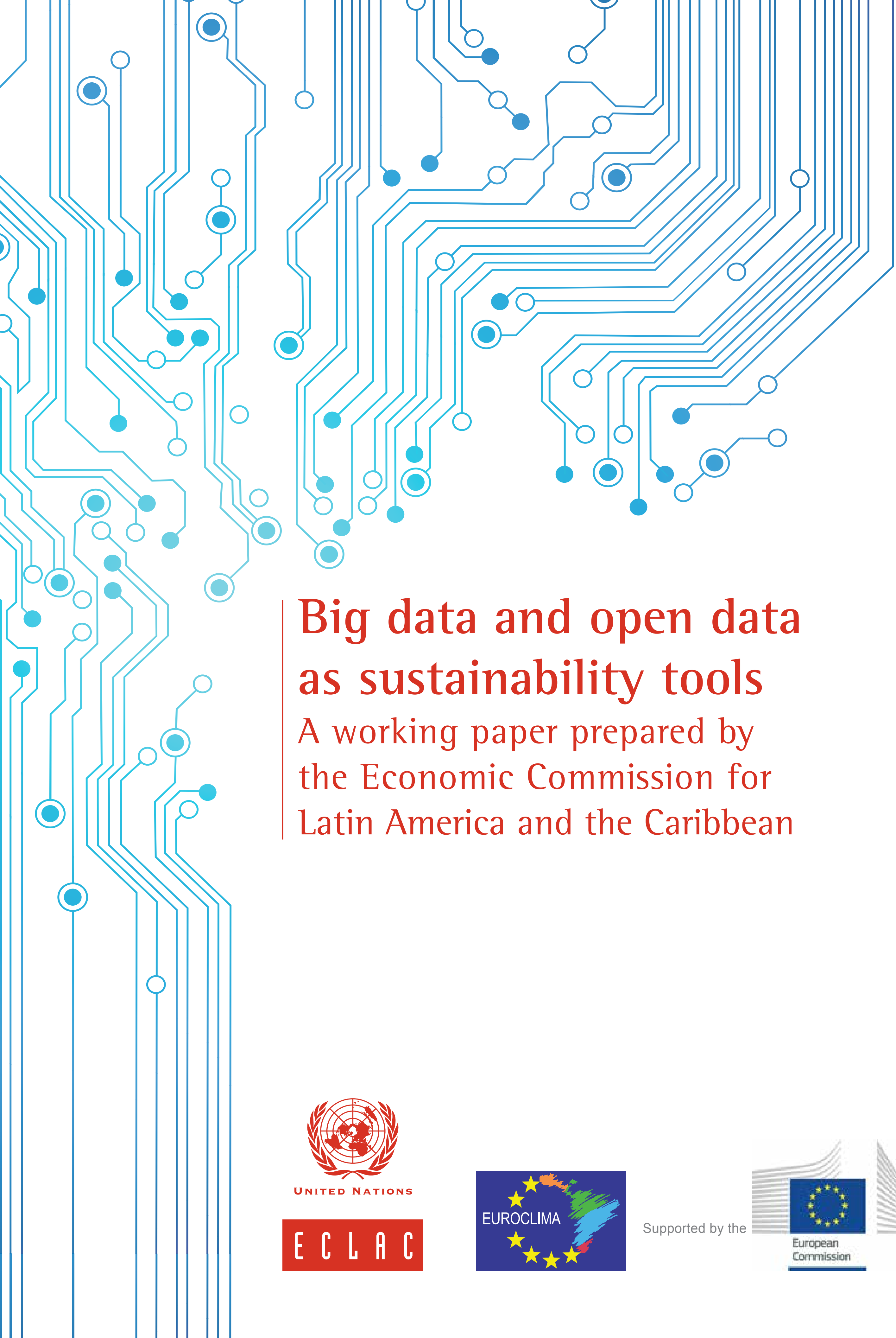


Big data and open data as sustainability tools

A working paper prepared by the Economic Commission for Latin America and the Caribbean

United Nations

ECLAC

EUROCLIMA

Supported by the

European Commission

Project Document

Big data and open data as sustainability tools

A working paper prepared by the Economic Commission for Latin America and the Caribbean

United Nations
ECLAC

EUROCLIMA

Supported by the European Commission

Economic Commission for Latin America and the Caribbean (ECLAC)

This paper was prepared by Wilson Peres in the framework of a joint effort between the Division of Production, Productivity and Management and the Climate Change Unit of the Sustainable Development and Human Settlements Division of the Economic Commission for Latin America and the Caribbean (ECLAC). Some of the imput for this document was made possible by a contribution from the European Commission through the EUROCLIMA programme.

The author is grateful for the comments of Valeria Jordán and Laura Palacios of the Division of Production, Productivity and Management of ECLAC on the first version of this paper. He also thanks the coordinators of the ECLAC-International Development Research Centre (IDRC)-W3C Brazil project on Open Data for Development in Latin America and the Caribbean (see [online] www.od4d.org): Elisa Calza (United Nations University-Maastricht Economic and Social Research Institute on Innovation and Technology (UNU-MERIT) PhD Fellow) and Vagner Diniz (manager of the World Wide Web Consortium (W3C) Brazil Office).

The European Commission support for the production of this publication does not constitute endorsement of the contents, which reflect the views of the author only, and the Commission cannot be held responsible for any use which may be made of the information obtained herein.

LC/W.628

Printed at United Nations, Santiago, Chile

Contents

Tables

Figures

Boxes

Diagrams

Introduction

The two main forces affecting economic development are the ongoing technological revolution and the challenge of sustainability. Technological change is altering patterns of production, consumption and behaviour in societies; at the same time, it is becoming increasingly difficult to ensure the sustainability of these new patterns because of the constraints resulting from the negative externalities generated by economic growth and, in many cases, by technical progress itself.

Reorienting innovation towards reducing or, if possible, reversing the effects of these externalities could create the conditions for synergies between the two processes. Views on the subject vary widely: while some maintain that these synergies can easily be created if growth follows an environmentally friendly model, summarized in the concept of green growth, others argue that production and consumption patterns are changing too slowly and that any technological fix will come too late.

These considerations apply to hard technologies, essentially those used in production. The present document will explore the opportunities being opened up by new ones, basically information and communication technologies, in terms of increasing the effectiveness (outcomes) and efficiency (relative costs) of soft technologies that can improve the way environmental issues are handled in business management and in public policy formulation and implementation.

The main conclusion of this paper is that the use of big data analytics and open data is now indispensable if the exponential growth in data available at very low or zero cost is be taken advantage of. The ability to formulate and implement public policies and projects and management techniques and models conducive to sustainability could suffer if these technological developments are not rapidly exploited.

I. Big data

A. Volume, variety, imprecision, correlation

The concept of big data includes data sets with sizes beyond the ability of commonly used software tools to capture, store, manage and process data.[1] In database terms, they involve a move from relational databases[2] and Structured Query Language (SQL) to programming models based on MapReduce, which have been popularized by Google and are distributed as open-source software through the Hadoop open code platform (see table 1).

Table 1
From relational databases to Hadoop

Relational databases	Hadoop
Multipurpose: useful for analysis and data update, batch and interactive tasks	Designed for large clusters: 1,000 plus computers
High data integrity via ACID (atomicity, consistency, isolation, durability) transactions	Very high availability, keeping long jobs running efficiently even when individual computers break or slow down
Lots of compatible tools, e.g. for data capture, management, reporting, visualization and mining	Data are accessed in "native format" from a file system, with no need to convert them into tables at load time
Support for SQL, the language most widely used for data analysis	No special query language is needed; programmers use familiar languages like Java, Python and Perl
Automatic SQL query optimization, which can radically improve performance	Programmers retain control over performance, rather than counting on a query optimizer
Integration of SQL with familiar programming languages via connectivity protocols, mapping layers and user-defined functions	Open-source Hadoop implementation is funded by corporate donors and will mature over time as Linux and Apache did

Source: Joseph Hellerstein, "The commoditization of massive data analysis", *O'Reilly Radar*, 19 November 2008 [online] http://radar.oreilly.com/2008/11/the-commoditization-of-massive.html.

Big data originated from rapid growth in the volume (velocity and frequency) and diversity of digital data generated in real time as a result of the increasingly important role played by information technologies in everyday activities (digital exhaust).[3] Properly handled, they can be used to generate

[1] What is meant by big data is not measured in absolute terms but in relation to the whole data set involved. It also varies over time as technology advances and by economic sector depending on the technology available. See [online] http://www.technology-digital.com/web20/digital-information-and-how-is-it-expanding.

[2] A relational database enables interconnections (relationships) to be established between data stored in tables.

