European Conference on Quality in Official Statistics (Q2016)
Madrid, 31 May-3 June 2016

Data Ecosystems: a new challenge for official statistics

Donatella Fazio[1], Enrico Giovannini[2], Marina Signore[3]

[1] Italian National Statistical Institute, Rome, Italy; dofazio@istat.it
[2] University of Rome Tor Vergata, Italy; enrico.giovannini@uniroma2.it
[3] Italian National Statistical Institute, Rome, Italy; signore@istat.it

Abstract

Nowadays, NSIs are called to open the doors to new ways of constructing statistics in an era characterised by an infinity of new data sources available on Internet or coming from satellites, sensors, etc.. The very heterogeneous *liquid data* floating on the web is supplied by several providers. On one side, the Big and Open data are available as indirectly generated by citizens, companies and administrations for many reasons (such as use of Social Media, on-line shopping, mobile phones, etc.) and belong to specific providers (private and public). On the other side, the crowd-sourced data is voluntarily collected by civil society communities through Web2.0 collaborative platforms sharing information for different scopes. Recently, the use of new data sources is at the forefront of methodological and experimental studies carried out by NSIs, Academia, research institutes and private sector. Most of the experiences developed so far are using the Big and Open data in a way to save time, money and to reduce the response burden, namely as inputs to: i) construct new statistics - as for administrative data; ii) replace existing surveys; iii) produce now casts. The exploitation of the crowd-sourced data has to follow a different direction. This data can have great potential (especially at sub-national level). The grass root local data generated by communities can provide valuable information for taking a "closer" picture of the local reality and emerging phenomena, even though its quality cannot simply be assessed following traditional quality frameworks. This paper explores how non-official data, especially crowd-sourced data, can be used by official statistics. It argues that they can represent "satellite" information to complement official statistics and stimulate social innovation. The construction of Data Ecosystems is a great opportunity to build up a knowledge-driven society where the communities can raise their voices and be proactive actors for a conscious societal sustainable progress.

Keywords: complement official non-official data, web civil society communities, social innovation.

1. New technologies for official statistics for society at large

New technologies are opening new ways for constructing statistical information. The new sources of data available on Internet or coming from satellites, sensors, etc. are calling the National Statistical Institutes (NSIs), jointly with the research community, to explore how to exploit these data in order to enrich the statistical information. These new data floating on the web are supplied by several providers (public and private). Moreover, most of these data are *indirectly* generated by citizens, companies and administrations for several reasons (such as use of Social Media, on-line shopping, mobile phones, etc.), while some valuable data and information are *voluntarily* collected by civil society communities through Web2.0 collaborative platforms, designed to share information for different scopes. While a lot of projects are being carried out by NSIs and others on the first type of data, the so-called "crowd-sourced data", locally grass root generated by citizens, are less explored by NSIs.

1.1. Web2.0: a new relationship between NSIs and society.

The interactivity provided by Web2.0 technology has become crucial for NSIs, also because it leads to strengthen the relationship between NSIs and society at large. The involvement of citizens and stakeholders in official statistics has deeply evolved over time, going through several phases described below (considering and indicative timeline with respect to the first developed NSIs).

Since their establishment (1926 for Istat) to the '80s, the NSIs invested a lot in methodological research to produce robust statistics following a "top-down" approach: the data collection were carried out by sample surveys and censuses submitting a questionnaire - on paper - to citizens, enterprises and other institutions. The questionnaires were designed by committees of "experts" defining ex-ante the societal information needs and the ICT solutions were primarily devoted to process the data, mainly released on paper. Citizens and stakeholders were essentially *respondents*, while statistics were mainly used by policy makers and businesses.

The '80s and the '90s are characterized by a great methodological effort to use administrative data – generated by national and local administrations for their specific scopes - as input to construct the official statistics. The NSIs understood the great importance of looking at sources of data which were continuously available. During this phase ICT played an increasing role and IT solutions for data collection and processing, using (also) administrative data, were used to modernize the collection of data through on line questionnaires, especially those designed for businesses. A faster release of final statistics became possible and paper publications were progressively replaced by Web1.0 tools. Citizens and stakeholders were still *respondents*, but became more and more *users* of final data.

