Adding Value to Statistics in the Data Revolution Age

Giorgio Alleva*
Italian National Institute of Statistics – Istat, Rome, Italy – giorgio.alleva@istat.it

Abstract

As many statistical offices in accordance with the European Statistical System commitment to *Vision 2020*, since the second half of 2014 Istat has implemented its internal standardisation and industrialisation process within the framework of a common Business Architecture. Istat modernisation programme aims at building services and infrastructures within a plug-and-play framework to foster innovation, promote reuse and move towards full integration and interoperability of statistical process, consistent with a service-oriented architecture. This is expected to lead to higher effectiveness and productivity by improving the quality of statistical information and reducing the response burden. This paper addresses the strategy adopted by Istat which is focused on exploiting administrative data and new data sources in order to achieve its key goals enhancing value to users. The strategy is based on some priorities that consider services centred on users and stakeholders as well as Linked Open Data, to allow Machine-to-Machine data and metadata integration through definition of common statistical ontologies and semantics.

Keywords: Modernisation of official statistics; users and stakeholders; Linked Open Data.

1. Introduction

In the last half decade Istat has undertaken a major effort to innovate the statistical production process and its products with a view to raising efficiency, enhancing data quality, and fostering the ability to respond to emerging user needs. The strategic vision is consistent with *Vision 2020*, the medium-term transformation programme of the European Statistical System. It has its foundation in the role of technology as a driver for innovation and builds on the Enterprise Architecture model based on lean management and agile organisation to respond to change promptly.

The key pillar of the strategy is process industrialisation based on Service-Oriented Architecture (SOA) and on the standardisation of statistical methods, tools and IT services for data collection, process and analysis, and dissemination. In these main areas of the production process, the modernisation strategy focuses on key priorities. For example a centralised data collection system (including web portals) with related tools is provided in order to reduce response burden and supply tailor-based information to users as well as extensive use of administrative data and new data sources, such as Big Data, for register-based integrated statistical production.

In addition, a unified metadata system is available as driver of production industrialisation based on standards and reuse of data and statistical tools.

Finally, dissemination systems are offered based on integrated data warehouse for data browsing, accessible micro data for research, Linked Open Data portals and related data analytics services, visual analytics tailored to user needs, and data analytics based on tools for stakeholders and policy makers. In this framework a great importance is attributed to multidimensional analyses to share best practices, methods and tools, thus adding value to data.

In the light of the above, key challenges and lessons have emerged such as the key role of the Business Architecture (BA) model for designing process innovation based on agreed corporate principles and ensuring the elimination of domain-specific silos. A sound governance to steer the strategy and change management process are instrumental for implementation based on comprehensive portfolio and project management frameworks. For this purpose, promoting innovation through agile and horizontal corporate networks drives corporate energy for change and builds innovation facilitators for the

strategy. Eventually, communicating modernisation to stakeholders is relevant to demonstrate the benefits of change and build momentum for political support and corporate consensus.

2. The reasons for change

A transformation of Istat structure is currently ongoing in order for it to adapt to the rapidly changing external context and be able to continue producing top-quality statistical information. Various challenges ensue from the change in the demand for statistical data, from the increasing wealth of information available, also unstructured, and from the availability of new methodological and technological tools.

As new IT and communication tools make it easier for most individuals, companies and institutions to access data, the demand for tools allowing to transform them into higher value-added knowledge to take decisions has grown rapidly. This, in turn, puts pressure on data providers to adjust their products and increase data accessibility, allow data management and mining by users, and support result dissemination. Traditional data collection systems are not appropriate any longer due to high costs and high response burden, which is causing decreasing response rates.

Data collection tools are being developed with the purpose of minimising response burden, reusing information that has been already collected from respondents through a variety of sources and building platforms that allow seamless exchange and data integration, in a register-based statistics approach.

These changes are occurring in a context where the provision of statistics by the private sector has also improved. New competitors able to produce statistical information based on such wealth of data have also appeared on the market. Both supply-side factors, such as innovation in technologies to collect and analyse data, and demand-driven forces have increased the scope for unofficial statistical productions. Thus the traditional monopoly of statistical agencies has changed, although the need for high-quality statistical data based on international standards has not been replaced. However, in this new environment public and private producers of statistical data can compete and cooperate in a market that has grown in size and complexity.