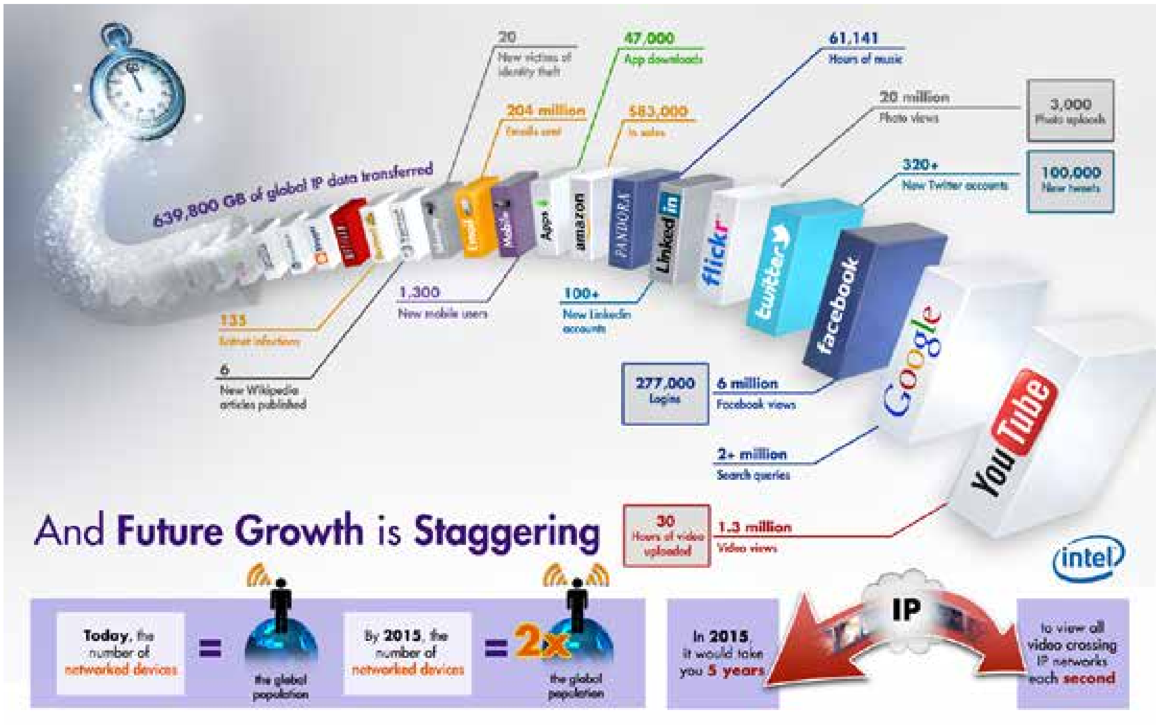
[3] Digital data may be kept in different types of format: (i) structured, where data are organized into transactional systems with a well-defined fixed structure and stored in relational databases, (ii) semistructured, where data have no regular structure or one that can evolve unpredictably, are heterogeneous and may be incomplete, and (iii) unstructured, where data have not been incorporated into regular structures (Lóscio, 2013).

information and knowledge based on full information in real time, meaning a time period short enough for the new information to be used to change decisions before they are irreversible.

Although the exponential growth in the volume of information processed and the rate at which it is generated (see diagram 1) may be obvious, some facts are worth highlighting.

- Google processes about 1 petabyte (10^{15} bytes) an hour.
- A total of 1,200 exabytes (10^{18} bytes) of digital data already existed in 2010, up from some 160 exabytes four years earlier. This 2006 figure was already equivalent to 3 million times the amount of information contained in all the books ever written and more than 30 times as many words as have been uttered in the history of humanity (Mehta, 2012).
- Data are increasingly unstructured, consisting of information contained in pictures, video and text rather than alphanumeric tables. In 2013, videos streamed by Netflix took up over 32% of fixed broadband in the United States, while YouTube accounted for a further 17%. In Europe, over 24% of fixed broadband was accounted for by YouTube and a further 16% by various video sources (Telefónica, 2014).
- Information is generated by a great variety of devices, essentially machines and sensors, of which there were over 4 billion in 2010. This has resulted in a growing share for machine-to-machine (M2M) traffic and given rise to the concept of the Internet of Things, meaning interconnection via the infrastructure of the Internet of unequivocally identified computer-like instruments incorporated into all kinds of devices. This process is the start of the "industrial revolution of data" era.[4]

Diagram 1
What happens in an Internet Minute?



Source: Krystal Temple, *InsideScoop*, Intel, 13 March 2012 [online] http://scoop.intel.com/what-happens-in-an-internet-minute.

[4] "We are at the beginning of what I call The Industrial Revolution of Data. We're not quite there yet, since most of the digital information available today is still individually 'handmade': prose on web pages, data entered into forms, videos and music edited and uploaded to servers. But we are starting to see the rise of automatic data generation 'factories' such as software logs, UPC scanners, RFID, GPS transceivers, video and audio feeds. These automated processes can stamp out data at volumes that will quickly dwarf the collective productivity of content authors worldwide" (Hellerstein, 2008).

The information giving rise to big data is generated by traditional sources, particularly firms and individuals in their daily activities, and is used for a purpose other than the intended one.[5] These sources include, in particular:[6]

- *Web and social media*: this includes Web content and information obtained from social networks such as Facebook, Twitter, LinkedIn and blogs.
- *Transaction data*: this includes billing, credit card and financial operation records. These data are available in both semistructured and unstructured formats.
- *Human-generated*: information kept by a call centre when a telephone call is made, voice notes, electronic mail, electronic documents and medical records.
- *Biometrics*: this includes fingerprints, retinal scans, facial recognition, genetics, etc.
- *Machine-to-machine*: this is when devices such as sensors or meters capture some particular event (for example, speed, temperature, pressure, meteorological variables or chemical variables such as salinity) and transmit it over wireline, wireless or hybrid networks to other applications that turn it into information.

Traditional sources gather data for one or a few specific purposes; with big data, by contrast, data are reused for purposes other than those intended when they were generated. The concept of reuse is thus fundamental (Mayer-Schönberger and Cukier, 2013).

As quantification and storage of the results of human activities (including even relationships, experiences and moods) have become routine and widespread, it has become possible to "datify" all kinds of phenomena, enabling them to be tabulated and analysed quantitatively.

The increased volume of data and their unstructured character go together with a third characteristic, which is the growing importance of correlations between events as a forecasting tool. This is progressively replacing the scientific paradigm whereby the first step towards predicting the outcome of a process is to explain it. The exponential increase in the number of data more than makes up for their messiness, opening up a choice of ways to improve decision-making by overcoming the constraints implicit in efforts to explain phenomena by developing and solving complex systems.

Accordingly, the great shift has been from using small samples of highly refined data to working with data that cover the whole universe in question, even if they are of lower quality.

This approach also tends to ignore the concept of causality (Mayer-Schönberger and Cukier, 2013; Hilbert, 2013). This is particularly important in the field of sustainability, where the complexity of the systems involved (number of variables and feedback) makes it very difficult to identify causal relationships. If the methodological approach taken to big data is the right one, the existence of plausible correlations provides a basis for designing, implementing and evaluating public policies without waiting for consensus on causality.

In summary, as Wang (2012) emphasizes, the fundamental variables of the big data universe are usually considered to be the three shown in diagram 2 (also known as the three V's):

