital universe" is in the process of generating 1.8 Zettabytes of information - with continuing exponential growth – projecting to 35 Zettabytes in 2020	Increase in data sources, higher resolution sensors
Velocity	How fast data is being produced and changed and the speed with which data must be received, understood, and processed	• Accessibility : Information when, where, and how the user wants it, at the point of impact • Applicable : Relevant, valuable information for an organization at a torrential pace becomes a real-time phenomenon • Time value : real-time analysis yields improved data-driven decisions	• Increase in data sources • Improved connectivity • Enhanced computing power of data generating devices
Variety	The rise of information coming from new sources both inside and outside the walls of the organization creates integration, management, governance, and architectural pressures on IT	• Structured – 15% of data today is structured, row, columns • Unstructured – 85% is unstructured or human generated information • Semistructured – The combination of structured and unstructured data is becoming paramount • Complexity – where data sources are moving and residing	• Mobile • Social Media • Videos • Chat • Genomics • Sensors
Veracity	The quality and provenance of received data	The quality of Big Data may be good, bad, or undefined due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations	Data-based decisions require traceability and justification

Source : Adapted from [12]

Regarding the definition of "**Official Statistics",** we can use the definition provided by the United Nations, the Fundamental Principles of Official Statistics, Principle 1 : "Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation." [15]

Official statistics are statistics that describe a situation, provide a *snapshot* of a country, its economy, its population etc. When Big data is used as an additional source of information, this *snapshot* needs to be considered.[13]

2.2 Sources

The Big data phenomenon makes one realize that our world is now full of data. This cannot be ignored, hence the interest for official statistics. So far, Big data sources can be categorized into six groups, according to their type and origin :

- Commercial or transactional : arising from the transaction between two entities, e.g. credit card transactions, on-line transactions (including from mobile devices), etc.
- Administrative : related to the administration of a program, be it governmental or not, e.g. hospital visits, electronic medical records, insurance or bank records, etc.
- From sensors, e.g. satellite imaging, road sensors, climate sensors, etc.
- From tracking devices, e.g. tracking data from mobile phones, GPS, etc.
- Behavioural, e.g. online searches about a product/service or any other type of information, online page view, etc.
- Opinion, like comments on social media, etc.

It should be noted that administrative data is one of the main data sources used by NSO's for statistical purposes [11]. Administrative data is collected at regular periods of time by statistical offices and is used to produce official statistics. Traditionally, it has been received, often from public administrations, processed, stored, managed and used by the NSOs in a very structured manner. Can one consider administrative data "Big" in accordance with the definition given above? For the moment the response would be, probably not. Administrative data can become "Big" when the velocity increases, e.g. using extensively administrative data where data is collected every day or every week instead of the usual once a year or once a month.

3. Opportunities and uses

Big data offers many opportunities for NSOs such as improving the timeliness and frequency of its offerings, creating new or richer statistics that meet emerging statistical data needs, reducing costs of statistical production, etc. Big Data has also the potential to complement and improve official statistic activities by replacing particular indicators and measurement processes. Big Data can feed into the statistical process through its descriptive function (referred to as diagnostic, via maps, descriptive statistics, visualizations...), in addition to its predictive and forecasting function (making inferences about current conditions and forecasts about future events, where the likelihood of some events in the near or distant future is assessed).

A number of applications of Big Data may be identified by drawing parallels with the well-established use in official statistics of administrative data, provided that the sources meet the benefit criteria and statistical validity issues [11]. These applications include for examples : providing auxiliary information such as stratification variables for samples frame, allowing full data substitution to replace surveys or partial data substitution for a population subgroup to reduce sample sizes, assisting the detection and treatment of anomalies in survey data, enriching the dataset without the need for statistical linking, creating richer datasets, ensuring the validity and consistency of survey data, generating new analytical insights, enhancing the measurement and description of economic, social and environmental phenomena.

Over the past years, NSOs have begun to look at the possibility of incorporating some of these sources in their production of economic and statistical indicators. Several experiments have been launched for this purpose. For example, INSEE (French National Institute of Statistics and Economic Studies) began exploring the possibility of using cash data (credit card slips from supermarkets) in the calculation of the consumer price index. This project aims to replace information gathered by the price collectors in the shops by an automatic process .The project covers the industrial food, hygiene and cleaning products sold in supermarkets and hypermarkets. One of the project's goals is to improve the accuracy of disaggregated indices. The project is currently at a stage of experimentation [8]. INSEE is also interested in the possibility of using user requests to forecast economic conditions, specially household consumption [2]. The potential of mobile data to improve tourism statistics has been studied as part of a project led by Eurostat [1] ; Statistics Netherlands built a household confidence indicator from social networking data [3]; Australian Bureau of Statistics is interested in forecasting agricultural production from satellite data [11].

