ISTAT INTERNATIONAL AND NATIONAL INITIATIVES ON BIG DATA

Paolo Righi

1. Introduction

The National Statistical Institutes (NSIs) have recently started investigating Big Data (BD) as potential data sources for generating official statistics. Especially beginning in 2013, numerous international projects and initiatives have been undertaken with the main purposes of defining strategies for adopting the BD sources in the production process. These strategies have to deal with some issues as legislation (privacy, confidentiality, etc.), communication (build public trust in the use of private sector BD for official statistics), partnership (especially with data providers) and financial aspects, IT infrastructure (to process huge amount of data), new or unusual (for the NSIs) quality and methodological framework, training NSI staff on new statistical tools and the definition of the governance of the BD inside the NSI. Several working group and task force have been set up for promoting the practical use of BD sources. The objective is to find the solutions for these challenges and support the capacity building, training and sharing of experience.

The Italian National Statistical Institute (ISTAT) attends to some international activities.

In particular at European level ISTAT is involved in the Task Force on Big Data and Official Statistics chaired by Eurostat. The Task Force implemented the road map and the action plan of the Scheveningen Memorandum on Big Data and Official Statistics (September 2013). This Memorandum aims to define a global strategy for introducing BD sources in the statistical production process and it has been the first step to provide a European strategic vision of the BD phenomenon.

Another initiative, where ISTAT participated, is the BD project of the United Nations Economic Commission for Europe (UNECE) launched in March 2014. The UNECE project ended in 2014 but the worldwide initiative of Global Working Group (GWG) on Big Data for Official Statistics coordinated by United Nations Statistics Division (UNSD) aims to continue some of the works already done by the UNECE project. The working group has been launched in 2015 and should work for the next two or three years.

In the GWG, ISTAT is leader of Task Team on Cross-cutting issues, Classifications, Frameworks and Taxonomy.

At National level, ISTAT closed (March 2015) a two year Technical Commission devoted to define a Roadmap for the adoption of BD in production process. Several actors coming from Academia, IT and Media companies have been involved and three pilots using BD sources have been carried out.

In the 2014, ISTAT has begun an internal project on the use of scanner data for the Harmonized Index of Consumer Prices compilation (involved in the European project "Multipurpose price statistics").

Other separated agreements with Mobile phone companies, University and IT companies have been established or are in progress.

The paper recaps the targets and the achieved outputs produced by these activities. Section 2 introduces the phenomenon of BD, with the definition of useful concepts for the official statistics. Section 3 focuses on the reasons of considering BD in the NSI. Their uses raise some issues shown in Section 4. Section 5 describes the national and international projects, task force and working group, where these issues are tackled.

2. Big Data: definition and facts

NSIs have been using several types of data sources in the production process of official statistics, including designed data sources such as censuses and survey sampling, and found data sources such as administrative and transactional data.

Recently as a result of more and more interaction with digital technologies by citizens, and the increasing capability of these technologies to provide digital trails, new sources of data have emerged and are increasingly available (IDC, 2014; Khan *et al.*, 2014; Chen *et al.*, 2014).

Some findings highlight what we are talking about:

- every two days we create as much information as we did from the beginning of time until 2003;
- in 2012 over 90% of all the data in the world was created in the past two years;
- it is expected that by 2020 the amount of digital information we have in existence will grow from 3.2 zettabytes today to 40 zettabytes and it will be 44 times larger of the amount produced in 2009, with yearly rate of 50%-60%;
- every minute we send 204 million emails, generate 18 million Facebook likes, send 278 thousand Tweets and up-load 200,000 photos to Facebook.

The volume of these new sources has naturally inspired to call these data as Big Data (BD). Nevertheless, an important question is how much is this new information useful for creating statistics. The definition of a BD source is more intricate taking into account that the phenomenon is complex and relevant dimensions changes in accordance with the field of interest. In the Official Statistics, the following definition is been proposed at international level:

Data that is difficult to collect, store or process within the conventional systems of statistical organizations. Either, their volume, velocity, structure or variety requires the adoption of new statistical software processing techniques and/or IT infrastructure to enable cost-effective insights to be made (UNECE, 2014).