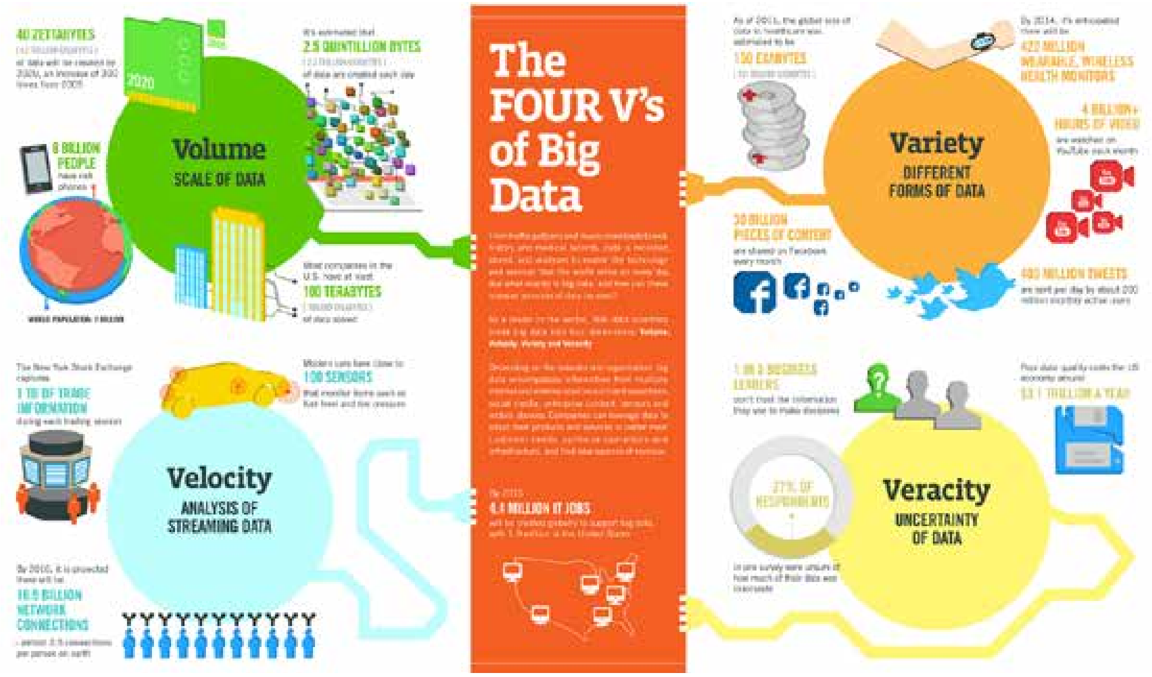
- *Volume*: the quantity of data
- *Velocity*: the speed at which data are generated, captured and shared
- *Variety*: the number and nature of different data types

[5] A piece of data is usually defined as a value without an explicit meaning, whereas by information is meant a meaning associated with or deduced from a set of data and the associations between them (Lóscio, 2013).

[6] IBM, "¿Qué es Big Data?" [online] http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/.

Another two concepts are sometimes added to these, although they are not incontrovertible. First, there is the veracity of data, i.e. their predictive accuracy in the universe under consideration (Hurwitz and others, 2013), although the trade-off between volume and veracity is usually tilted towards the former, as large quantities of data can more than make up for a lack of precision. The second is their value, which implies a reduction in their complexity to make them usable in decision-making.[7]

Diagram 2
The three, four or more V's


The FOUR V's of Big Data
Volume
SCALE OF DATA
40 ZETTABYTES
2020
It's estimated that
2.5 QUINTILLION BYTES
of data are created each day
6 BILLION PEOPLE
have cell phones
100 TERABYTES
of data stored
Velocity
ANALYSIS OF STREAMING DATA
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
18.9 BILLION NETWORK CONNECTIONS
4.4 MILLION IT JOBS
Variety
DIFFERENT FORMS OF DATA
150 EXABYTES
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
Veracity
UNCERTAINTY OF DATA
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
27% OF RESPONDENTS
$3.1 TRILLION A YEAR

Source: IBM, "The Four V's of Big Data" [online] http://expertintegratedsystemsblog.com/wp-content/uploads/2013/12/OSC-Big-Data.bmp.

B. Use in value creation and development

Apart from the characteristics of big data, what matters is how far they contribute to economic, social and environmental goals. This section shows some of the impacts on business efficiency and productivity and on economic development, while the next considers sustainability effects in detail.

These impacts are the result of reality mining, involving the monitoring and measurement of complex social systems. This entails (i) continuous analysis of data flows (streaming) using web scraping instruments, for example, to gather real-time goods prices, (ii) online processing of semistructured and unstructured data such as news or product reviews and (iii) rapid real-time correlation of the data flow with repositories of historical data that can only be accessed slowly (Nathan and Pentland, 2006; Global Pulse, 2012).

At the microeconomic level, having large volumes of data and so being able to go beyond inferences based on small samples enables a business's value added to be increased via different

[7] The list of big data V's has been growing unchecked. Wang (2012) himself adds viscosity (resistance to the flow of data and instruments for reducing this) and virality (how quickly data spread in person-to-person (P2P) networks).

mechanisms (Manyika and others, 2011). In the first place, it means that markets can be segmented more precisely and personalized products and services offered by firms. This greater segmentation allows productivity and profitability to increase, with positive effects on business investment and growth, and will result in the creative destruction of current business models (A.T. Kearney, 2013).

Increased availability of information at ever-lower unit costs, even tending to zero, means that the quality of business products and services can be improved and new goods developed that combine mass production (which cuts costs) with personalization of these goods (which increases consumer welfare).

Big data support for decision-making using intelligent software facilitates the development of new business models, often by making efficient use of long-tail models (Anderson, 2004).[8] These new business models can operate in both government and private services, as with the anticipatory product shipping developed by Amazon.

Lastly, increased transparency and efficiency from data sharing means that positive results can be achieved both at the microeconomic level (better and timelier analysis of organizations' performance and adjustments to actions) and at the production chain level, creating the conditions for integrated management of these both economically (for example, revenues and costs) and environmentally, as will be seen in the next section.

Microeconomic impacts are relevant at the individual or inter-personal levels. An example of the former is a model developed by physicists at Northwestern University which can predict with more than 93% accuracy where someone will be at a given time by analysing mobile phone information generated during their past movements (Hotz, 2011).

To take an example that directly concerns company workforce performance, Eagle, Pentland and Lazer (2009), after analysing 330,000 hours' worth of data on the mobile phone behaviour of 94 people and comparing these with relational data reported directly by these individuals, presented a method of measuring behaviour based on proximity and communication data and identified characteristics that allowed them to predict reciprocal ties of friendship with 95% accuracy.

They used these behaviour signals to predict individual outcomes such as job satisfaction and showed that observations about mobile phone use provided clues not only to observable behaviour but also to other variables such as friendship and individual satisfaction that are strongly related to workplace productivity.

At a more aggregate level, scientists at Johns Hopkins University analysed over 1.6 million health-related tweets (out of a total of over 2 billion) in the United States between May 2009 and October 2010 and found a correlation of 95.8% between the proportion of influenza sufferers estimated from their data and the official rate (Paul and Dredze, 2011).

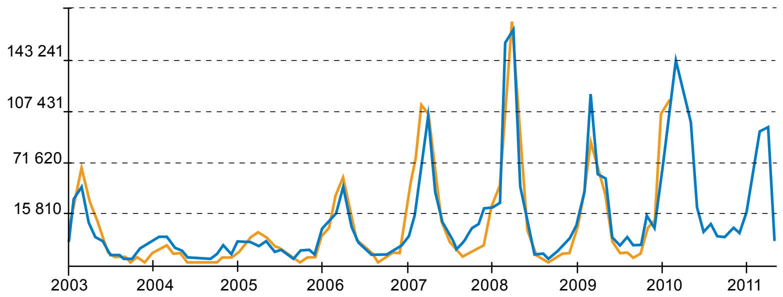
This example and others, such as the accuracy of predictions made on the basis of Google data about dengue activity in Brazil[9] and the now classic case of the estimation of influenza incidence in regions

[8] In a universe with exponentially more small units than large ones (a few big players and a lot of small ones), more value may be generated by small producers than by large ones. Thus, Anderson (2004) writes: "Forget squeezing millions from a few megahits at the top of the charts. The future of entertainment is in the millions of niche markets at the shallow end of the bitstream."

[9] Google discovered that there was a close relationship between the number of people carrying out dengue-related searches and the number of people actually presenting symptoms of the disease. When all searches relating to the disease were added up, a pattern emerged. Comparing query counts with traditional dengue surveillance systems showed that search queries tended to be popular exactly when dengue season was in progress. Counting how often these search queries were seen made it possible to estimate how much dengue was circulating in different countries and regions around the world [online] http://www.google.org/denguetrends/intl/en_gb/about/how.html.

of the United States,[10] show the great predictive capacity of models based on big data that work with correlations between variables without attempting explanations, let alone addressing issues of causality.

Figure 1
Brazil dengue activity, 2003-2011: Google estimates and government data



Source: Google.org, “Dengue Trends” [online] http://www.google.org/denguetrends/intl/es/about/how.html.

Note: The darker line shows Google estimates and the lighter line data reported by the health ministry.

In the macroeconomic dimension, big data use can positively impact development in at least two ways. First, it makes it possible to produce intelligent information that saves resources in the process of generating the statistics that underpin development policies and then serve to implement and evaluate them. Second, by turning imperfect, unstructured and complex data on people’s welfare into processable information, it can reduce time lags between design, execution and evaluation and can narrow knowledge gaps, thus enabling timely policy decisions to be made in response to particular situations and rapid feedback to be obtained on their impact, operating, as noted earlier, in real time. Thus, for example, Helbing and Balietti (2011) have shown that a country’s GDP can be estimated in real time by measuring remotely detected night-time emissions of light.

