May | 2018

Big Data for development
Year 10 – Number 1

Internet Sectoral Overview

Big Data & development: an overview[1]

Emmanuel Letouzé [2] *(Data-Pop Alliance)* [3]

Facts, Figures and the Big Picture

Despite the buzz, 'Big Data for development'—the field of research and practice about the applications and implications of Big Data for policymaking and development—remains in its intellectual and operational infancy. Is this "new oil" poised to be a blessing or a curse for human development and social progress? Optimists have called it a revolution that will change how we live, think and work, and have even expressed the hope that "Africa's statistical tragedy" may be partly fixed by Big Data. But skeptics and critics have been more circumspect, or plainly antagonistic, referring to Big Data as a big ruse, a big hype, a big risk, and, of course, 'Big Brother'. The Big Data buzz could just be a bubble, some observers point out - automated analysis of large datasets is not new. So what is new here?

[1] This is an edited version. To read the original article, please visit:http://datapopalliance.org/wp-content/uploads/2015/12/Big-Data-Dev-Overview.pdf

[2] Emmanuel Letouzé is the Director and co-founder of Data-Pop Alliance, a Visiting Scholar at the MIT Media Lab, a Research Affiliate at Harvard Humanitarian Initiative, and a Research Associate at Overseas Development Institute. He is the author of UN Global Pulse's White Paper "Big Data for development" (2012) and of the 2013 and 2014 OECD Fragile States reports. His research and work focus on Big Data's application and implications for official statistics, poverty and inequality, conflict, crime, and fragility, climate change, vulnerability and resilience, and human rights, ethics, and politics.

[3] Data-Pop Alliance is a think tank on Big Data and development jointly created by the Harvard Humanitarian Initiative (HHI), the MIT Media Lab, and the Overseas Development Institute (ODI) to promote a people-centered Big Data revolution.

From the 3 Vs to the 3 Cs of Big Data

Although there is no single agreed-upon definition of Big Data, it must be approached as a new ecosystem that is part and parcel of a larger social phenomenon driven by digital technology. This Big Data ecosystem can be characterized as the union of the 3 Cs of Big Data: crumbs, capacities and communities. This characterization is more accurate and less confusing than the 3 Vs of Big Data (volume, velocity and variety) used in the early years of Big Data, circa 2010-2012. The major limitation of the 3 Vs was their exclusive focus on Big Data as being just "Big Data." Another was their emphasis on Big Data being essentially a quantitative, rather than a qualitative, shift. It cannot be said loudly and clearly enough: Big Data cannot be constrained to Big Data, and the 3Vs ought to belong to history.

What are the 3 Cs? The first C, "crumbs," refers to "digital breadcrumbs," as suggested by MIT professor and Data-Pop Alliance Academic Director Alex "Sandy" Pentland. Unlike traditional survey data, these data are not produced for the purpose of statistical inference; instead, they are for the most part passively left behind by humans using digital devices and services, many of which were unavailable 5 or 10 years ago. Each of these actions leaves a digital trace; added up, they make up the bulk of Big Data as data. The volume of these data is growing fast and the proportion of digital data produced recently is growing ever faster —it is commonly reported that up to 90% of the world's data was created in the last year alone, although the exact source of the claim and estimation methodology are unclear.

These data come in 3 main types. One is small, "hard," structured data that can be easily quantified and organized (in columns and rows. for instance) for systematic analysis, and that cannot be edited by their emitters. Examples include call detail records (CDRs) and credit card transactions, as well as EZ pass or subway records. Some argue that this kind of data constitutes the real novelty and promise of Big Data; as explained by Pentland, "The power of Big Data is that it is information about people's behavior instead of information about their beliefs."

A second kind of data includes videos, online documents, blog posts and other social media content. These are "unstructured" data—harder to analyze in an automated fashion. They are also more subject to their authors' editorial choices: Someone may blog about boycotting a certain product, but their credit card statement may tell a different story.

A third type is gathered by digital sensors that pick up human actions, such as electric meters or satellite imagery that can pick up deforestation. Some consider the universe of Big Data as data to be much wider, to include administrative records, price or weather data, for instance, or books that have been previously digitized— which, taken collectively, may constitute a fourth kind. Importantly, as mentioned, what these data have in common is that they were not collected or sampled with the explicit intention of drawing conclusions from them. So the term Big Data is fundamentally misleading: Size isn't a defining feature, it is only a corollary of their nature. Even a small "Big Data" dataset can be Big Data if it doesn't stem from fully controlled processes like surveys and statistical analyses undertaken by official bodies.

> Although there is no single agreed-upon definition of Big Data, it must be approached as a new ecosystem that is part and parcel of a larger social phenomenon driven by digital technology.

CALL DETAIL RECORDS (CDRS)

Call detail records (CDRs) are metadata (data about data) that capture subscribers' use of their cell phones. These records include an identification code and, at a minimum, the location of the phone tower that routed the call for both caller and receiver and the time and duration of the call. Large operators collect over six billion CDRs per day.