4. Challenges

The use of Big data in official statistics faces several challenges. According to [4] and [13], the main challenges can be classified as follows :

- ***Legislative***, i.e. with respect to the access and use of data. Legislation in some countries (e.g. Canada) may provide the right to access data from both government and non-government organizations while others (e.g. Ireland) may provide the right to access data from public authorities only. The right to access admin data, established in principle by the law, is not adequately supported by specific obligations for Big data. Many potential Big Data sources are collected by non-governmental organizations or are ‘freely’ available on the web; situations that may not be covered by existing legislation ;

- ***Privacy***, i.e. managing public trust and acceptance of data re-use and its link to other sources. Privacy is generally defined as the right of individuals to control or influence what information related to them may be disclosed. The problem with Big data is that the users of services and devices generating the data are most likely unaware that they are doing so, and/or what it can be used for. The data would become even bigger if they are pooled, as would the privacy concerns.

- ***Financial***, i.e. potential costs of sourcing data vs. benefits. It is likely there will be a cost to acquire Big data, especially Big data held by the private sector and especially if legislation is silent on the financial modalities surrounding acquisition of external data. As a result, the right choices have to be made by NSOs, balancing quality (which encompasses relevance, timeliness, accuracy, coherence, accessibility and interpretability) against costs and reduction in response burden. Costs may even be significant for NSOs but the potential benefits far outweigh the costs, with Big data potentially providing information that could increase the efficiency of government programs.

- ***Management***, Big data for official statistics means more information coming to NSOs that is subject to policies and directives on the management and protection of the information to which NSOs must adhere. Another management challenge is one related to human resources.

The data science[1] associated with Big data that is emerging in the private sector does not seem to have connected yet with the official statistics community. The NSOs may have to perform in-house and national scans (academic, public and private sectors communities) to identify where data scientists are and connect them to the area of official statistics.

- ***Methodological*,** i.e. data quality and suitability of statistical methods. Representativity is the fundamental issue with Big data. The difficulty in defining the target population, survey population and survey frame jeopardizes the traditional way in which official statisticians think and do statistical inference about the target (and finite) population. With a traditional survey, statisticians identify a target/survey population, build a survey frame to reach this population, draw a sample, collect the data etc. They will build a box and fill it with data in a very structured way. With Big data, data comes first and the reflex of official statisticians would be to build a box! This raises the question is this the only way to produce a coherent and integrated national system of official statistics? Is it time to think outside of the box?

Also, the subpopulations covered by Big Data sources studied are not the target populations for official statistics. Therefore such data are likely to be selective, not representative of a relevant target population. Assessing representativity of Big Data may prove problematic, as often there are no characteristics readily available to conduct such comparison. Next, including the information content of Big Data sources in the statistical production process (often without unique statistical ID keys) makes integration challenging.

Another issue is both IT and methodological in nature. When more and more data is being analysed, traditional statistical methods, developed for the very thorough analysis of small samples, run into trouble; in the most simple case they are just not fast enough. There comes the need for new methods and tools :

* methods to quickly uncover information from massive amounts of data available, such as visualisation methods and data, text and stream mining techniques, that are able to ‘make Big data small’. Increasing computer power is a way to assist with this step at first ;

* methods capable of integrating the information uncovered in the statistical process, such as linking at massive scale, macro/meso-integration, and statistical methods specifically suited for large datasets. Methods need to be developed that rapidly produce reliable results when applied to very large datasets.

- ***Technological***, i.e. issues related to information technology. Dedicated and specialized computing infrastructures are required to cope with Big Data and enable processing and speed up analysis of large amounts of data. Certainly for the exploratory phase, during which the content and structure of Big Data sets has to be understood, fast technology certainly speeds up this process and more quickly enable the revelation of their use for statistics.

[1] Wikipedia defines data science as a science that incorporates varying elements and builds on techniques and theories from many fields, including mathematics, statistics, data engineering, pattern recognition and learning, advanced computing, visualization, uncertainty modelling, data warehousing, and high performance computing with the goal of extracting meaning from data and creating data products.

5. What about Morocco?

The data revolution is not restricted to the industrialized world [14]. The spread of mobile phone technology to the hands of billions of individuals may be the single most significant innovation that has affected developing countries in the past decade. Across the developing world, mobile phones are used daily to transfer money, buy and sell goods, and communicate information including test results, stock levels and prices of commodities. Mobile technology is used as a substitute for weak telecommunications and transport infrastructures as well as underdeveloped financial and banking systems.