In poverty reduction policies, measuring the short-term evolution of poverty is essential both for designing mechanisms to respond to any sudden deterioration and for evaluating (measuring) the effectiveness of policy instruments in real time. On one level, Soto and others (2013) show that information from aggregated mobile phone usage records can be used to identify socioeconomic levels, presenting a model that yields accurate predictions for over 80% of an urban population of 500,000. Even more significant are the considerations set out by WEF (2012), showing that a drop in mobile telephone prepayment amounts tends to indicate a loss of income among the population. This information would flag up increases in poverty long before they were reflected in official indicators. In another related field, a study for Indonesia in 2010-2011 finds a very strong positive relationship between Twitter conversations about food (rice) and food price increases (Global Pulse/Crimson Hexagon, 2011). Furthermore, not only can price rises be predicted, but it is possible to monitor in real time how and where people are changing their behaviour and the allocation of their resources. These conclusions are of huge importance for public policy design.[11]

[10] Ginsberg and others (2009) state in their conclusion that “we present a method of analysing large numbers of Google search queries to track influenza-like illness in a population. Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a patient presents with influenza-like symptoms, we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day. This approach may make it possible to use search queries to detect influenza epidemics in areas with a large population of web search users.” The model has lost predictive capacity since then, overestimating the incidence of influenza, which highlights the need for constant updating of correlation-based models (Tempini, 2013).

[11] A wide-ranging review of the use of mobile phones as sensors for social research can be found in Eagle (2011).

C. Big data for sustainability

Handling large volumes of data means that actions can be coordinated and monitored right along value chains, allowing for efficient oversight of products and externalities. In particular, it enables large firms to understand, measure and act upon major environmental impacts they have that are outside their direct control, i.e. those caused by their suppliers of physical inputs and services.

There are important examples of leading firms that have developed big data-based systems to understand the impact of their business right along the value chains they operate in, particularly Nike, Ikea and Hitachi. The last of these has created an online platform that makes it easier for its suppliers to report on their compliance with sustainability criteria, a process that is believed to encourage accountability among small suppliers and that facilitates overall monitoring of the chain.[12]

Besides these cases, British Telecom (BT) seems to have gone furthest, at least judging by the information available online. The firm collaborated with the Carbon Trust to study the total carbon footprint of its business and concluded that 92% of all emissions in its value chains were outside its direct control, with two thirds of all emissions originating in the operations of 17,000 suppliers. It also identified so-called carbon hotspot areas where there were business opportunities to cut costs and carbon emissions (see box 1).

Box 1
Monitoring and operations in the BT value chain to cut carbon emissions

BT with the Carbon Trust

BT is working in partnership with the Carbon Trust to measure the life-cycle carbon emissions of its flagship consumer products, engaging suppliers in carbon reduction and helping reduce the footprint of the communication services sector. BT is one of the world's leading communications services companies, delivering fixed-line, broadband, networked IT, mobile and TV products and services to businesses and consumers in the United Kingdom and more than 170 countries worldwide. From systems that manage energy use in buildings to video conferencing that helps reduce the need for air travel, communication technology can be used to reduce the pressure on resources and cut carbon emissions.

Net Good: sustainability leadership

BT's Net Good programme, launched in June 2013, demonstrates its sustainability leadership and shows the company's pioneering commitment to carbon abatement. Through its Net Good project, BT aims to use its products and people to help society live within the limits of the planet's resources, with the 2020 goal of helping customers reduce carbon emissions by at least three times the end-to-end carbon impact of its own operations. At the same time, BT is targeting an 80% reduction in the carbon intensity of its global business per unit turnover by continuing to work on improving the sustainability of its own operations and extending the influence to its supply chain.

Working in partnership with the Carbon Trust, BT has:

- Measured the full life-cycle carbon emissions of three flagship consumer products: BT Home Hub, BT DECT digital cordless phone and BT Vision set-top box;
- Undertaken an extensive supplier engagement programme involving a number of workshops;
- Developed a climate change procurement standard that applies to all of BT's suppliers, encouraging suppliers to use energy efficiently and reduce carbon during the production, delivery, use and disposal of products and services;
- Collaborated with the Greenhouse Gas (GHG) Protocol ICT Sector Guidance initiative to forge global agreement on a consistent approach to assessing the life-cycle GHG impacts of ICT products and services, and used this approach to understand the full life-cycle footprint of the communication services for the London 2012 Olympic and Paralympic Games.

[12] See Hsu (2014), Carbon Trust (2013) and "BT - Reducing footprint of the communication services sector" [online] http://www.carbontrust.com/our-clients/b/bt.

Box 1 (concluded)

Reducing impacts

BT is one of only a handful of companies globally to have measured and reported its full GHG Protocol Scope 3 emissions against all 15 categories that detail emissions both upstream (e.g. from purchased inputs) and downstream (e.g. distribution to the point of sale). Measuring full life-cycle carbon emissions in three of its flagship consumer products has informed BT and helped the company reduce impacts in successive models. For example, the BT Home Hub 5 has VDSL capability integrated into the unit (avoiding the need for a separate modem), dual wireless technology (improving short-range transmission) and intelligent power management technology including a power save mode when not in use. Altogether these changes are expected to reduce power consumption by 30%, saving 13,000 tons of carbon dioxide a year.

BT has also:

- Created Designing Our Tomorrow, a framework of sustainable design principles to bring the benefits of responsible and sustainable business practice into commercial and customer experience processes;
- Launched the Better Future Supplier Forum to drive energy efficiency and carbon reduction in the supply chain.

Investment in the future

BT's focus on being a responsible and sustainable business leader has led to a 44% reduction in operational emissions and a 15% reduction in supply chain emissions, along with a 40% reduction in waste to landfill since 2011. At the same time, BT has decreased operating costs by 14% and boosted earnings before interest, taxes, depreciation and amortization by 6%, building a strong investment in the company's future and that of the United Kingdom's telecommunications infrastructure (see [online] Opportunities in a resource constrained world: How business is rising to the challenge).

Source: Carbon Trust, "BT - Reducing footprint of the communication services sector" [online] http://www.carbontrust.com/our-clients/b/bt [retrieved 16 July 2014].

Apart from the example of cooperation between BT and the Carbon Trust, big data have been used to reduce negative environmental externalities by two major firms in the information technology sector: Cisco and IBM.

In May 2014, Cisco signed partnership agreements on big data management for environmental purposes with three local governments in Denmark, a country that is a leader in this area. The municipalities of Copenhagen, Albertslund and Frederikssund have signed memoranda of understanding with the firm to develop the digital infrastructure of tomorrow, the Internet of Everything, a concept similar to that of the Internet of Things mentioned earlier.[13]

The strong green profile of these three municipalities was the main reason for Cisco's decision to enter into these partnerships. In particular, Copenhagen has set itself the goal of becoming the world's first CO2-neutral capital by 2025 via the CPH 2025 Climate Plan.[14]

The agreement includes technical solutions to improve service to citizens and meet the environmental targets set for 2025. In particular, in Copenhagen and Albertslund it should be possible to validate and scale the cities' solutions through a single network, including solutions such as outdoor lighting, parking, mobile services, traffic light systems, location-based services, sensor-based cloudburst mitigation, physical infrastructure monitoring and control and intelligent energy technologies (see box 2).

[13] See [online] http://www.stateofgreen.com/en/Newsroom/Cisco-Enters-into-Big-Data-Partnership-with-Three-Danish-Municipalities.

[14] According to the main municipal executive authority, Frank Jensen, "Copenhagen is among the very best when it comes to green solutions and this agreement demonstrates how sustainability can be the direct path to economic growth. Over the next couple of years, the capital area of Denmark will test and develop new solutions including intelligent street lighting, green waves in traffic for bicycles and buses and energy-saving technologies for offices and households. These innovative and green solutions will increase the quality of life for citizens while creating growth and employment."

Box 2
Cisco's partnership with Danish municipalities

Copenhagen

This partnership also means that Cisco invests money, time and expertise in the capital area. The intention is that Cisco and Copenhagen should learn from each other and develop new products and solutions in collaboration with other municipalities, companies and institutions. Cisco has also announced a global fund of 110 million euros to assist start-ups and ecosystems surrounding the "Internet of Everything".

"The Internet of Everything's data revolution is now enacted in Copenhagen, which already is one of Europe's most innovative cities. The city has an ambitious green vision and forms the perfect setting for a green lab. The projects in and around Copenhagen will act as pioneering examples that increase efficiency, cost reductions and sustainability, which other world cities can re-use," said Cisco's Executive Vice-President, Industry Solutions and Chief Globalization Officer Wim Elfrink.