CALLER ID	CALLER CELL TOWER LOCATION	RECIPIENT PHONE NUMBER	RECIPIENT CELL TOWER LOCATION	CALL TIME	CALL DURATION
X76VG588RLPQ	2°24' 22.14" 35°49' 56.54"	A81UTC93KK52	3°26' 30.47" 31° 12' 18.01"	2013-11-07T15:15:00	01:12:02

www.unglobalpulse.org/Mobile_Phone_Network_Data-for-Dev

If these data constitute the core of Big Data as an ecosystem, they do not constitute its whole. The second C of Big Data stands for capacities—tools, methods, software and hardware. As expressed by Harvard University professor Gary King, "Big Data is not about the data." These capacities include powerful computers, parallel computing systems, and statistical machine-learning techniques and algorithms that are able to look for and reveal patterns and trends in vast amounts of complex data.

The third C is for community. Big Data is also made up of the movement of individual and institutional actors that operate largely outside of traditional policy and research spheres. They are multidisciplinary teams of social and computer scientists with a "mindset to turn mess into meaning," as data scientist Andreas Weigend puts it. This also includes regular people using Google Maps to decide whether they will take their car or the subway to go to a meeting. More than ever, in the age of Big Data, everyone is a decision-maker.

This characterization suggests that Big Data is a complex system with feedback loops. New methods will yield new data; new data will give someone the idea of creating a data science start-up. It also shows that the phrase "using Big Data" is fundamentally missing the point, unless it is explicitly meant to say that one considers using the Big Data ecosystem to achieve certain goals. Rather, the question and challenge are why and how to engage with Big Data—to try to become part of it, affect its evolution, and/or benefit from its innovations.

The promise: supply and demand factors

The excitement over Big Data has stemmed from two factors: supply of ever more data and analytics capacities; and demand for better, faster and cheaper information. In other words, there has been and remains both a push for and a pull toward Big Data.

Availability of reliable and up-to-date data has been improving significantly over time; however, in many instances gaps remain. For instance, a good indicator of a region's poverty or underdevelopment is lack of poverty or development data. Some countries (most of them with a recent history of conflict) haven't had a census in four decades or more. Their population size, structure and distribution can often only be approximated with triangulation of information from different sources. Even though official figures exist, they are often based on incomplete data. Poor data also mean that some countries' official GDP figures get an overnight boost — 40% for Ghana in 2010, or 60% for Nigeria in 2014 — when changes in the structure of their economies, such as the rise of the technology sector, are finally taken into account.

This lack of reliable data led to the call for a "data revolution" that led to the publication of a report by a UN-appointed experts group. The basic and somewhat simplistic rationale is that, in the age of Big Data, economies should be steered by policymakers relying on better navigation instruments and indicators that let them design and implement more agile and better-targeted policies and programs. Big Data has even been said to hold the potential for national statistical systems in data-poor areas to leapfrog ahead, much as many poor countries skipped the landline phase to jump straight into the mobile phone era.

The excitement over Big Data has stemmed from two factors: supply of ever more data and analytics capacities; and demand for better, faster and cheaper information.

The appeal of potentially leaping ahead is also shaped by the supply side of Big Data. There is early practical evidence and a growing body of work on Big Data's novel potential to help understand and affect human populations and processes. For example, Big Data has been used to track inflation online, estimate and predict changes in GDP in near real-time, and monitor traffic or the outbreak of epidemics. Monitoring social media data to analyze people's sentiments is opening new ways to measure welfare, while email and Twitter data could be used to study internal and international migration. And an especially rich academic literature uses CDRs to study migration patterns, socioeconomic levels, and the spread of disease, among others. Because smartphones will soon overtake regular cell phones around the globe, CDR analysis will recede and new crumbs will become the next frontiers.

Meanwhile, taxonomies have been proposed to clarify how Big Data could benefit development. One taxonomy distinguishes early warning uses from real-time awareness, or from real-time monitoring of the impact of a policy or program. Another contrasts its descriptive function (such as a real-time map) from predictive and prescriptive applications. The following table provides examples of applications falling under each use category.

The predictive use can be understood in two senses of the term: as inference or nowcasting—predicting what is happening right now (such as when cell phone activity is used to predict socioeconomic levels); or forecasting (in a fashion very similar to what meteorologists do). The prescriptive use requires making causal inferences; i.e., establishing the existence, direction and magnitude of a causal link between some treatment or variable X and some effect or variable Y.

The grey side of Big Data: risks and challenges

The promise of Big Data's applications to real-world problems has been met with warnings about its perils, and more broadly, active discussions about its social implications. Perhaps the most severe risks are to individual and group rights, privacy, identity, and security. In addition to the obvious intrusion of surveillance activities and issues around their legality and legitimacy, there are important questions about data anonymization: what it means and its limits. An early study of movie rentals showed that even anonymized data could be de-anonymized —linked to a known individual by correlating rental dates of as few as three movies with the dates of posts on an online movie platform. Other research has found that CDRs that record location and time, even when free of any individual identifier, could be re-individualized, which is referred to as re-identification. In that case, four data points were theoretically sufficient to uniquely single out individuals in an entire dataset with 95% accuracy. Recent research using credit card transactions yields very similar conclusions: Our behaviors are unique, and predictable enough to make it very hard for any given individual to hide in the digital crowd.