Along with this definition the following classification of BD is proposed:

- Human-sourced information (Social Networks): Facebook, Twitter, Tumblr etc.; blogs and comments; personal documents; pictures: Instagram, Flickr, Picasa etc.; Videos: YouTube etc.; internet searches; mobile data content: text messages, user-generated maps, E-Mail. This information is the record of human experiences, previously recorded in books and works of art, and later in photographs, audio and video. Data are loosely structured and often ungoverned.
- Process-mediated data (Traditional Business Systems and Websites): data produced by Public Agencies; medical records; data produced by businesses; energy consumption; commercial transactions, banking, stock records, E-commerce, Credit cards. The process-mediated data thus collected is highly structured and includes transactions, reference tables and relationships, as well as the metadata that sets its context. Some sources belonging to this class may fall into the category of "Administrative data".
- Machine-generated data (Automated Systems or Internet of Things): data from sensors such as: home automation, weather and pollution sensors, traffic sensors and webcams, scientific sensors, security sensor; mobile sensors (tracking) such as mobile phone location, cars and travel sensors satellite images. Data from computer systems logs and web logs. Its well-structured nature is suitable for computer processing, but its size and speed is beyond traditional approaches.

3. Why do the NSIs deal with the use of BD sources?

Afterward the adoption of administrative data in statistical process, the NSIs should treat with a new class of secondary data collected by agencies or private companies with not statistical purpose. Organizational and management challenges relate to the legal framework, privacy, knowledge and capacities of handling a dynamically evolving data and metadata ecosystem have to deal with. Despite these problems the NSIs are approaching to the BD sources for different reasons (UNECE 2014):

Reputational drivers: BD sources have the potential to significantly impact the statistics industry. It is important that NSIs continue to demonstrate their relevance and remain competitive with other emerging sources of data if governments are to continue to see value in Official Statistics. These drivers seek to exploit new opportunities to keep pace with possibilities. This leads to a data-oriented approach where statistical organizations ask how they can make use of new sources.

Efficiency drivers: budget cuts in the NSIs and at the same time producing improved outputs lead to consider new data sources, technologies and methodologies. BD are sought to:

- identify and provide information about survey population units (sample frame and register creation);
- replace survey collection, reduce sample size, or simplify survey instruments (full or partial data substitution);
- ensure the validity, consistency and accuracy of survey data (data confrontation, imputation and editing).

The need for new statistics or statistics with an improved timeliness or relevance: use of new data sources that fill a particular information need to extend the existing measurement of economic, social, environmental phenomena to a high quality for use in policy making. There may be a range of demands that can be assisted through the use of BD:

- Improve timeliness of outputs;
- Enhance relevance or granularity of outputs;
- Increase accuracy or consistency of outputs.

Alternatively it may enable statistical organizations to produce statistics where high quality is less appropriate but can meet public demand on issues of the day.

4. Issues of using BD

Several issues can be identified when using BD: legislation, protection of privacy, confidentiality, intellectual property; communication to manage public

trust and acceptance of data re-use and linking to other sources; communication to tackle a negative public opinion; partnership with data providers, scientific community and IT providers; internal IT infrastructure; governance of access and use of the data; new skills for the NSI staff.

Considering the statistical view-point the paradigm shift from designed data for planned statistics to data-oriented or data-driven statistics offers new challenges. Beyond the descriptive statistics it will be necessary to determine under which conditions valid inferences can be made. Undercoverage and selectivity typically affect BD sources i.e., human generated data may suffer from several biases such as self-selection, self-reporting; absence of metadata obstacles the inference process since many of the interest characteristics like gender or age, are not known; automated systems suffers from a placement bias.