Internally, there is room for improvement in terms of efficiency and effectiveness. The current organisational structure is still based on domain-specific silos, fragmentation of the basic organisational units and on local or vertical know-how which does not promote reuse. Research and innovation are based at separate into each structure rather than at corporate level. This creates duplication and lack of consistency of solutions, limited interoperability and limited capacity to exploit technological opportunities.

The governance system is weak: general services are not easily accessible and there is a drive to obtain services at the local level, generating redundancies and inefficiencies.

3. Modernisation at Istat

Istat modernisation project pursues the following main objectives:
- give incentive to the development and exploitation of methodological, technological and organisational innovation;
- enhance and reorient staff skills;
- develop a specific policy of corporate social responsibility;
- reduce response burden through the reuse of available data and information;
- enrich the range of statistical information and services on offer sharing methods and tools to add value to data;
- improve the Institute performance, by reducing costs and response times, and the quality of statistical production and processes.

Starting from September 2014, Istat has developed a model for defining a new organisational and production structure. The modernisation process has already started with centralisation of some functions, extensive access to administrative data, and standardisation under way. An in-depth assessment of the *as is* situation was conducted by an internal survey on Istat structure functions and capabilities followed by the definition of a *to be* model in terms of principles, activities, data, organisation and governance.

In partnership with external experts and through study visits in some National Statistical Institutes (NSIs), also an analysis of international best practices was carried out that highlighted the relevance to invest in a register-based approach to industrialisation.

NSIs of Nordic countries have started developing statistical registers from administrative data already since the 1960s: today in Sweden about 80% of statistics are register-based with a notable advantage in terms of reduced costs and response burden, while quality, timeliness and completeness are also safeguarded. At the same time register data offer a large potential because different registers can be linked together on the basis of clearly defined keys. Base statistical registers identify the statistical population: they are founded on and linked with administrative registers (gathered in satellite registers) and also partly based on surveys. An organised governance mechanism is needed so that there are clearly defined responsible persons for the system variables, the core variables that define the population of the base register, and thematic variables (mainly in satellite registers).

Relations between responsible persons and users should be regulated by a Committee for Register Management, which can consult the Business Architect or the IT Architect when needed.

The new model governing such a transformation is based on a BA (see Barcaroli, Falorsi, Fasano and Mignolli in *STS071* of this Congress) which can guide the operational changes and all stages in the production chain, highlighting the required infrastructure that should be put in place.

The BA model used at Istat has been developed as a joint task of the *Statistical Network* (Statistical Institutes of Australia, Canada, Italy, New Zealand, Norway, United Kingdom) and the *ESSNet on Standardisation* (Eurostat and Statistical Institutes of France, Hungary, Italy, Latvia, Lithuania, The Netherlands, United Kingdom). It is a reference model to optimise processes within an Organisation and make them more efficient, based on the reuse of data, methods, processes, and tools. It covers both statistical activities and strategic organisational tasks and capabilities, and is based on an activity model composed of four business lines: Strategy, Corporate Support, Capabilities and Production.

The new organisational structure of Istat is divided into two main areas in compliance with the BA model:

- a Statistical production area dealing with statistical themes, the integrated register system based on administrative data, surveys (decreasing progressively in the next future), processes and analyses (mainly focusing on multidimensional analyses);
- a Corporate support service area subdivided into two main groups of centralised services, one dealing with specialised support (methodology, IT, data collection and dissemination) and more related to the production processes, the other one dealing with more general and legal support (management/administration).

A specific governance policy is also assured through a strategic and operational planning in order to make these two areas working efficiently and effectively on the basis of a portfolio and project management.

4. Operationalising the project

There are many lines of action for the achievement of the modernisation objectives at Istat that attribute a central role both to the strengthening of the governance mechanisms (in particular for strategic and operational planning) through a coordinated management of a portfolio of innovation projects and to human resources. In general, they include also redesigning both of production

processes according to a register system and of an integrated system for the management of skills implying the definition of specific training courses to support the new production model.

In more detail with regard to human resources, specific actions are focused on the reduction of the organisational fragmentation and of the static nature of responsibilities, enhancing employee mobility within Istat internal structures. In addition, particular attention is given to the realisation of a corporate social responsibility system exploiting experiences and good practises already in use internally.