Critics also point to the basic risks associated with basing decisions on analyses lacking external or internal validity.

Another risk is that analyses based on Big Data will focus too much on correlation and prediction, at the expense of cause, diagnostics or causal inference, without which policy is essentially blind. A good example is predictive policing. Police and law enforcement forces in some US and UK cities have crunched data to assess the likelihood of increased crime in certain areas for years, predicting rises based on historical patterns. Forces dispatch their resources accordingly, and this has reduced crime in most cases. However, unless there is knowledge of why crime is rising, it is difficult to put in place a preventive policy that tackles root causes or contributing factors. At the same time, proponents argue that cracking down on crime in an area may have a cumulative structural effect.

Yet another big risk that is receiving growing attention is Big Data's potential to create a new digital divide that may widen rather than close existing gaps in income and power worldwide. One of the three paradoxes[4] of Big Data is that because it requires analytical capacities and access to data that only a fraction of institutions, corporations and individuals have, it may disempower the very communities and countries it promises to serve. People with the most data and capacities are in the best position to develop Big Data systems to their economic and political advantage, even as they claim to use them to benefit others.

A last basic challenge is that of putting the data to use—having a fundamental understanding of how data has affected societies historically. Most discussions about the data revolution assume that data matter, and that poor data are to

Perhaps the most severe risks are to individual and group rights, privacy, identity, and security.

[4] To delve into the three paradoxes, see: Richards, N.M. and King, J.H. (2013). Three paradoxes of Big Data. Stanford Law Review.

blame for poor policies. But lack of data has historically played only a marginal role in the decisions leading to bad policies and poor outcomes. And a blind algorithmic future may undercut the very processes that are meant to ensure that the way data are turned into decisions is subject to democratic oversight. At the same time, there is tremendous potential for societies to understand and affect processes that they have grappled with for centuries. Fulfilling that promise will require profound reframing and rewiring of our political, ethical, technological and legal systems.

Big Data: risks to drawing valid conclusions

A key challenge in Big Data is that the people generating it have selected themselves as data generators through their activity. In technical terms this is a selection bias, and it means that analysis of Big Data is likely to yield different results from traditional surveys (or polls), which seek out a representative cross-section of the population. For example, trying to answer the question "Do people in country A prefer rice or chips?" by mining data on Twitter would be biased in favor of young people's preferences, since they make up more of Twitter's users. So analyses based on Big Data may lack external validity, although it is possible that individuals who differ in almost all respects may have similar preferences and display identical behaviors (young people may have the same preferences as older people). Another risk comes from analyses that are flawed because they lack internal validity. For instance, a sharp drop in the volume of CDRs from an area might be interpreted, based on past events, as signaling a looming conflict. But it could actually be caused by something different, such as a mobile phone tower having gone down in the area.

Big Data's future or the future's Big Data?

Since the growth in data production is highly unlikely to abate, and human creativity and curiosity are almost limitless, the Big Data bubble is unlikely to burst in the near future. The world can expect more discussions and controversies about Big Data's potential and perils for development and societies at large. The future of Big Data will probably be shaped by three main related strands: academic research; legal and technical frameworks for ethical use of data; and larger social demand for greater accountability and participation.

Research will continue to examine whether and how methodological and scientific frontiers can be pushed, especially in two areas: drawing stronger causal inferences and measuring and correcting sample bias.

Policy debate will develop frameworks and standards — normative, legal and technical — for collecting, storing and sharing Big Data streams and sets. These developments fall under the umbrella term "ethics of Big Data." Technical advances will help, for example, by injecting noise into datasets to make re-identification of the individuals represented in them more difficult, although they will probably never make it impossible. But a comprehensive approach to the ethics of Big Data would ideally encompass other humanistic considerations such as privacy and equality, as well as champion data literacy and human-centered design.

A third related influence on the future of Big Data will be how it engages and evolves alongside the open data movement and its underlying social drivers — where open data refers to data that is easily accessible, machine-readable, accessible for free or at negligible cost, and has minimal limitations on its use, transformation, and distribution.

For the foreseeable future, the Big Data and open data movements will be the two main pillars of a larger data revolution. Both have arisen against a background of increased public demand for more openness, agility, transparency, accountability and participation. The political overtones — so easily forgotten — are clear. A true Big Data revolution should be one where data can be leveraged to change power structures and decision-making processes, not just create insights.

> The future of Big Data will probably be shaped by three main related strands: academic research; legal and technical frameworks for ethical use of data; and larger social demand for greater accountability and participation.

REFERENCES

References and selected bibliography can be accessed from the original document, available at http://datapopalliance.org/wp-content/uploads/2015/12/Big-Data-Dev-Overview.pdf.