Albertslund

One of the main projects in Albertslund is the Danish Outdoor Lighting Lab (DOLL) Living Lab,[a] a new platform for testing innovative and intelligent lighting solutions: "We have just been in Barcelona to witness Cisco's Internet of Everything projects. There are some very interesting possibilities for development of tomorrow's cities. In Albertslund, we have been preparing for intelligent lighting, and DOLL Living Lab opens up the city space for technologies that push the green transition," said the Mayor of Albertslund, Steen Christensen.

Frederikssund

Frederikssund will primarily use the data network in connection with the establishment of the city of Vinge, which will be the first in Denmark to obtain 100% of its electricity from renewable sources: "We have worked on sustainable initiatives for citizens and business life for a long time. With up to 20,000 citizens, the new city of Vinge will become Denmark's first fully electric city based entirely on renewables. It is natural for us to make Vinge available for implementation of tomorrow's digital infrastructure and smart city solutions."

Source: Cisco, "Cisco Enters into Big Data Partnership with Three Danish Municipalities" [online] http://www.stateofgreen.com/en/Newsroom/Cisco-Enters-into-Big-Data-Partnership-with-Three-Danish-Municipalities.

[a] DOLL Living Lab, located in Hersted Industry Park, Albertslund, offers a 1:1 experience of outdoor lightning.

At the other environmental extreme, big data-based instruments are also being used in Beijing to reduce the impact of negative growth externalities. The municipality of the Chinese capital announced on 15 July 2014 that it had reached a 10-year agreement with IBM, called Green Horizon, to use its advanced weather forecasting and cloud computing technologies to solve the problem of smog. Since handling pollution and fog problems requires improvements to data collection and monitoring and prediction capabilities, Beijing reported that it had set up an early warning system, using data from 35 monitoring stations, which could flag up acute pollution episodes three days in advance, enabling traffic volumes to be adjusted in time.[15]

At the same time, since China has sought to reduce the proportion of coal-fired power generation, the IBM cloud computing analysis system will optimize and adjust renewable energy source goals. In particular, the demonstration project of the State energy network in Zhangbei, Hebei, shows that the IBM system of supply and demand management can cut energy wastage rates by between 20% and 30%. The project is also expected to create business opportunities in the sectors of renewable energy and pollution control.

Lastly, in another shift crucial to sustainability, this time in the management of natural resources, Google announced in June 2014 that it was planning to use Skybox satellites in the near future to update Google Maps and increase their accuracy. According to *Wired*, this will make it possible to estimate the dynamic of natural resource reserves in real time.[16] One example is the measurement of Saudi oil reserves from space using Skybox Imaging. Because oil is typically stored in tanks with floating lids to avoid losses from evaporation in the gap between the top of the oil and the lid, Skybox satellite images

[15] *China Tech News.com* [online] http://www.chinatechnews.com/2014/07/15/20725-big-data-used-by-beijing-government-to-alleviate-pollution.

[16] *Wired*, June 2014 [online] http://www.wired.com/2014/06/googles-new-satellites-have-amazing-potential-and-potential-for-abuse.

have been used to estimate the volume and level of oil in each tank by observing the movement of the lid. Increasing integration between Skybox and Mapbox will simplify access to and analysis of images of this type (Sánchez-Andrade and Berkenstock, 2014).

In summary, major examples in recent years involving firms such as BT, Cisco, IBM and Google show that big data is now being used as a sustainability instrument in real-life situations both to optimize environmental outcomes in advanced countries and to prevent extreme situations from arising.

D. Analytics

Just as the characterization of big data is based on the volume-imprecision-correlation triad, their analysis is summarized in another: data-proxies-visualization. Big data analytics encompasses all the tools and methodologies used to turn masses of raw data into metadata for analytical purposes.[17] Its origin lies in disciplines such as computing-intensive biology, biomedical engineering, medicine and electronics.

At the heart of this analytics are algorithms for detecting data patterns, trends and correlations over various time horizons, along with the ability to generate significant proxies, for example by linking the content of tweets about food prices to increases in the price of rice in a specific community, as shown earlier. Two elements are important in this construct: first, creativity for identifying new relationships to be explored and, second, the high contextual content of this identification. Both generate opportunities for local value creation on the basis of metadata analysis.

The essential data analysis instruments are visualization (histograms, scatter or surface plots, tree maps, parallel coordinate plots, etc.), statistics (hypothesis tests, regressions, principal component analysis, etc.), data and association mining and machine learning methods such as clustering and decision trees.

Advanced visualization techniques (visual data analysis) involve the selection of a visual representation of abstract data to enhance knowledge; these data may be numeric or non-numeric, such as text or geographical information. The results of the correlations identified may sometimes present in spaces of more than two dimensions, which makes them much harder for non-specialists to understand. Accordingly, an important part of big data analysis is the development of visualization instruments that make sense to non-technical policymakers.

Furthermore, information visualization can also work as a system for generating hypotheses to be continued by more formal analyses, such as hypothesis tests.

In a recent document entitled “Big Data: New Tricks for Econometrics”, Varian (2014), now chief economist at Google, offers important conclusions about the need for new economists (postgraduate students) to incorporate big data tools into their toolbox. On the basis of a highly technical analysis, he argues that there are analytical elements that only arise in the field of big data and that require tools different to and more powerful than those econometrics can provide.

The document highlights two points. The first is the need for progress with the selection of variables, as the volume of data available means that the research process may have more potential predictors than are needed for an estimate. The second is the need to go beyond the linear model by incorporating tools such as decision trees, neural networks and complex relationship modelling.

On the basis of these considerations, Varian concludes that there is a vital need for cooperation between areas working with causal inferences (econometrics) and those working with correlation-based prediction (machine learning). This cooperation is needed not just in the development of postgraduate studies curricula but also and particularly in the world of economic research.

[17] Metadata are data about data; they aid understanding of the relationships between data and increase their usefulness.

Big data can supplement data generated from traditional sources and thus drive the modernization of statistical systems. Two important examples from Brazil and Colombia are presented in United Nations (2014). In the former, an agreement was signed in 2012 between the Brazilian Geographical and Statistical Institute (IBGE), the National Water Agency (ANA) and the Ministry of the Environment to set up a committee with a mandate to develop water accounts using the large volume of data collected and processed by ANA each day, which are accessible online. In the latter, satellite images are being used as a source of big data to supplement the interviews forming part of the national agricultural census and to measure and monitor coca crops. Use is also being made in the country of data from electronic road toll gantries to improve traffic flows and as an input for transport statistics. Outside the Latin American and Caribbean region, examples can be found in Australia (agricultural statistics), Bhutan (consumer price index), Estonia (international travel statistics based on mobile location data) and the Netherlands (social media as a source for official statistics).

II. Open data

A. The characteristics of open data

The second tool covered in this paper is open data for sustainability. Open data are defined as data which can be reused and distributed without restrictions of any kind, in particular those of an administrative or technical nature. Open data are thus part of the universe of large data by virtue of their size and the fact that they entail reuse of information collected for other purposes. Implicit in the concept of open data is a distinctive characteristic: they are assumed to be held by governments. Meeting in California in December 2007, the Open Government Working Group laid down eight principles that characterize open government data:[18]

- They are complete
- They are primary and unprocessed
- They are timely and released as soon as possible
- They are usable for any purpose
- They are machine-processable
- Access is non-discriminatory, open and anonymous
- They are in a non-proprietary format
- They are free of licences, patents, trademarks, etc.

Several of these characteristics are obvious from the name “open data”, but some warrant special attention, particularly the fact that these have to be primary, readable and machine-processable data. This means that, as explained below, they can be reused by developing applications based on a combination of databases (mash-ups). It is crucial to distinguish here between government transparency and open data. The publication of government information in formats such as PDF may be an exercise in transparency, but does not meet the conditions to be considered an exercise in open data. The potential for interoperability and data linking entails more than transparency alone.

[18] See [online] http://opengovdata.org/ and [online] https://public.resource.org/8_principles.html.

The three actors involved in opening up and employing this type of data are the government, which opens and publishes the data in the formats necessary for reuse, developers of applications, and users, who benefit from increased transparency in public administration, something that can facilitate their participation in this process and help empower them.