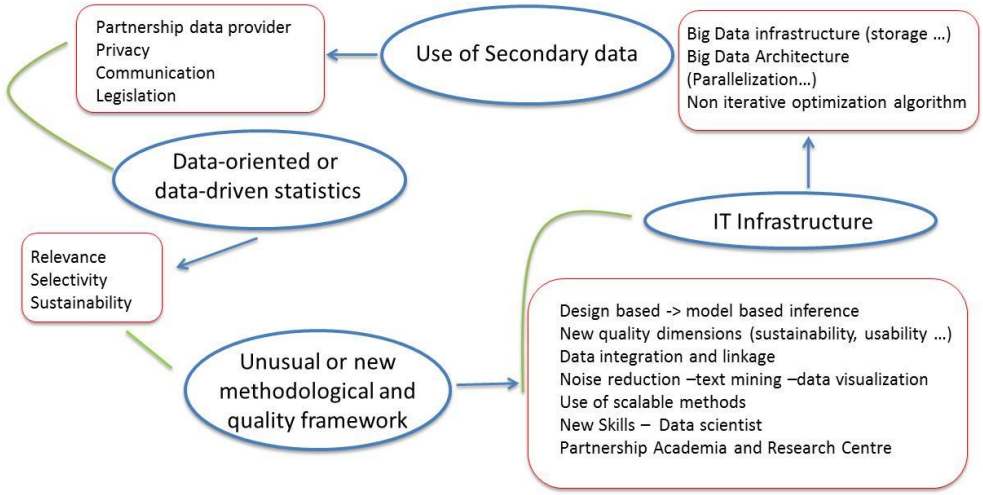
In general, the NSI staff is few experienced on non-probabilistic samples and statistical tools to extraction, transformation and load methods to take unstructured data to a processable form (statistical learning, data mining, data visualization).

The interconnection with the IT aspects yields a lot of efforts to create platforms and services specifically built to handle vast amount of data. Parallelization algorithms (like MapReduce, Hadoop, RHadoop, consistent hashing) to make possible the computation on distributed environments and non-iterative optimization algorithms have to be applied.

Summing up all these aspects the methodological and quality framework to manage the BD can be new or unusual for a NSI.

Figure 1 highlights the innovations or peculiarities and the related implications of the statistical production process when using BD sources. They are: the use of secondary data; data driven statistics; methodological and quality framework; IT infrastructure.

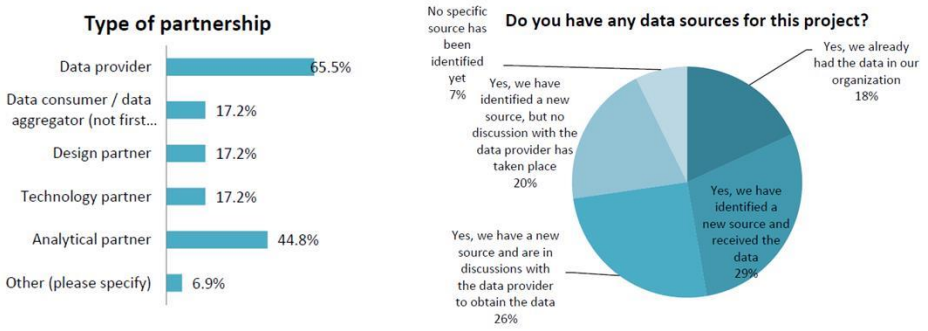
Figure 1 – *Challenges, innovations and related issues with use of Big Data sources in the Official Statistics*



The 2015 survey (first edition) on Big Data projects for Official Statistics conducted by the United Nations Statistics Division (UNSD) and the United Nations Economic Commission for Europe (UNECE) gives a picture of the development of the BD strategies in the Official Statistics. The questionnaire investigated the overall strategies and the specific projects carried out by the NSIs and International Organization. The respondents were 32 NSIs and 3 Organizations per 57 BD projects. The first evidence is that very few respondents have defined a long-term strategy for adopting BD sources and […] *More than two thirds of the organizations explained that they do not yet have defined a quality assessment framework for Big Data sources or the output of analysis* […] (United Nation Statistical Commission, 2015). Likely, to deal with these issues the experiences with real data should be useful. But there is a concrete difficulty to have BD available: partnership with BD providers and the legislation are felt as the main problems to work with the new sources. Follow, skills, IT infrastructure and methodology.

In particular for the 58% of the projects a partnership has been defined or is still in discussion but only the 65.5% of these partnerships are with a data providers (figure 2 left) and the 29% of the projects received data after a partnership agreement (figure 2 right). The 68% of these 57 projects have privacy and confidentiality issues to be tackled.

Figure 2 – *Partnership in the 57 Big Data projects investigated in the Big Data Project Survey*



Source UNSD/UNECE.