Table 1 –TAXONOMIES OF ACTUAL AND POTENTIAL USES OF BIG DATA FOR DEVELOPMENT

	APPLICATIONS	EXPLANATION
UN GLOBAL PULSE REPORT TAXONOMY (Letouzé, 2012)	1- EARLY WARNING	Early detection of anomalies in how populations use digital devices and services can enable faster response in times of crisis.
	2- REAL-TIME AWARENESS	Big Data can paint a fine-grained and current representation of reality which can inform the design and targeting of programs and policies.
	3- REAL-TIME FEEDBACK	The ability to monitor a population in real time makes it possible to understand where policies and programmes are failing, and make the necessary adjustments.
ALTERNATIVE TAXONOMY (Letouzé et al., 2013)	1- DESCRIPTIVE	Big Data can document and convey what is happening.
	2- PREDICTIVE	Big Data could give a sense of what is likely to happen, regardless of why.
	3- PRESCRIPTIVE	Big Data might shed light on why things may happen and what could be done about it.

	EXAMPLES	COMMENTS
	Predictive policing is based upon the notion that analysis of historical data can reveal certain combinations of factors associated with greater likelihood of crime in an area; it can be used to allocate police resources. Google Flu trends is another example, where searches for particular terms ("runny nose", "itchy eyes") are analyzed to detect the onset of the flu season — although its accuracy is debated.	This application assumes that certain regularities in human behaviours can be observed and modelled. Key challenges for policy include the tendency of most malfunctiondetection systems and forecasting models to over-predict — i.e. to have a higher prevalence of 'false positives'.
	Using data released by Orange, researchers found a high degree of association between mobile phone networks and language distribution in Ivory Coast — suggesting that such data may provide information about language communities in countries where it is unavailable.	The appeal for this application is the notion that Big Data may be a substitute for bad or scarce data; but models that show high correlations between 'Big Databased' and 'traditional' indicators often require the availability of the latter to be trained and built. 'Realtime' here means using high frequency digital data to get a picture of reality at any given time.
	Private corporations already use Big Data analytics for development, which includes analysing the impact of a policy action — e.g. the introduction of new traffic regulations — in real-time.	Although appealing, few (if any) actual examples of this application exist; a challenge is making sure that any observed change can be attributed to the intervention or 'treatment'. However highfrequency data can also contain 'natural experiments' — such as a sudden drop in online prices of a given good — that can be leveraged to infer causality.
	This application is quite similar to the 'real-time awareness' application — although it is less ambitious in its objectives. Any infographic, including maps, that renders vast amounts of data legible to the reader is an example of a descriptive application.	Describing data always implies making choices and assumptions — about what and how data are displayed — that need to be made explicit and understood; it is well known that even bar graphs and maps can be misleading.
	One kind of 'prediction' refers to what may happen next —as in the case of predictive policing. Another kind refers to proxying prevailing conditions through Big Data—as in the cases of socioeconomic levels using CDRs in Latin America and Ivory Coast.	Similar comments as those made for the 'early-warning' and 'real-time awareness' applications apply.
	So far there have been few examples of this application in development contexts.	Most comments about 'real-time feedback' apply. An example would require being able to assign causality. The prescriptive application works best in theory when supported by feedback systems and loops on the effect of policy actions.

Big Data in practice: Cetic.br projects

DIGITAL ECONOMY INDICATORS PROJECT: USE OF WEB SCRAPING[5] FOR PRODUCING ICT INDICATORS FOR COMPANIES

TOPIC	Electronic commerce.
PARTNERS	The project falls within a context of collaboration between the Economic Commission for Latin America and the Caribbean (ECLAC) and the Regional Center for Studies on the Development of the Information Society (Cetic.br) to encourage and develop methodologies for measuring the digital economy in Latin America and the Caribbean through Big Data and data analytics tools.
CONTEXT	Conducted annually by Cetic.br since 2005, the ICT Enterprises survey measures the presence of information and communication technologies (ICT) in companies with ten or more employed persons. The objective of the survey is to investigate access to infrastructure, as well as the use and appropriation of new technologies by the private sector. Among the different topics studied, the survey applies a module on electronic commerce, resulting in online buying and selling indicators.
USE OF BIG DATA	The Big Data project seeks to produce indicators on electronic commerce in companies from the automated collection of data, using data scraping on the websites of companies contacted for the ICT Enterprises survey. Some of the data collected includes: proportion of enterprises that offer product and services catalogues, price lists, ordering systems, online payment and customer support on their websites; proportion of enterprises that buy on the Internet; and proportion of enterprises that sell on the Internet and the channels used – email, social networks, or group purchasing websites. Among the expected results were: accuracy assessment of the modeling of data from Big Data sources for estimating certain ICT indicators in companies; development of a tool for automated collection of data on the Internet; and development of a keyword dictionary (semantic context).
STATUS	Pilot project in progress.

[5] Web scraping is a data extraction technique used for collecting data/content from websites. Through automated processes, this type of information scraping is a way to obtain copies of data from websites, converting it into structured information for subsequent analysis.