The movement towards opening up government data and developing methodologies for this has been a result of the work done by Tim Berners-Lee through the World Wide Web Consortium (W3C), an international community that has been developing open standards since 1994 with a view to ensuring the long-term growth of the Internet.[19]

Berners-Lee and W3C have developed a system to grade the extent to which databases have been configured to qualify as open. This system yields a ranking of from zero to five stars,[20] associated with file formats as detailed in table 2:

- *No star*: data are not accessible under an open licence, even if they are online.
- *One star*: data are accessible on the Web and are readable by the human eye but not by a software agent, since they are in a closed format and cannot be easily reused.
- *Two stars*: data are accessible on the Web in a structured, machine-readable format. They can be processed, exported and published easily, but reusers must still rely on proprietary software such as Microsoft Word or Excel.
- *Three stars*: reuse of data is not dependent on proprietary software, as the comma-separated value (CSV) format is used instead of Excel.
- *Four stars*: data are in the Web rather than on the Web; the use of a uniform resource identifier (URI) gives fine-granular control over the data, by means of bookmarking and linking for example.
- *Five stars*: data are not only in the Web but are also linked to other data, taking full advantage of network effects. This interlinking places data in context and means they can be discovered by other sources, such as via links to Wikipedia.

Table 2
Open data file formats

One star	Portable document format (PDF)
Two stars	Microsoft Excel
Three stars	Comma-separated value (CSV)
Four stars	RDFa (Resource Description Framework in attributes), which adds a set of attribute-level extensions to HTML, XHTML and various XML-based document types for embedding rich metadata within Web documents. It enables these to be used by embedding RDF subject-predicate-object expressions within XHTML documents. It must have at least three URIs.
Five stars	RDFa with several URIs and semantic properties allowing for linked data reuse.

Source: Epsi Platform, “The Five Stars of Open Data” [online] http://www.epsiplatform.eu/sites/default/files/The%20 5%20stars%20of%20Open%20Data_MdV_PR2.pdf.

[19] The main boost to the movement for open data is considered to have been the presidential memorandum issued on the first day of the Obama administration (21 January 2009) calling on the heads of federal agencies to open up as much data as they could: “In the face of doubt, as much data as possible” (Mayer-Schönberger and Cukier, 2013). The Open Government Partnership, an international platform to promote more open government, was launched in 2011. Sixty-four governments are taking part, with Brazil, one of its founders, playing a prominent role (see [online] http://www.opengovpartnership.org).

[20] For a more detailed presentation, see [online] http://5stardata.info/, http://opendatamanual.org/ and W3C Brasil/CGI.br/ NIC.br (2011a and b).

B. Open data for sustainability

Open data enable the developer community to produce mash-up applications, i.e. applications making use of content from more than one source to create a new service that is displayed on a single graphical interface.

The number of mash-ups on the Internet is growing rapidly, with the most popular being map-based (36% of those detected by ProgrammableWeb in 2014). There are two reasons for this growth. First, major companies such as Yahoo!, Google and Amazon have opened up their data to make them usable by other data sources without the need for lengthy licensing negotiations. Second, new tools have been developed that make it easy to create mash-ups even for users with little in the way of technical expertise.

A search made on ProgrammableWeb for this paper on 24 April 2014 found the following numbers of mash-ups in the categories listed: (i) environment: 80 mash-ups, 90 application programming interfaces (APIs)[21] and 2 source codes; (ii) pollution: 6 mash-ups and 2 APIs; (iii) carbon footprint: 1 mash-up and 2 APIs; (iv) carbon emissions: 3 mash-ups and 1 API.

Table 3 identifies some of the main mash-ups found that directly concern environmental issues.

Table 3
Examples of environmental mash-ups found on ProgrammableWeb, 24 April 2014

Map of United States Toxic Air Emissions: displays data from the Environmental Protection Agency (EPA) Toxic Release Inventory database on a simple map that is searchable by city or address.
United States Green Power Map: combines the EPA Emissions and Generation Resource Integrated Database (EGRID) 2006 with Google Maps to provide a search interface for displaying power plants and their associated emissions and generation information, state by state.
United Kingdom Carbon Footprint Project: a platform for a carbon calculator gadget, built in a partnership between Google, the Energy Saving Trust and the Royal Society for the encouragement of Arts, Manufactures and Commerce. It uses the AMEE engine and draws on the United Kingdom carbon calculator for its data, calculations and approved methodologies.
Carbon Emissions Commute: estimates carbon emissions and fuel costs based on the driving itineraries provided by users, who enter details of their daily commute or any other trip. It also provides users with various ideas on how to reduce their emissions.
Greenspace United Kingdom Carbon Emissions Map: provides a map of carbon emissions in the United Kingdom down to the square kilometre level, together with the source of emissions. It also includes a simple emissions reduction to-do list for individuals and communities.
Air Travel Emissions Calculator: a Google Maps mash-up that allows users to calculate the per-passenger greenhouse gas emissions created by a commercial airline flight between any two airports.

Source: ProgrammableWeb [online] http://www.programmableweb.com/.

To provide greater detail on georeferenced applications using open data for sustainability, examples from Mexico, Bolivia, Uruguay and Brazil will now be given.

In Mexico, an application called Infoaqua was developed in 2012 to reuse the databases of Mexico's National Institute of Statistics and Geography (INEGI) and National Water Commission (CONAGUA) and government maps[22] with a view to addressing a lack of awareness and concern about water overconsumption and pollution and more general water-related issues such as droughts, scarcity, inadequate treatment and overexploitation. An interactive solution was chosen to show society the state of the country's water infrastructure and surface water in the form of dynamic real-time infographics and educational games to illustrate water issues around the world.

Recognizing that lack of awareness among individuals and institutions is a cause of environmental damage, some of it reversible but a great deal of it not, in 2012 Econsciencia developed an application called The Source in Bolivia to provide information on activities that are good or bad for the environment.[23]

[21] Please see the next section for the definition of an API.

[22] See [online] http://labs.ratio.mx/infoaqua/. For the databases used, see [online] http://www.inegi.org.mx, [online] http://www.cna.gob.mx and [online] http://mapas.inecc.gob.mx/.

[23] See [online] http://2012.desarrollandoamerica.org/portfolio/econsciencia/. Use was made of the GeoBolivia database [online] http://www.geobolivia.abc.gob.bo and the World Database on Protected Areas (WDPA) [online] http://www.wdpa.org. A short video showing the presentation of The Source at the collaborative hackathon "Desarrollando América Latina" ("Developing Latin America") 2012 can be viewed at [online] http://www.youtube.com/watch?v=MZjRnHxZyCQ.

The idea was to use a map of the country showing rivers, lakes and protected areas to identify pollution hotspots at specific points along a river, show places where a reforestation initiative is being carried out in a protected area and create mechanisms for reporting spots where refuse or other waste is dumped. Different types of institution will be able to use this tool to carry out environmental activities and publish details on the project website, with the information being available to all users, registered or otherwise.

In an urban context, a firm called Data has developed "¿Dónde reciclo?" ("Where do I recycle?"), a mobile device-compatible application that facilitates the sorting and recycling of household waste in Montevideo by providing the location of containers for recycling glass, plastic, batteries and metal anywhere in the city. The application ascertains the user's location from their mobile phone GPS or their computer's geolocation information and provides a map showing their position and the nearest recycling points using open data from the city council. Although these data could be obtained using a tool provided by the city council itself, the developers decided to create a new one because they believed they could come up with a quicker and more efficient way of obtaining and displaying the information by making it accessible from mobile devices (see [online] http://datauy.org/proyecto/donde-reciclo/).

A fourth example is Linked Brazilian Amazon Rainforest Data, a database linked to the linkedscience.org site, which hosts and explores this dataset of deforestation statistics.[24] The application uses the variables of deforestation, the price of pasture and geographical coordinates to draw a grid of 8,441 cells measuring 25 square kilometres each, for which the information is presented. The data used are produced jointly by the Institute for Geoinformatics of the University of Münster in Germany and the Institute for Spatial Research (INPE) in Brazil, on the basis of information from INPE itself, the Brazilian Ministry of the Environment, Informa Economics FNP and the Brazilian Institute of Geography and Statistics (IBGE).

These four examples have certain characteristics in common: the use of information referenced by time and place, display techniques, the combining and reuse of databases produced by various types of institutions, and open access for all types of users. To sum up, the following information is needed to analyse processes and operations in systems as complex as the environment: strongly interconnected data on the system, statistical computing techniques and other types of reasoning to find new information, and methods for its display.[25]

[24] See [online] http://www.semantic-web-journal.net/content/linked-brazilian-amazon-rainforest-data.

[25] See [online] http://linkedscience.org/tools/sparql-package-for-r/tutorial-on-sparql-package-for-r/.

III. Tools, problems and proposals

A. Tools: APIs, web crawlers and metasearch engines

The management and efficient use of big data require new tools for data capture and transmission, in addition to those involved in data analysis. Three kinds of tools deserve special attention: the application programming interfaces (APIs) offered by large companies such as Amazon, web crawlers, and metasearch engines.