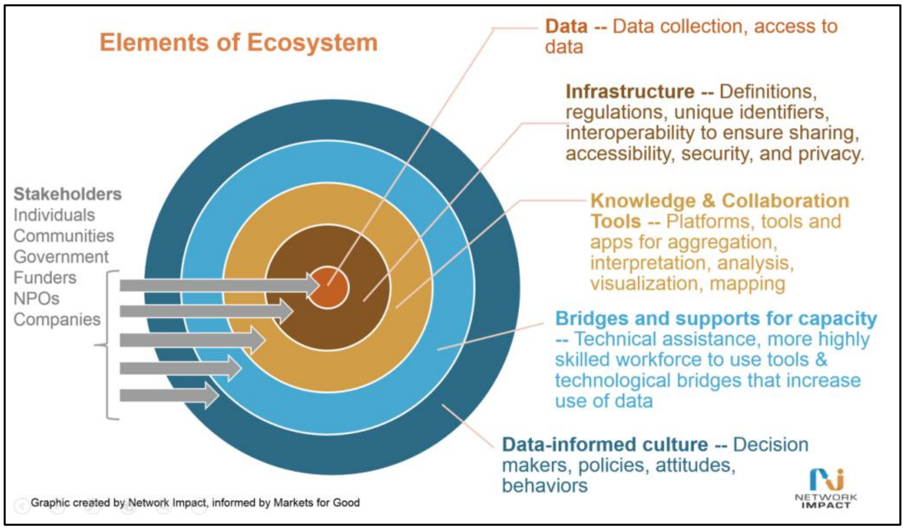
Since the beginning of the new millennium, the research has faced two epochal revolutions: the Web2.0 and the "data revolution", on one side, and the debate on "GDP & Beyond", on the other side. Over the last 10 years, the "GDP & Beyond" debate is having profound effects on NSIs and the research community. The necessity to combine GDP data with a set of multiple indicators to have a comprehensive measure of economic and societal progress is radically changing the way of "making statistics". It helped to introduce new concepts in the measurement of well-being (both subjective and objective) and of its sustainability. As these concepts may be varying over time and across regions, the involvement of citizens and policy makers in defining what well-being means became unavoidable. This made NSIs more and more aware that all sectors of society must be part of this discussion: the experiences carried out by Istat (http://www.istat.it/it/misure-del-benessere) and Office for National Statistics of UK (https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing) on "stakeholder inclusion" through a deliberative process co-managed with the society at large represent an important step towards a more "bottom-up" approach. In this way, citizens and stakeholders become *respondents*, *users* and also *co-designer* and *interpreters* of data.

Moreover, the Data Revolution [1] opens huge opportunities to combine society-generated non-official data (Big, Open and crowd-sourced data) with the data produced by NSIs. Of course, there are risks associated to this process, but non-official data and information can also represent a great opportunity for the construction of "better statistics". Great part of these data are available via the digital initiatives (crowd sourced platforms) fed by components of the society at large, such as civil society organisations, social entrepreneurs, and several communities representing various categories of local actors. Presently, NSIs are actively engaged with studies and pilot projects to treat the usage of Big and Open Data at global, national and sub-national level [2], but grassroots locally generated data have a great potential to provide valuable information for taking a "closer" picture of the local reality and discover emerging phenomena. In this way, citizens and stakeholders are *respondents*, *users*, *interpreters* and *producers* of data (i.e. *prosumer*, to use a classical Web 2.0 wording).

2. Data Ecosystems

The growing demand for real-time data and information and people's willingness to be protagonists in the construction of better statistics can lead to a new model for constructing and accessing statistical information. Overcoming the "one-way" approach (from data producers to final data users) the reality is moving towards the so-called "Data Ecosystem approach" [3]. An exemplification of the architecture of a Data Ecosystem is illustrated in Figure 1.