The numbers of real-time information streams and people using social media are growing rapidly in developing countries as well. Based on the percentage of individuals using the internet, Morocco is listed as one of the most connected countries in Africa, the total number of internet users has reached 14.5 million at the end of 2015, representing a penetration rate of 42.8% and mobile phone subscribers have grown to 43 million, representing a penetration rate of 127.3% (Source : ANRT, the National Telecommunications Regulatory Agency of Morocco).

In Morocco, the use of Big Data is still in the embryonic state, but the awareness of its importance exists. In the private sector, a very limited number of companies and start-ups offers services and big data solutions (IBM, Bold Data...) and others particularly in the telecoms, insurance and banking sectors are closely interested in it.

In this paper, we have considered that is pertinent to highlight the strengths and weaknesses of the Moroccan statistical system, with the purpose of integrating and implementing Big Data sources into its information architecture, and how it can take advantage of external opportunities and deal with potential threats, in the form of a SWOT analysis– summarized in the table 2.

Table 2 : **SWOT** analysis for the moroccan national statistical system for integrating BIG DATA into its information architecture

STRENGTHS	WEAKNESSES
Moroccan statistical system has a solid, long tradition of census and surveys and has developed significant and relevant expertise in collecting and processing large amounts of data	Limited experience in the BIG DATA
	Methods for providing reliable official statistics based on BIG DATA sources are not well controlled or not known
The High commission for Planning of Morocco (main producer of official statistics) is empowered under the legislation to compel the provision of information by providers for the purposes of producing official statistics	Weak research and development (R&D) of key software technologies such as NoSQL(Not only Structured Query Language) and big data analytics.
	The culture of the public sector may be less tuned to what is required than the culture of the BIG DATA.
	Lack of technology to support all formats, current implementation has complex logic.
	Traditional IT infrastructure lacks scalability required for BIG DATA, new infrastructure is required
	Weak in vertical know-how as well as best practices of BIG DATA.
	Analytic skills should be developed in order to break the barrier between such amount of data and the meaningful information
	Official Institutions and Statistics departements have long and slow programming and budget cycles compared to the flexibility and responsiveness to developments- and new data needs- that characterises BIG DAT
OPPORTUNITIES	**THREATS**
Big Data can provide new statistical products and services, increase the frequency in the production of official statistics, more regular and timely information on interesting patterns such early indicators of epidemics, economic upturns or downturns, unemployment or housing boom etc.	Faced with increasing budget pressure, NSOs are not willing to invest a lot in Big Data
Big Data has the potential to increase the cost efficiency of NSOs, thanks to the lower unit cost of acquiring Big Data sources than the traditional direct data collection methods used by NSOs.	Potential scepticism from the staff of Official Institutions and Statistics departements towards new data as they see new technologies and Big Data as potential threats to their jobs.
Big Data may provide an opportunity for NSOs to better fulfil its mission in the provision of official statistics for informed decision making.	Is “Big Data technology” sufficiently mature to warrant an investment by the NSO?
Some Big data sources already exist in Morocco and could be used e.g credit card transactions, insurance and bank records, electronic medical records, motorway sensors in toll stations, online searches about a product/service, comments on social media...	The question of the sustainability and the independence of access to Big data sources arises. The stability of some of these data is not always assured (the case for the data of social networks, the use of which is volatile and subjected to the effects of mode)
Based on the percentage of individuals using the internet, Morocco is listed as one of the most connected countries in Africa, the total number of internet users has reached 14.5 million at the end of 2015, 42.8% penetration rate and 43 million of mobile phone subscribers 127.3% penetration rate (Source : The National Telecommunications Regulatory Agency (ANRT)	Technological risks because of the strong dependence on it
Moroccan universities are increasingly starting master’s programs on data science (Data scientists, Data analysts).	

Source : own elaboration

6. Conclusions and recommendations

This section gives conclusions and proposes some recommendations for NSOs. Presently, whilst not all Big Data variety is suitable for the production of official statistics, they have the potential to increase the cost efficiency of NSOs, provide new statistical products and services, and increase the frequency in the production of official statistics at little additional cost to NSOs. Big Data may also provide an opportunity for NSOs to better fulfill its mission in the provision of official statistics for informed decision making.

However, Big data represents a number of challenges and responsibilities for international and national statistical organisations, both methodological, technological, managerial, legal, and skills issues. Moroccan statistical organisations will thus be called upon in the years ahead to deal with Big data.

For many NSOs, big data projects would require significant investment in IT infrastructure. The challenge is that resources are reported to be limited due to budget constraints and human capital limitations (e.g. the need to train staff and develop specialized expertise to be able to handle and analyze large data sources). Moreover, it was felt that the IT implications of big data are complex, as specific software and hardware are required to ensure the adequate collection, storage, analysis and reporting of such information.