In the new model, as outlined in the Business Architecture, the entire statistical production process is driven by the required output and metadata. Assessing user needs represents the very first stage for tailoring products and services, developing new tools and partnerships - also involving the private sector - thus allowing Istat to assess user real needs.

The development of a generalised data collection system can already count on a business statistical portal that collects data effectively via web through a one-stop-shop and uses all available information consistent with the e-government strategy. This web portal manages back-office operations, moving away from a vertically integrated stovepipe model; structured interactions with respondents are also provided, including returning key tailored information to them.

Exploiting administrative data is ensured also through a Single Entry Point, an infrastructure for all the information used in the statistical production that consolidates the technologies of the different access systems in place and provides on line documentation to enhance data integration.

In addition, Istat is testing Big Data mining engines and web-scraping infrastructures (also in partnership with research centres and the private sector) to collect auxiliary or substitutive information for statistical production, using new sources available on the web and other sources. These experiments will enter the production process in the medium term.

Furthermore a study is being carried out on the integration of Istat social survey system with the permanent population and housing Census.

Processing and analysis components of the modernisation programme are focused on the creation of generalised tools to carry out data integration, validation and estimation functions. There are several activities that are conducive to establishing a set of standard methods and tools that would constitute the library (a real *Istat Store*) to be used by analysts for their tasks, addressed to the development of SOA-compliant standard software to perform data processing functions and to achieve a gradual replacing of existing instruments, so as to move away from the stovepipe model. Moreover, new methods are available to improve the support of the integration between survey data and administrative sources in the statistical production process, thus enhancing estimate quality.

For this purpose, huge efforts are aimed at the construction of a generalised conceptual methodology to use administrative data for sampling, validation and estimation as well as at the production of enhanced methodological documentation and manuals, on line and user-friendly software repository tools, standard certification guidelines and procedures.

Data dissemination components represent some of the most advanced elements of the programme. The strategy followed is to build on existing infrastructures in order to enhance their capacity and strengthen the links between production and dissemination data warehousing tools.

Within this framework, a relevant role is played by the expansion of aggregate data dissemination tools so as to link the corporate data warehouse to geo-analytics visualisation tools with a view to reinforcing authoring tools and interactive access to all data by users, also profiling them.

Importance is also given to strengthening machine-to-machine data communication tools and hubs to support information exchange and enhance access to data produced by the Italian National Statistical System network, as well as to centring micro data dissemination on the availability of Public Use Files and Files for Research, so as to enhance the access to Istat Research Data Centres and develop the Italian Data Archive in partnership with other public and private institutions.

Finally, the link between communication tools and dissemination platforms is ensured, providing consistency and constant updating of data on the website, rapidly passing from physical to virtual assets.

Istat is also experimenting with new dissemination techniques in a dedicated portal relating to short-term economic statistics in order to increase integration, broaden contents and promote this information through collaborative work among different internal structures.

For each single topic, this section of the website contains thematic pages with descriptive analysis, automated chart and file creation, time series, as well as direct access to press releases and data at national and international level. This project has low implementation costs and large potential spillovers.

In the effort of putting users in the driver's seat, Istat is also experimenting with analytics based on Linked Open Data (LOD), which can add value through integration: common ontologies are defined to drive internal data and metadata integration, while statistical ontologies (together with data) act as a critical link for LOD Cloud. For example, Istat in partnership with the Agency for Digital Italy (Agid) is experimenting the production of official classifications under LOD form on the Agid website.

Furthermore, a Linked Open Data portal with three access points will cover the requirements of different possible user tailoring for different needs, linking official statistics with other data.

5. Harnessing administrative and new data sources

More abundant and higher-quality data are required to meet increasingly complex knowledge demands from users and stakeholders in modern societies. At the same time, budget pressures and costs associated with response burden limit the scope for new surveys. Consequently exploiting administrative data should play a central role in a register-based system characterised by both base and satellite statistical registers.

Base statistical registers (Persons and households, Economic units, Geographical/territorial units, Activities) have links that relate each other and connect them to the administrative sources from which they are derived. An administrative register can provide information both for social and economic statistics allowing several integrations (for example persons and businesses).

Satellite statistical registers contain other variables that can be derived from administrative registers or obtained by integrating information from sample surveys.