BROADBAND INDICATORS: INTERNET TRAFFIC MEASUREMENT SYSTEM (SIMET) DATABASES	
TOPIC	Broadband quality.
PARTNERS	The project is the result of a partnership between the Center of Study and Research in Network Technology and Operations (Ceptro.br), a department of the Brazilian Network Information Center (NIC.br), which is responsible for data collection with the Simet[6] tool, and Cetic.br, which is responsible for the data analysis.
CONTEXT	Different data sources can be used to measure broadband Internet quality, ranging from administrative records to the perceptions of users regarding the quality of services provided by operators. Cetic.br produces data on Internet connection that derive from the information provided by respondents for their sample surveys. However, for a more complete evaluation, other broadband quality data sources were examined.
USE OF BIG DATA	For ten years, Simet has been periodically collecting data on Brazilian Internet quality, in real time, through tests performed by users. The measurements, taken at 11-second intervals, occur through sending and receiving data packets in order to produce various quality indicators in relation to the operation in progress. The database resulting from these measurements has Big Data characteristics, since there is a large volume of data that is updated very quickly (new measurements every moment) and generated by user-defined events, i.e., information that is collected if the user wishes, and not from a sample selection. Based on the analysis of the data from this database, and taking into account its characteristics and limitations, the data can be supplemented through sample surveys. Therefore, it is possible to obtain a broader perspective on broadband evolution in Brazil.
STATUS	Project in progress.

[6] The Internet Traffic Measurement System (Simet) is a set of systems that tests Internet quality. It takes independent measurements, automated or manually, triggered by users and fully supported on the NIC.br infrastructure. These measurements are currently taken through three different devices/applications: Simet Box, Simet Mobile and Simet Applet. For more information, go to: https://simet.nic.br/index.php

Interview

Roberto Olinto is the President of the Brazilian Institute of Geography and Statistics (IBGE)

Big Data and official statistics: challenges and opportunities

Roberto Olinto, President of the Brazilian Institute of Geography and Statistics (IBGE), comments on the potential opportunities and challenges that technological changes, especially Big Data, have on the production of official statistics.

I.S.O._ What actions is IBGE implementing in relation to [the use of] Big Data?

R.O._ Before implementing actions, an official statistics institution has to allot time on its agenda for reflecting on the impact of Big Data on the institution, i.e., what the information generated by Big Data causes, which is not a very clear phenomenon. How should a statistics institution position itself in relation to this amount of information? Our position at IBGE focuses on the area of communication, which is how we present ourselves in an increasingly comprehensive manner, using different media, but also as increasingly transparent, explaining what we produce, given the amount of information available nowadays and that can be contrary to our work.

An internal committee has been created at IBGE. We tried to recruit representatives from all the departments to discuss exactly what IBGE's internal position would be in relation to Big Data. This will inform a larger policy and even incorporate international discussions on the use of Big Data for official information, for which a work group already exists. However, there are already initiatives in certain departments of IBGE for dealing with Big Data.

The first key question is: How will this technology enable us to gather data more quickly and efficiently and, fundamentally, according to the methodology that we have adopted, and not be a chaotic search for data? Then, we will need to reflect on whether generating data will help us in some way and how it will help. However, this is a later step, because, first, we have to look at the amount of information that is being generated and in what way it can impact the statistics institution, in terms of communication; and second, implement the use of these tools, search methods and other things of this nature, to improve how we capture information.

I.S.O._ What are the main opportunities and advantages of [using] Big Data for official statistics?

R.O._ It is essentially about speed in obtaining information. Another more complicated point that has been discussed internally is whether we can use private databases for creating statistics. The biggest example is the issue of mobile telephony: What if we could have access to mobile phone databases? It would obviously be without [user] identification, since we would not need the informant's identification, which is also not a problem, because we are protected by secrecy. Then we would be able to develop statistics on mobility, flow, influence of cities, and so on.

The main issue is that this information is not provided free of charge, and the statistics institution does not have money or, I should say, their budget for this is limited. Nor can you be left, at any given moment, at the mercy of database producers. This matter, in particular, has to be discussed, and some countries are already addressing it: Whether it is possible to have a law that would enable moving forward with this idea of database sharing.

There are international experiences on the use of databases from major supermarkets for price tracking. Now, Big Data can be very useful particularly in terms of the collection and handling of information, through electronic invoices in Brazil.

Electronic invoices contain an enormous amount of information that can be utilized in the new processes, whether new systems or new storage capacity, which is one of the characteristics of Big Data. We are discussing this information, which is now public, with finance secretariats. At the state level, we will discuss this at Confaz (the National Council of Financial Policy). In my opinion, this would be the first major statistical operation using Big Data: Use the entire electronic invoice to improve IBGE statistics.

I.S.O._ What are the main challenges to be tackled to be able to harness the potential benefits of [using] Big Data?

R.O._ The challenges include: access to technology and funding for doing updates; having a technical team that is large enough, since at IBGE teams are hired on the basis of competitive examinations, and in order to have a profile where it is possible to incorporate data analysts, the own team will need to be trained. This is a challenge given the speed with which technology advances.

As I said earlier, it is a challenge to have ongoing access to private databases or create public databases, for example, based on electronic invoices. This would require having legislation that looks at the mass of data generated today, through administrative records, and being able to operate it with Big Data tools and have access to information. Perhaps a side challenge would be changing the culture within the statistics institute, introducing Big Data logic.