It is imperative for NSOs to foster the most important types of skills for working with Big Data : IT skills (noSQL databses, SQL databases and Hadoop), statistical skills (methodology and standards for processing Big Data, data mining) and other skills like creative problem solving, data governance and ethics. At present, there is insufficient training in these skills in Moroccan statistical offices. Researchers, statisticians and technicians should be able to master different tools and be ready to deal with the huge amount of data, so strong IT skills are needed. NSOs should develop the necessary internal analytical capability through specialised training. In fact, the processing of more and more data for official statistics requires "data scientists" and "data analytics", e.g. statistically aware people with an analytical mind-set, an affinity for IT and a determination to extract valuable ‘knowledge’ from data. International collaboration in this regard would be very beneficial for the official statistical community.

While a limited number of NSOs are actively engaged with technological aspects of Big data, it is mainly the private sector which leads the work on Big data analytics tools and solutions. Adapting Big data analytics tools and systems to official statistics will inevitably require the involvement of NSOs. Partnerships between NSOs and the private sector is of critical importance and it concerns issues such as privacy, trust and corporate competitiveness, as well as the legislation framework of the NSOs.

NSOs could play a new role in the future in terms of certification of statistics derived from Big data and used for public policy. With the proliferation of Big data into many aspects of life, is it NSOs that should assume this responsibility alone or as members of an independent multi-disciplinary authority?

Finally, we hope this article will be useful for those in charge of making policies or leading Big Data projects in developing countries and will stimulate discussion. The number of Big data projects and people using big data nowadays are increasing in developing countries indicating a diffusion of big data technologies. Diffusion of innovations can greatly accelerate adoption and utilization of Big Data, even though there are challenges faced by developing countries which limit capability and utilization of these technologies effectively. The Big Data adoption issues highlighted imply there is need for strategies for promoting technology diffusion including information dissemination, development of legal frameworks to deal with matters such as privacy concerns, cultural change, increasing digitization levels and investment in human capital particularly technical skills development.

References

[1] Ahas R., Armoogum J., Esko S., Ilves M., Karus E., Madre J.-L., Nurmi O., Potier F., Schmucker D., Sonntag U. and Tiru M., Feasibility study on the use of mobile positioning data for tourism statistics -consolidated report. Tech. rep., Consortium, June 2014.

[2] Bortoli C. and Combes S., Apports de Google trends pour prévoir la conjoncture française : des pistes limitées. In Note de Conjoncture. Insee, Mars 2015, pp. 43_56.

[3] Daas P., Puts M., Buelens B. and Van Den Hurk P., Big data and official statistics. In New Techniques and Technologies for Statistics, Eurostat, 2013.

[4] Daas P., Van Der Loo M., Big Data (and official statistics). Meeting on the Management of Statistical Information Systems, Paris, France and Bangkok, Thaïlande, 2013.

[5] Diebold F., On the Origin (s) and Development of the Term 'Big Data'. Penn Institute for Economic Research, 2012. Available at http://economics.sas.upenn.edu/pier/working-paper/2012/origins-and-development-term-%E2%80%9Cbig-data.

[6] Eurostat, United Nations, Economic and Social Council, Economic Commision for Europe, Conference of European, Big data - an opportunity or a threat to official statistics ?, April 2014.

[7] Laney D., 3d data management : Controlling data volume, velocity and variety. Technical Report 949, META Group (now Gartner), 2001.

[8] Leonard I., Varlet G. and Sillard P., Data editing and scanner data. In Conference of European Statisticians, United Nations Economic Commission for Europe, 2014.

[9] NIST Big Data Workshop (September 30, 2013). NIST Big Data Program. Retrieved April 10, 2015, Available at http://bigdatawg.nist.gov/home.php:

[10] Podesta J., Pritzker P., Monitz E., Holdren J. and Zients J., Big Data : Seizing opportunities, preserving values, Executive Office of the President, Washington, 2014. Available at http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

[11] Tam S.-M. and Clarke F., Big data, official statistics and some initiatives by the Australian Bureau of Statistics. ABS, 2014.

[12] TechAmerica Foundation, Federal Big Data Commission, Demystifying Big Data, a practical guide to transforming the business of government.

[13] United Nations Economic Commission for Europe, Conference of European Statisticians, What does "Big Data" mean for official statistics ?, March 2013.

[14] United Nations Global Pulse, Big Data for Development : A primer, 2013.

[15] United Nations, Fundamental Principles of Official Statistics, available at http://unstats.un.org/unsd/dnss/gp/FP-NEW-e.pdf.

[16] Zikopoulos P., Parasuraman K., Deutsch T. & Giles J. C., Harness the power of big data. The IBM big data platform. New York, McGraw Hill Professional, 2012.