5. International and National initiatives on BD

ISTAT is involved in different international activities on the use of BD.

At European level, in September 2013, has been set up the Scheveningen Memorandum on Big Data and Official Statistics which encourages members of the European Statistical System to develop a BD strategy, share experiences and collaborate at the level of the European Statistical System and beyond. Furthermore the Memorandum outlines a global strategy for introducing BD sources in the statistical production process. ISTAT attended on drafting the Memorandum and to the subsequent Task Force for defining the road map and the action plan implementing the Memorandum in concrete. In September 2014, the European Statistical System Committee endorsed the Roadmap and Action plan 1.0 and integrated them into the ESS Vision 2020 portfolio (Eurostat, 2014)

In 2016 and since 2020 two European Statistical System network (ESSnet) projects should be launched. Projects will investigate practical aspects by implementing pilots and applications on BD sources at EU level with many NSIs including ISTAT. The overall aim of these ESSnets is to test parts of the ESS Big Data Action Plan and Roadmap 1.0. In particular, the following results are expected: identify pilots for generating statistics from at ESS level; identification and analysis of output portfolio of BD sources; identification and definition of skills and competences; exchange of information with stakeholders within the statistical system and the research community; development and review of methodological and quality frameworks for BD sources in official statistics; identification, definition and implementation of IT infrastructures for BD

processing; access to BD sources, identification and preparation of non-legal and legal conditions for access and use of BD within the ESS. Experiences obtained in each pilots will contribute to horizontal topics as: methodology, quality and metadata, IT infrastructure, skills, partnerships and communication. Another initiative, where ISTAT participated, is the BD project of the United Nations Economic Commission for Europe (UNECE) launched in March 2014 (UNECE 2014). Four Task Teams have ended their works in the 2014. They focused on: partnership with BD owners; privacy and legal issues; quality of statistics (UNECE, 2014); a "Sandbox" that provided a technical platform for the NSIs to jointly experiment with BD sets and tools. Sandbox activity will continue in the 2015. Moreover, ISTAT is involved in worldwide initiative of Global Working Group (GWG) on Big Data for Official Statistics coordinated by United Nations Statistics Division (UNSD). The purpose is of continuing some of the works already done by the UNECE project. The activities of GWG on these topics began in the mid of 2014 with a Global Survey for investigating the challenges relating to methodology, privacy and access to data, partnerships and skills that the NSIs have to dealt with (United Nation Statistical Commission, 2015). The organization of the working group provides many Task Teams. ISTAT is the leader of Cross-cutting issues, Classifications, Frameworks and Taxonomy Task Team. Next deliverable of the Task Teams is the new questionnaire of the Global Survey. The results will be shown at the Global Conference on Big Data for Official Statistics in Abu Dhabi, UAE the 20-22 October 2015. At National level, ISTAT closed (March 2015) a two year Technical Commission devoted to define a Roadmap for the adoption of Big Data in production process. Several actors coming from Academia, IT and Media companies have been involved and three pilots using BD sources have been carried out:

Web Scraping for the ICT usage and e-Commerce in enterprises (Barcaroli *et al*., 2015): use of crawlers and scrapers for collecting data for the ISTAT Survey on ICT Usage and e-Commerce in Enterprises and text mining techniques in order to estimate the services offered by an enterprise through its web sites. In particular the pilot is focused on the online ordering or reservation or booking. The main target of the pilot is to verify a new technique for collecting information from the enterprises. Partnership with Cineca (consortium of Italian universities, National Research Council and Ministry of Education and Research) has been settled on;

Persons and Places, use of mobile phone data to estimate mobility flows (Furletti *et al*., 2014): **t**he ongoing ISTAT project "Persons and Places" estimates the residences and flows of people by means of administrative registers (Residence register - Work register – Study register). The first release has been an Origin Destination matrix at municipality level. The objective of the project is to produce statistics that are comparable with those obtained in the ongoing project deploying

the massive and constantly updated information carried by mobile phone call data records (CDRs) for estimating population statistics related to residence and mobility. The project has been carried out jointly by the Italian National Research Council (CNR), University of Pisa and a Mobile phone company (partnership agreement with University of Pisa). In 2014 Italian Data Protection Authority authorized ISTAT to utilized anonymized data only for the project Person and Place but there are some problems (ISTAT has not yet used the CDR data);

Google Trends, u**se of query shares (Google Trend Index) for nowcasting labour force statistics** (Bacchini *et al*., 2014). The ISTAT Labour Force survey produces monthly estimates of unemployment rate at National level with a time lag of 1 month. Purpose of the experiment has been to use the Google query shares "job offer" as auxiliary variables for nowcasting and improving the current provisional estimates.