‘The first key question is: How will this technology enable us to gather data more quickly and efficiently and, fundamentally, according to the methodology that we have adopted, and not be a chaotic search for data?’.

I.S.O._ What is IBGE’s position in relation to integration between the various statistical data producers?

R.O._ IBGE is a strong and deeply committed advocate of this idea. Nowadays, the issue of interoperability and database sharing is key to advancement of statistical systems. Integration of information producers and databases, as well as transparency in methodologies, is the path for countries to take in their information systems. We defend this idea and are working for it to happen as quickly as possible. Obviously, technical challenges are easy to solve. However, the biggest issue is institutional challenges. The biggest problem is having laws and a culture that accepts the logic behind database sharing without any major issues, or sees it as a natural step in the evolution of the information system, i.e., in the statistics and geoinformation system.

I.S.O._ How can the use of Big Data help national statistics institutions to measure the Sustainable Development Goals (SDGs)?

R.O._ It will help us with all the information, since the SDGs are an additional demand today and require us to improve and expand our statistical production. This means that Big Data tools are essential, not only for achieving the SDGs, but also for meeting all new demands, which are not restricted to the SDGs, but arise from a number of other sources.

Interview II

Ronald Jansen is Director and Chief of Data Innovation and Capacity Branch, United Nations Statistics Division (UNSD)

Communities and Big Data: the role of data collaboratives

To comment on the role of multi-stakeholder collaboration models being developed for the use of Big Data, we interviewed Ronald Jansen, Assistant Director and Chief of Data Innovation and Capacity Branch, United Nations Statistics Division (UNSD).

I.S.O._ What are data collaboratives and why are they important for using Big Data for official statistics?

R.J._ Data collaboratives – within the context of the global statistical community – are projects in which national statistical offices, the private sector and other partners work together in a secure environment on Big Data and other data sources to experiment with producing statistics and indicators, which could be used for official policy purposes. One example of a data collaborative is the collaboration of several statistical offices, the UN Statistics Division and Nielsen, a private sector company specialized in collecting price data, to test new statistical methods using scanner data (provided by Nielsen) for the calculation of price indices, most notably the consumer price index. Within this data collaborative, not all partners have the same access rights to all data, but everyone will have access to the algorithms and code developed in the tested methods.

The purpose of the data collaborative is to develop trusted methods and applications with trusted data and trusted partners, the results of which can be transferred to the wider community through trusted learning. In other words, the data collaborative results in tested methods, new tools, new algorithms and new applications, which are made available to the whole statistical community.

The broader objective of such collaboration on Big Data is to stay relevant as a statistical community. Statistical offices need to innovate and keep up with changes in technology, available data sources, and the expectation of the current society. The public nowadays expects very fast and detailed information of good enough quality. The commercial world is providing a lot of data very quickly, and statistical offices have to compete to stay relevant.

I.S.O._ How can national statistical offices, technology companies and data owners collaborate in a mutually beneficial way in a changing world, in which data are seen as the most important source for creating wealth and development for all?

R.J._ I gave an example on the use of scanner data. Another example of a data collaborative is the use of satellite data for the measurement of freshwater extent (SDG indicator 6.6.1[7]). In this case, several statistical offices teamed up with UN Environment, the UN Statistics Division, and Google. Google brought into the project the use of the Google Earth engine, as well as the necessary Cloud server space and computing power. The statisticians brought their subject matter and methodological expertise to develop and test algorithms. In principle, the outcome of this data collaborative could help each country in the world with the calculation of SDG indicator 6.6.1.

I.S.O._ What are the experiences and lessons learned from existing data collaboratives in relation to coverage, inclusion (and exclusion) of partners, activities, management and financing?

R.J._ The two data collaboratives that I mentioned involve mostly offices of developed countries. Although the overall governance lies with the UN Global Working Group (GWG) on Big Data, the participating project partners will determine the direct management of resources. Financial implications have been kept to a minimum for now.

The GWG has set itself a goal to develop a proof of concept by March 2020. This proof of concept of working in data collaboratives with the private sector should include a viable and sustainable business model (or various business models), which could be working on a not-for-profit basis, or on a commercial basis, with credits and discounts for those who can contribute less. It should therefore also include evidence that developing countries can participate and can benefit from the new data sources, methods, tools and applications.

I.S.O._ What is the Cape Town Global Action Plan for Sustainable Development Data and how does it relate to the so-called "data revolution"?

R.J._ The "data revolution" was a call for action to energize the statistical community (and the data community in a broader sense) to innovate and to make use of all the data that was automatically being generated in the digitized world.

The Cape Town Global Action Plan is a more comprehensive call to strengthen national statistical offices. The use of new data sources is only one part of it. Some of the points of the action plan are to modernize governance, institutional frameworks, and statistical standards; to facilitate the application of new technologies and new data sources in mainstream statistical activities; to expand the use of administrative records; and to strengthen partnerships with governments, academia, civil society, the private sector and other stakeholders involved in the production and use of data for sustainable development.

> 'The public nowadays expects very fast and detailed information of good enough quality. The commercial world is providing a lot of data very quickly, and statistical offices have to compete to stay relevant'.