In addition, harnessing new data sources is decisive to provide an adequate supply of information and knowledge while enhancing the effectiveness of statistical production processes. Big Data and the Internet as data sources represent new opportunities and challenges for official statistics that should incorporate all innovative potential data sources as much as possible into their conceptual design (see G. Barcaroli's contribution for *IPS044*). In this context some experiments are being conducted at Istat which will be fully operational and enter the production process in the medium term.

In more detail, these experiments concern the use of web scraping and text mining techniques. They have been applied in the survey on "*Information and Communication Technology in enterprises*" with the aim of substituting traditional instruments of data collection and estimation or of combining them in an integrated approach to produce the required estimates, harnessing both survey data and data from the Internet.

In the framework of an integrated research project aimed at evaluating the potential of Big Data, Istat has also experimented with the production of preliminary estimates using Google Trends indicators for nowcasting purposes in the Labour Force domain.

Another ongoing experiment within Istat is based on the use of mobile phone (tracking) data for mobility statistics with the aim of producing origin/destination matrix of daily mobility for work and study at the spatial granularity of municipalities. The results obtained using Big Data as a proxy of the presence and mobility of individuals are generally encouraging and promising, in comparison to analogous statistics obtained with official data.

Other areas of work include the use of alternative social media indicators (e.g. work on Twitter and Facebook), application to other variables (e.g. retail sales, employment), calculation of road traffic estimates using webcams (mainly on motorways) and smart city indicators (see De Santis, Fasano, Mignolli and Villa in *STS072* of this Congress).

6. Concluding remarks

In the current situation characterised by a strong dynamism the need for varied and timely information at multidimensional level increases and official statistics is called to respond efficiently and effectively by intervening and transforming the production data mode.

For this reason in the last decade NSIs have revisited their business model to remain relevant in a fast-changing world where advanced technologies and communication impose challenging structural reforms also in the statistical production chain, sharing a common vision.

Achieving modernisation is a task that combines different skills and needs to generate political consensus from staff for reforms and innovation targets.

Key challenges within the implementation of this strategy include changes related to legislation with regard to the disclosure control affecting data access and use, as well as privacy matters involving the management of public trust and the acceptance of data reuse and its links to other sources.

At same time, it is necessary to continue ensuring high data quality together with integrity and suitability of statistical methods within this innovative framework, also investing in human resources by developing new skills. Human capital indeed is the core ingredient in innovating and creating a strong and sustainable support for this ambitious transformation agenda. Consequently, a good staff training is one of the most relevant factor to put in place modernisation programmes and should build human resources and strategy managers together, with a view to generating an environment prone to innovative thinking.

Furthermore investing in communication and sharing is very relevant both to increase statistical culture among users and to foster international collaboration while engaging stakeholders in decisions and creating partnerships that involve research centres, university departments and the private sector in order to improve global synergies.

References

Alleva, G., (2014). *Integration of business and trade statistics: limitations and opportunities*. DGINS Conference, Riga 24-25 September 2014.

Arezzo, M. F., Alleva, G., (2012). *Estimating the Probability of Moonlighting in Italian Building Industry*. In: Di Ciaccio, A., Coli, M., Angulo Ibanez, J.M. (Eds.). *Advanced Statistical Methods for the Analysis of Large Data-Sets*. Berlin: Springer-Verlag.

Baldacci, E. (2013). *Innovate or perish, Italy's Stat2015 modernisation programme*. 59th World Statistics Congress, Hong Kong, 25-30 August.

Barcaroli, G., Falorsi, P.D., Fasano, A., Mignolli, N. (2014). *A Business Architecture Model to Foster Standardisation in Official Statistics*. European Conference on Quality in Official Statistics - Quality 2014, Vienna 2-5 June 2014.

Citro, C.F. (2014). *From multiple modes for surveys to multiple data sources for estimates*. Survey Methodology, Statistics Canada.

FAO (2014). *The Global Strategy to improve Agricultural and Rural Statistics*. Technical Report on the Integrated Survey Framework, Technical Report Series GO-02-2014.

McCandless, D. (2010). *Data is the new soil*. TED talk on visualising data, Columbia Journalism Review.

Wallgren, A. and Wallgren, B. (2013). *Register - based Statistics: Administrative Data for Statistical Purposes*. John Wiley & Sons, Ltd, Chichester, UK (second edition).