An API specifies how certain software components should interact. It enables a program to accomplish a specific task by automatically gathering or modifying data. In other words, it acts as an intermediary between data and the systems that will exploit them.

The supply of APIs is growing, with ProgrammableWeb reporting that the number rose from 1 in June 2005 to 10,302 in October 2013, this latter figure in turn being more than double that for January the year before. The categories showing the most growth are those associated with social networks, firms and the financial sector, followed by maps, e-commerce and government.

Notable APIs include those offered by Amazon, Twitter and Google. Amazon Mobile App SDK provides a variety of APIs, tools and resources so that developers (usually small-scale) can access Amazon Appstore resources to sell their applications and games and build user loyalty. Its Product Advertising API enables small companies to access the Amazon catalogue and include or sell their products via websites and third-party applications.[26]

Twitter offers APIs for almost every feature on its website. Programmers can use them to develop applications, websites, widgets and other projects that communicate with Twitter via Hypertext Transfer Protocol (HTTP). In the same field, Google offers programmers a range of APIs enabling them to create applications that read and write data from a large number of services, such as Google Apps, Google Analytics, Blogger, Google Base, Google Book Search, Google Calendar, Google Code Search, Google Earth, Google Spreadsheets, Google Notebook and Picasa Web Albums.

[26] Other Amazon APIs include the In-App Purchasing, Mobile Ads, Mobile Associates and GameCircle APIs [online] https://developer.amazon.com/.

The importance of these services' databases for the analysis and formulation of sustainability policies means there is a need to develop human resources capable of capitalizing on access to these kinds of data.

For years now, more sophisticated tools than mass market search engines such as Google, Bing and Yahoo! have been required to access information on the Web efficiently. M2M communication using web crawlers (also known as web spiders, ants, automatic indexers or web scutters) has become the norm here. These are Internet bots[27] that systematically trawl the World Wide Web, usually for indexing purposes.

The major search engines and similar sites use web crawlers to update their own content or to index content from other sites. This kind of software is able to copy every page visited for later processing by search engines, which index the pages downloaded so that users can find them much more quickly.

Automatic extraction of World Wide Web data simulates the human browsing process, using software to, for example, compare prices, identify contacts, monitor weather data, detect changes to web pages, conduct research, mash up content and integrate data.[28]

Metasearch engines are one kind of web crawler. They consist of systems that locate information in the most commonly used search engines, acting in fact as search engines within search engines. Metasearch engines have no databases of their own, instead using those of other search engines and displaying a combination of the best pages returned by each. They usually present search results from engines such as About, Ask.com, FindWhat, Google, LookSmart, MSN Search, Teoma, Yahoo! and Bing, including their audio and video content.[29]

Lastly, a tool widely used in the field of open data is SPARQL (SPARQL Protocol and RDF Query Language), a standardized query language for RDF graphs. It is a key technology in the development of the Semantic Web, enabling information available on the Web to be accessed through platforms such as Sig.ma, Sindice.com and DBpedia.org, this last being used to extract structured data from Wikipedia.[30]

Those responsible for carrying out research and proposing options for action in areas where data are largely held online will need to make greater use of these tools. A major ECLAC initiative to this end was the training in Mozenda (*Mo*re *Zen*ful *Da*ta) data extraction software (see [online] www.mozenda.com) that it provided in June 2014. As a provider of software as a service (SaaS), Mozenda offers companies and individuals a fast, easy and low-cost way for non-programmers to extract information from the Web. Along the lines discussed earlier, the firm also offers its Mozenda Web Services Rest API, which enables users to connect to their account in a fully automated process to perform operations such as setting search parameters and gathering and updating results.[31] Another important activity was the first course on open data for public bodies in Latin America and the Caribbean, held at ECLAC headquarters in June 2013 as part of the ECLAC-International Development Research Centre (IDRC)-W3C Brazil project on Open Data for Development in Latin America and the Caribbean (see [online] www.od4d.org).

[27] An Internet bot, also known as a web robot, a WWW robot or simply a bot, is a software application that runs automated tasks over the Internet. Typically, bots perform simple and repetitive tasks, but at a much faster rate than a human being. The main use of bots is in web spidering, where an automated script fetches, analyses and files information from web servers.

[28] Web scraping (web harvesting or web data extraction) is a technique for extracting information from websites. Such software generally simulates human browsing either by implementing low-level HTTP or by embedding a web browser.

[29] Examples of metasearch engines include (i) Blucora's InfoSpace, which offers metasearch and private-label Internet search services (solutions for web publishers), and whose main metasearch site is dogpile.com, and (ii) Ixquick (startpage.com), which gives particular importance to the first 10 results returned by multiple search engines and uses a system of stars to rank the results. One of its characteristics is respect for the privacy of users, as it does not record their IP addresses.

[30] The Semantic Web is a web of data linked in such a way that it can be processed by machines all over the world. It is based on data associated with metadata, common metadata vocabularies and resources with unique identifiers. It is based on three standards: the RDF data model, the SPARQL query language and the Web Ontology Language (OWL) (Lóscio, 2013).

[31] A large base of individual programmers, often in India, has grown up around solutions such as Mozenda and Kaggle (a company that organizes predictive modelling (data mining) competitions open to all comers, [online] www.kaggle.com), and they offer their services online to meet the search needs of users who do not wish to get involved in software writing.

B. Problems: asymmetries, lack of privacy, apophenia and quality

While it holds out the promise of access to all information available on the Internet at low or zero cost and in real time, big data use can also exacerbate certain problems.

First, both data availability itself and data access are highly asymmetrical. Although the big social networks and the main search engines open up their data via APIs or content on their pages, they also decide what to open and by how much, as well as the monetary or information cost of user access. Governments are also opening up their data, but slowly, while no great progress has been seen in the private sector, aside from some transparency efforts (and it has already been seen that transparency is not the same as data openness). As long as social actors pursue such different open data policies, any positive externalities will be some time in coming.[32]

What is more, highly disparate data processing and analysis capabilities may lead to even greater asymmetry. Although advanced analytical tools will probably become more affordable over time, the capacity gap between companies of different sizes, and between individuals and companies, suggests that very few agents will be able to exploit the exponential growth in the stock of structured and unstructured data to the full. In short, as boyd and Crawford (2012) put it: “Limited access to Big Data creates new digital divides…between the Big Data rich and the Big Data poor.”

In business, the accumulation of big data may become a major barrier to the entry of new competitors, in addition to the well-known effects of scale and network economies and product and brand differentiation. If a new company sought to compete with Google or Amazon, it would need not only the financial resources to do so but also the stock of accumulated data that is their true competitive advantage.

Second, as the use of big data and open data applications becomes widespread, which may take longer than expected for individuals and small firms,[33] the negative impact on privacy will certainly increase. Experience has shown that when data are collected and stored, they are eventually reused for purposes undreamt of by the person or company that generated them, or even by some of the big companies that manage them. The anonymization techniques used for small samples are of limited use here, as the innumerable links to and from each piece of data will make its origin easily identifiable by someone with the right technology.

Another potential risk in big data analysis is apophenia, or the identification of non-existent patterns (Global Pulse, 2012). Great masses of data have links in all directions, increasing the likelihood that Type I errors (false positives) will occur.

The risk involved in working increasingly by way of correlations alone is rising because many new data sources reflect only perceptions, intentions and desires. Although these may be precisely what interests researchers, there is a greater risk of false positives to the extent that they are not separated from data about realities. In the end, the issue comes down to the dichotomy already mentioned between the quality and the quantity of data, which is difficult to resolve and will pervade the debate between big data optimists and pessimists.

While these considerations must be given due weight, the generation of data from digital exhaust will continue to grow despite qualms about lack of privacy, and large firms, research centres and government agencies with the ability to capture and analyse these data will continue to use this powerful instrument of social knowledge and oversight.

[32] Concepts such as corporate data philanthropy, proposed by WEF (2012), would be unlikely to reduce this asymmetry significantly, at least somewhere like Latin America.

[33] One problem for the community pressing for government data openness is that only a small number of applications are of professional quality and really in demand from users, even when they are available at very low or zero cost.

C. Proposals for action

For big data to be used effectively for empirical research and in the formulation, implementation and evaluation of policy proposals by an organization such as ECLAC, the following areas of action are proposed.