Fig. 1: Elements of a Data Ecosystem


Elements of Ecosystem
Data -- Data collection, access to data
Infrastructure -- Definitions, regulations, unique identifiers, interoperability to ensure sharing, accessibility, security, and privacy.
Knowledge & Collaboration Tools -- Platforms, tools and apps for aggregation, interpretation, analysis, visualization, mapping
Bridges and supports for capacity -- Technical assistance, more highly skilled workforce to use tools & technological bridges that increase use of data
Data-informed culture -- Decision makers, policies, attitudes, behaviors
Stakeholders
Individuals
Communities
Government
Funders
NPOs
Companies
Graphic created by Network Impact, informed by Markets for Good
NETWORK IMPACT

Source: http://www.networkimpact.org/leveragingtech/

The figure shows how the Ecosystem is a "portal" based on a fulcrum of multi-sources open data - uploaded by traditional data producers (NSIs, etc.) together with new data producers (communities), embedded in an interoperable infrastructure to allow data sharing, accessibility, security and privacy. The portal can be made available with tools and applications for treating data and information (aggregation, interpretation, analysis, visualization, mapping, narrative, storytelling, etc.) with the aim of building a data-informed culture to steer policy actions and drive the collective and individual behavior towards a knowledge-based social innovation. Going beyond the chart above, the management of the Data Ecosystems is based on the concept of *data cycles*, where users of data and information (expert and non-experts) are involved in re-shaping the statistical information in a jointly top-down/bottom–up approach. So the role of a *passive final user* is replaced by a *proactive community of users*, able to add quantitative and qualitative information that can *enrich* and *possibly correct* the official statistics. In this approach, data management and capacity building can be supported by the producers of official data and by new figures belonging to the community of the Ecosystem, the so called *infomediaries*, i.e. intermediate consumers of data such as builders of apps and data *wranglers* (http://blog.okfn.org/2011/03/31/building-the-open-data-ecosystem), that can develop applications to facilitate a better understanding and use of data and information.

3. Experiences of Data Ecosystems

Against this background, this section reports two recent experiences on the construction of Data Ecosystems. The first experience, at global level, is the Web-COSI Wiki of Progress Statistics, set-up within the activity of the FP7 Web-COSI project funded in 2014 by the European Commission. The second experience, OpenGrid, set-up by the City of Chicago, represents a best practice at local level.

3.1. The Web-COSI Wiki of progress statistics

The project Web-COSI [4] - (*Web Communities for Statistics for Social Innovation*)[1] was designed to engage communities (civil society, social entrepreneurs, NGOs) alongside

[1] The project was carried out by a four-partner Consortium, led by the Italian National Statistical Institute – coordinator D.Fazio- in partnership with the Organisation for Economic Co-operation and Development (OECD – www.oecd.org) and alongside Lunaria – an Italian Association for Social Promotion (www.lunaria.org), and i-genius – a Social Entrepreneur web community from UK (www.i-genius.org).

traditional stakeholders and institutions (NSIs, international organisations, businesses, governments, etc.) from around the world for fostering the statistical information “beyond GDP”, complementing official and non-official data. The main goal was the set-up of a Wiki of Progress Statistics: a crowd sourced portal of data and information on well-being and progress, built on Wikiprogress.org[2] (hosted by OECD). The portal was designed using the open-source software (CKAN) to enable the various developers to create and share new tools and solutions. In the platform there are no barriers to uploading data, as users are merely required to create an account with an email address, although content is checked for relevance, to reduce spamming. The portal can host open data, making them freely available to everyone, who can reuse data without restrictions from copyright, patents or other control mechanisms. The portal allows for the visualisation of well-being data and initiatives, easily enabling their geographical mapping. Actually, the portal is organized in five main areas: i) an interactive crow sourced map of initiatives and organizations around the world; ii) a data portal where is possible to download and upload official and non-official data; iii) a Youth Portal devoted to involve young people through on line discussions, target campaigns, blogs, competitions, etc.; iv) a European Wikiprogress University Programme to establish partnership among universities for relevant courses and training modules accessible to students interested in research on well-being and societal progress. Moreover, the portal provides a Forum for an open dialogue – using blogs, newsletters, and the Social Media - among old and new actors interested in developing a better knowledge around “beyond GDP”.

Since its first release in August 2014, the portal has mapped about 400 initiatives carried out by more than 200 organisations and collected 917 resources (including data and documents). So far, the impact of the portal is really impressive: 51.000 visits and about 74.000 pages views by a community of users of about 40.000 people.