Finally, ISTAT is implementing the use of scanner data for estimating the Harmonized Index of Consumer Prices. A partnership agreement with the Association of Modern Distribution (900 Associates, 32,000 outlets) has been established and two years (2013-2014) scanner data for six market chains have been obtained for more than 15 provinces of Italy. Data are referred to grocery products (almost 20% of the basket of products) and for each outlet by reference (identifies by the EAN code) weekly data (turnover and quantity) are available.

References

BACCHINI F., D'ALÒ M., FALORSI S. FASULO A., PAPPALARDO A. 2014. Does Google index improve the forecast of Italian labour market?, *Proceedings 47th Scientific Meeting of the Italian Statistical Society, Cagliari June 11-13*.

BARCAROLI G., NURRA A., SALAMONE S., SCANNAPIECO M., SCARNÒ M., SUMMA D. 2015. Internet as Data Source in the Istat Survey on ICT in Enterprises, *Austrian Journal of statistics*, Vol. 44, pp. 31–43.

CHEN M., MAO S., LIU Y. 2014. Big data: a survey. *Mobile Networks and Applications*, Vol. 19, no. 2, pp. 171–209.

EUROSTAT 2014. The European Statistical System (ESS) Vision 2020. Technical Report, http://ec.europa.eu/eurostat/web/ess/-/the-essc-comes-to-an-agreement-on-the-ess-vision-2020

FURLETTI B., GABRIELLI L., GAROFALO G., GIANNOTTI F., MILLI L., NANNI M., PEDRESCHI D., VIVIO R. 2014. Use of mobile phone data to estimate mobility flows. Measuring urban population and inter-city mobility using big data in an integrated approach, *Proceedings 47th Scientific Meeting of the Italian Statistical Society, Cagliari June 11-13*.

IDC 2014. Analyze the future, http://www.idc.com.

KHAN N., YAQOOB I., HASHEM I. A. T., INAYAT Z., MAHMOUD ALI K., ALAM M.,SHIRAZ M., GANI A. (2014). Big Data: Survey, Technologies, Opportunities, and Challenges, *The Scientific World Journal*, Vol. 2014, pp 1-18.

UNECE (2014). How big is Big Data? Exploring the role of Big Data in Official Statistics. Technical Report, http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=99484307.

UNECE (2014) A Suggested Framework for the Quality of Big Data. *Deliverables of the UNECE Big Data Quality Task Team*. December, 2014.

UNITED NATION STATISTICAL DIVISION (2015). Results of the UNSD/UNECE Survey on organizational context and individual projects of Big Data Prepared by the Statistics Divisions of UN/DESA and UN Economic Commission for Europe. Background Document, http://unstats.un.org/unsd/statcom/doc15/BG-BigData.pdf.

SUMMARY

Istat international and national initiatives on big data

The National Statistical Institutes have only recently started investigating Big Data as potential data sources for generating official statistics. Especially beginning in 2013, numerous international projects and initiatives have been undertaken. The scope of these activities is to deal with the issues related to the Big Data sources that are typically secondary data collected by agencies or private companies with not statistical purpose. Legislation, protection of privacy, confidentiality, communication, partnership with data providers, scientific community and IT providers, IT infrastructure, governance and new skills for the staff of the National Statistical Institutes are relevant matters. Furthermore, the paradigm shift from designed data for planned statistics to data-oriented or data-driven statistics offers new challenges in the methodological and quality framework. The paper shows the international and national project involving the Italian National Statistical Institute describing the main steps useful for defining a global strategy for introducing Big Data sources in the statistical production process.

Paolo RIGHI, ISTAT, parighi@istat.it