[7] Indicator 6.6.1 refers to the change in the extent of water-related ecosystems over time and is part of SDG 6 on ensuring availability and sustainable management of water and sanitation for all.

Domain Report

The dynamics of the registration of domains in Brazil and the world

The Regional Center for Studies on the Development of the Information Society (Cetic.br) carries out monthly monitoring of the number of domain names registered in the 16 largest country code Top-Level Domains (ccTLDs) in the world. Combined, they exceed 97.6 million registrations.

In May 2018, the domains registered under .tk (Tokelau) reached 20.71 million, followed by Germany (.de), China (.cn) and the United Kingdom (.uk), with 16.31 million, 11.06 million and 9.93 million records, respectively[8]. Brazil continues to occupy the seventh place on the list, with 3.95 million registrations under .br. With 1.91 million registrations, Spain (.es) ranked 16th, as can be seen in Table 2.

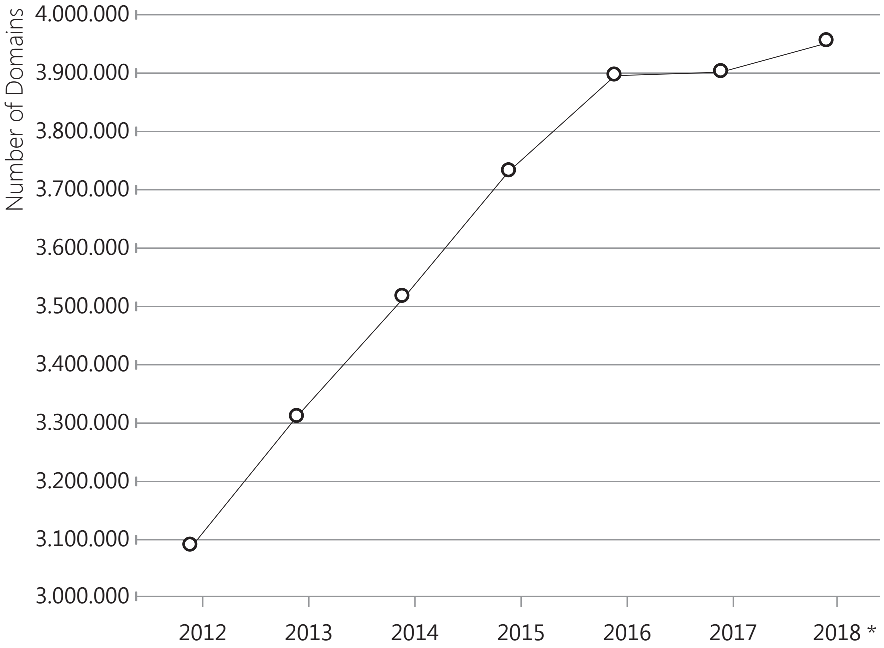
Table 2 – REGISTRATION OF DOMAIN NAMES IN THE WORLD – MAY 2018

Position	ccTLD	Domains	Ref.	Source
1	Tokelau (.tk)	20.715.276	May-18	http://research.domaintools.com/statistics/tld-counts/
2	Germany (.de)	16.308.345	May-18	www.denic.de/
3	China (.cn)	11.059.726	May-18	http://research.domaintools.com/statistics/tld-counts/
4	United Kingdom (.uk)	9.926.573	Jan-18	https://www.nominet.uk/uk-register-statistics-2018/
5	Netherlands (.nl)	5.806.100	May-18	www.sidn.nl
6	Russia (.ru)	5.239.190	May-18	www.cctld.ru
7	**Brazil (.br)**	**3.952.679**	**May-18**	**registro.br/estatisticas.html**
8	European Union (.eu)	3.717.165	May-18	http://research.domaintools.com/statistics/tld-counts/
9	France (.fr)	3.212.900	May-18	https://www.afnic.fr/en/resources/statistics/detailed-data-on-domain--names/
10	Australia (.au)	3.152.285	May-18	www.auda.org.au
11	Italy (.it)	3.133.248	May-18	www.nic.it/
12	Canada (.ca)	2.748.296	May-18	www.cira.ca/
13	Poland (.pl)	2.564.636	May-18	www.dns.pl/english/zonestats.html
14	Switzerland (.ch)	2.147.841	Mar-18	https://www.nic.ch/reg/cm/wcm-page/statistics/index.html?lid=em*
15	United States (.us)	1.966.460	May-18	research.domaintools.com/statistics/tld-counts/
16	Spain (.es)	1.911.574	May-18	www.dominios.es

[8] It is important to note that variations exist among ccTLD reference periods, although it is always the most updated one for each country that is used.

Graph 1 shows the performance of .br since 2012.

Graph 1 – TOTAL NUMBER OF DOMAIN REGISTRATIONS PER YEAR FOR .BR – 2012 TO 2018*



**Data in reference to May 2018.*
Source: Registro.br

In May 2018, the five generic Top-Level Domains (gTLD) totaled more than 167 million registrations. With 134.41 million registrations, the .com ranked first, as shown in Table 2.