First, at the technical level, more needs to be known about the availability of environmental data in Latin America. This will require detailed research into the content of open data catalogues with a sustainability dimension and the mash-ups and APIs available. To complement this, it is vital to have more knowledge about API use and to make web crawlers and metasearch engines routine tools in research, particularly where this suffers from a scarcity of structured data produced by government agencies (production chains, SMEs, etc.). With regard to economic analysis, it would be extremely useful to discuss and draw out the implications of the points made by Varian (2013) about the effects of big data on the practice of econometrics.

Second, in the political arena, in-depth research should be carried out into the level of social demand for an approach involving big data, open data and sustainability. Whether the demand is clearly large or only incipient, it would be useful to build partnerships with civil society organizations pressing for government data to be opened up, such as the World Wide Web Consortium (W3C).

Lastly, from a longer-term perspective, progress is needed in two areas. The first is knowledge of trends towards the Internet of Things and the accompanying generation of data from ubiquitous environmental sensors that will result. Second, there needs to be ongoing formal analysis of complex ecosystem models designed to increase knowledge of the biosphere, such as the Madingley model, and an effort to understand the policy implications of operating with these.[34]

The strategy should centre not only on strengthening analytical capabilities, but on developing a big data mindset in the economic, social and environmental domains of sustainability.

[34] See Purves and others (2013). The Madingley General Ecosystem Model (GEM) was developed by Microsoft Research and the United Nations Environment Programme World Conservation Monitoring Centre (UNEP-WCMC) [online] http://www.madingleymodel.org/about.html.

Bibliography

Anderson, Chris (2004), "The long tail", *Wired*, No. 12.10, October.

A.T. Kearney (2013), "Big Data and the Creative Destruction of Today's Business Models" [online] http://www.atkearney.com/strategic-it/ideas-insights/article/-/asset_publisher/LCcgOeS4t85g/content/big-data-and-the-creative-destruction-of-today-s-business-models/10192.

boyd, Danah and Kate Crawford (2012), "Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon", *Information, Communication & Society*, vol. 15, No. 5.

Carbon Trust (2013), "Big Business, Big Data, Big Sustainability" [online]http://www.carbontrust.com/news/2013/10/big-business-big-data-big-sustainability.

Eagle, Nathan, (2011), "Mobile phones as sensors for social research", *The Handbook of Emergent Technologies in Social Research*, Sharlene Nagy Hesse-Biber (ed.), Oxford University Press.

Eagle, Nathan and Alex Pentland (2006), "Reality mining: sensing complex social systems", *Personal and Ubiquitous Computing*, vol.10, No. 4, Massachusetts Institute of Technology (MIT).

Eagle, Nathan, Alex Pentland and David Lazer (2009), "Inferring social network structure using mobile phone data", *Proceedings of the National Academy of Science (PNAS),* vol. 106, No. 36, 8 September.

Ginsberg, Jeremy and others (2009), "Detecting influenza epidemics using search engine query data", *Nature*, vol. 457, 19 February.

Global Pulse (2012), *Big Data for Development: Challenges & Opportunities*, New York, United Nations [online] http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf.

Global Pulse/Crimson Hexagon (2011), "Twitter and Perception of Crisis Related Stress. A Methodological White Paper" [online] http://www.unglobalpulse.org/projects/twitter-and-perceptions-crisis-related-stress.

Hagen, Christian and others (2013), *Big Data and the Creative Destruction of Today's Business Models*, ATKearney.

Helbing, Dirk and Stefano Balietti (2011), "From social data mining to forecasting socio-economic crises", *European Physical Journal. Special Topics*, vol. 195, No. 1, Swiss Federal Institute of Technology Zurich (ETH), May.

Hellerstein, Joe (2008), "The commoditization of massive data analysis", *O'Reilly Radar*, November [online] http://radar.oreilly.com/2008/11/the-commoditization-of-massive.html.

Hilbert, Martin (2013), *Big Data for Development: From Information- to Knowledge Societies*, Economic Commission for Latin America and the Caribbean (ECLAC)/University of Southern California [online] http://www.geogecko.com/wp-content/uploads/2013/03/Big-data-for-development.pdf.

Hotz, Robert Lee (2011), "The really smart phone", *The Wall Street Journal*, 22 April.

Hsu, John (2014), "Why big data will have an impact on sustainability", *The Guardian*, 31 January.

Hurwitz, Judith and others (2013), *From Big Data for Dummies*, Wiley.

Kolb, Jason and Jeremy Kolb (2013), *The Big Data Revolution*, CreateSpace Independent Publishing Platform.

Lóscio, Bernadette (2013), *Integración de datos* y *Web Semántica* [online] www.od4d.org.

Manyika, James and others (2011), Big Data: The Next Frontier for Innovation, Competition and Productivity, McKinsey.

Mayer-Schönberger, Viktor and K. Cukier (2013), *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Houghton Mifflin Harcourt.

Mehta, Abhishek (2012), *Big Data: Powering the Next Industrial Revolution*, Seattle, Tableau Software.

Paul, Michael J. and Mark Dredze (2011), "You are what you tweet: analyzing twitter for public health¨, *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* [online] https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2880/3264.

Purves, Drew and others (2013), "Time to model all life on Earth", *Nature*, vol. 493, 17 January.

Sánchez-Andrade Nuño, Bruno and Dan Berkenstock (2014), "Monitoring Oil Reserves from Space", Skybox Imaging, 8 April [online] http://skybox.com/blog/monitoring-oil-reserves-from-space.

Soto, V. and others (2013), "Prediction of Socioeconomic Levels using Cell Phone Records", Telefónica Research, Madrid [online] http://www.vanessafriasmartinez.org/uploads/umap2011.pdf.

Telefónica (2014), *A Digital Manifesto. An Open and Safe Internet Experience for All* [online] www.digitalmanifesto.telefonica.com.

Tempini, Niccolò (2013), "Elephant in the glass shop", 5 May [online] http://bit.ly/bigdatareview.

United Nations (2014), *Big data and modernization of statistical systems*. Report of the Secretary General (E/CN.3/2-14/11), New York, Statistical Commission.

Varian, Hal R. (2013), "Big Data: New Tricks for Econometrics", Berkeley, University of California [online] http://people.ischool.berkeley.edu/~hal/Papers/2013/ml.pdf.

Wang, R. "Ray" (2012), *Monday's Musings: Beyond the Three V's of Big Data - Viscosity and Virality* [online] http://blog.softwareinsider.org/2012/02/27/mondays-musings-beyond-the-three-vs-of-big-data-viscosity-and-virality/.

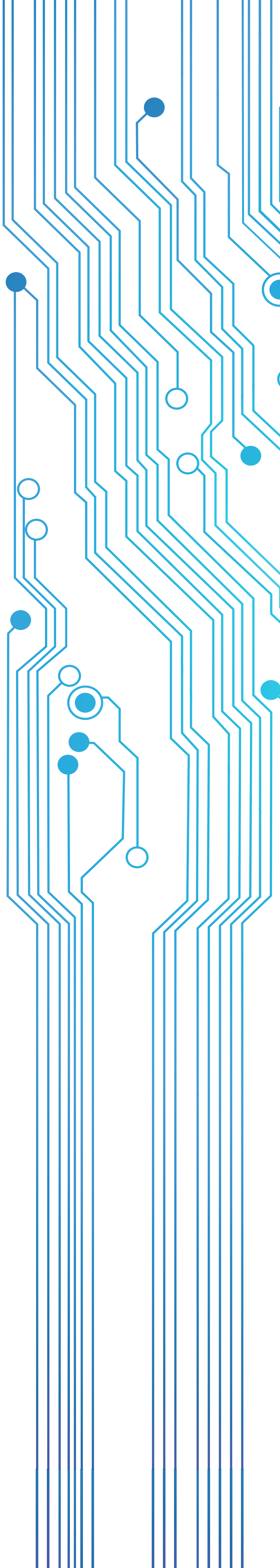
Washington, Anne (2014), *Big Data and Public Sector Information*, School of Public Policy, George Mason University.

WEF (World Economic Forum) (2012), *Big Data, Big Impact: New Possibilities for International Development*.

Wohlsen, Marcus (2014), "Google's New all-seeing satellites have huge potential—for good and evil", *Wired*, June [online] http://www.wired.com/2014/06/googles-new-satellites-have-amazing-potential-and-potential-for-abuse.

W3C Brasil/CGI.br/NIC.br (World Wide Web Consortium/Brazilian Internet Steering Committee/Brazilian Network Information Center) (2011a), *Manual dos dados abertos: desenvolvedores*, São Paulo.

_______(2011b), *Manual dos dados abertos: governo*, São Paulo.



ECLAC

Economic Commission for Latin America and the Caribbean (ECLAC)
Comisión Económica para América Latina y el Caribe (CEPAL)
www.eclac.org