Table 2 – MAIN GTLDS – MAY 2018

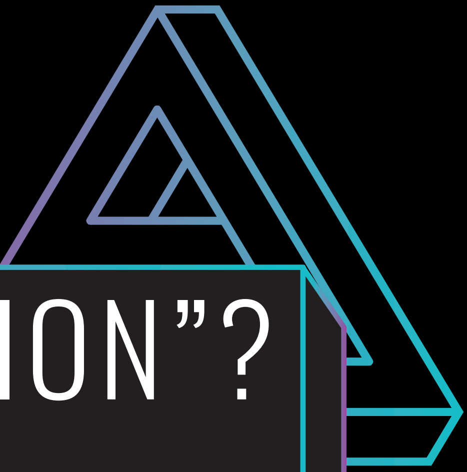
Position	gTLD	Domains
1	.com	134.407.747
2	.net	14.268.162
3	.org	10.396.758
4	.info	6.105.279
5	.biz	1.983.870

Source: DomainTools.com
Retrieved from: http://research.domaintools.com/statistics/tld-counts/

/Answers to your questions

BIG DATA? “DATA REVOLUTION”? WHAT IS THAT?

Here are definitions of key terms and concepts[9]:



Big Data

An umbrella term signifying one or more of three trends: the growing volume of digital data generated as a byproduct of the use of digital devices by people on a daily basis; new technologies, tools and methods available to analyze large datasets not originally designed for analysis; and the intention to extract from these data and tools insights that can be used for policymaking.

Data revolution

A common term in development discourse since the high-level panel of eminent persons on the Post-2015 Development Agenda called for a “data revolution” to “strengthen data and statistics for accountability and decision-making purposes.” It refers to a larger phenomenon than Big Data or the “social data revolution,” which is defined as the shift in, and implications of, human communication patterns toward greater personal information sharing.

Algorithms

In computer science, an algorithm is a series of predefined instructions or rules written in a programing language that is designed to tell a computer how to sequentially solve a recurrent problem, especially as it involves making calculations and processing data. There is a growing use of algorithms for decision-making purposes in an increasing array of activities and industries, such as policing and banking.

[9] Adapted from Letouzé, E. (2015). Big Data and development: General overview primer. Data Pop Alliance White Paper Series. Data Pop Alliance, World Bank Group, Harvard Humanitarian Initiative. Retrieved May 10, 2018 from http://datapopalliance.org/wp-content/uploads/2015/12/Big-Data-Dev-Overview.pdf

Statistical machine learning

Refers to the construction and study of computer algorithms — step-by-step procedures for calculations and/or classification — that can teach themselves to grow and change when exposed to new data. It represents the ability to "learn" to make better predictions and decisions based on what was experienced in the past, as with spam filtering, for example. The addition of "statistical" reflects the emphasis on statistical analysis and methodology, which is the predominant approach to modern machine learning.

Call detail records (CDRs)

The technical name of cell-phone data recorded by telecom operators. CDRs contain information about the time and duration of calls, and the locations of senders and receivers of calls or text messages transmitted through their networks.

(New)

Digital divide

Differential access to and ability to use new information and communication technologies between individuals, communities and countries, and resulting inequalities in socioeconomic and political opportunities and outcomes. The skills and tools required to absorb and analyze large amounts of data resulting from the growing use of technology may provide the foundation for the creation of a "new digital divide."

/Credits

TEXT
MAIN ARTICLE
Emmanuel Letouzé (Data-Pop Alliance)
DOMAIN REPORT
José Márcio Martins Júnior (Cetic.br)

EDITORIAL COORDINATION
Alexandre Barbosa (Cetic.br)
Tatiana Jereissati (Cetic.br)
Stefania L. Cantoni (Cetic.br)

ACKNOWLEDGMENTS
Emmanuel Letouzé (Data-Pop Alliance)
Roberto Olinto (Brazilian Institute of Geography and Statistics – IBGE)
Ronald Jansen (United Nations Statistics Division – UNSD)

GRAPHIC DESIGN AND LAYOUT
Comunicação NIC.br

REVIEW IN PORTUGUESE
Aloisio Milani (Magma Editorial Ltda.)
Alexandre Pavan (Magma Editorial Ltda.)

TRANSLATION INTO ENGLISH
Grant Borowik, Lorna Simons, Luana Guedes, Luisa Caliri
(Prioridade Ltda.)

PUBLISHING OF ENGLISH EDITION
Grappa Marketing Editorial

The Internet Internet Sectoral Overview is also available in Portuguese at cetic.br/publicacoes/indice/panoramas/

United Nations
Educational, Scientific and
Cultural Organization

Regional Centre of Studies for the
Development of the Information
Society under the auspices of UNESCO

Brazilian Network Information Center

Brazilian Internet Steering Committee

STRIVING FOR A BETTER INTERNET IN BRAZIL

CGI.BR, MODEL OF MULTISTAKEHOLDER GOVERNANCE

www.cgi.br

nic.br cgi.br

FOTOS: RICARDO MATSUKAWA E SHUTTERSTOCK