[2] The Wikiprogress portal was launched by OECD in 2009 and thanks to the Web-COSI project it has been revamped. The newly updated Wikiprogress site, was showcased at the 5th OECD World Forum on “Statistics, Knowledge and Policy: Transforming Policy, Changing Lives”, Mexico, October 2015, during the Lunchtime Panel "Opening up well-being statistics to new audiences: opportunities and challenges", 14 October (http://www.oecd-5wf.mx/)

3.2. OpenGrid by the City of Chicago

OpenGrid is an open-source geo-spatial *situational awareness* platform developed by the City of Chicago that lets users explore in real-time data and information about the locality (www.opengrid.io). The portal was extensively presented by Tom Schenk, Chief Data Officer of the City of Chicago (USA), at the 5th OECD World Forum, during the Panel session: New tools and approaches for well-being policy - Big and crowd-sourced data[3]. Chicago's open data portal provides almost 600 datasets that are updated on a daily basis. Data concern a wide range of subjects, from crimes to the quality of water on beaches and includes many other important items for the quality of life of residents (such as red light and speed camera violations, problem landlords, and public chauffeurs, etc.). For the City of Chicago the open data portal constitutes a component of governance: an Ecosystem around data of the locality aimed at including multiple stakeholders and initiatives, well beyond the aim of achieving transparency. The IT framework of OpenGrid is open-source so that can be downloaded to quickly allow automated updates. It can be freely used by other governments. Improvements and corrections can be submitted by users. The portal has a large, vibrant, productive, civic community of users, led by Chicago residents interested in technology and society. Smart Chicago Collaborative (http://www.smartchicagocollaborative.org/) and other non-profits organisations provide assistance for the implementation of the portal, meeting regularly with city officials. Through the Smart Chicago's Civic User Testing (http://www.cutgroup.org/), made up of regular residents who provide feedback on applications (ensuring they reach beyond a technical audiences), the community of users has produced several helpful apps to enrich the usage of the portal. OpenGrid, comprising a large number of apps and websites to the benefit of the citizens, has become a pillar to deliver services close to real citizens' needs. The portal serves also as the Town Square, providing common topics of conversation for everyone and also supporting the digital literacy.

[3] 5th OECD World Forum on "Statistics, Knowledge and Policy: Transforming Policy, Changing Lives", Mexico, October 2015, during the Parallel sessions on New tools and approaches for well-being policy –"Big and crowd-sourced data", chaired by Enrico Giovannini, 15 October (http://www.oecd-5wf.mx/)

The OpenGrid is a vivid example of how new technology, based on the Web2.0 interactivity and on Open Data, can strongly contribute to establish a constructive dialogue among administrators and citizens for better policies. A knowledge-based territory is the fundamental condition towards a social innovation based on a participative and inclusive democracy.

4. A new paradigm for quality

4.1. The drivers for quality

The concept of data quality has largely evolved over the years. Quality requirements and frameworks were developed and adjusted to meet emerging needs and changes in the society. The users' role in quality matters also changed dramatically.

At the start, quality corresponded to accuracy of sample estimates following methodological definitions and addressing topics such as optimal sample designs with regard to costs, sampling variances and desired level of detail of the estimates. Starting from the '60s of the last century, the non-sampling errors in sampling surveys and censuses were formalised in methodological papers [5]. This called for a shift from the product (or output) quality to the process quality, thus introducing in NSIs a quality management approach and the principles of Total Quality Management philosophy already developed in the industrial context [6]. Almost at the same time, users and stakeholders were given increasing importance in the quality management: they were recognised as being the starting and the ending point of the production process cycle (i.e. the statistical process starts by identifying users' needs and ends with the assessment of the degree of users satisfaction towards statistical information). Consequently, quality was no more expressed by a single number (e.g. coefficient of variation) but by a set of requirements taking into account users' needs (e.g. relevance, accuracy, timeliness, accessibility,…). Indeed, the users were not seen as the final passive "consumers" of statistical information but started to play an active role being consulted and gradually involved in the production activity itself (e.g. via the establishment of Committees of Users). The central role of the users introduced a new dynamic in quality: it was no more the problem of methodological optimisation but balancing and managing trade-offs among different quality dimensions that have different weights for different users and for different uses. More recently, the European Statistics Code of Practice further extended the quality framework to the institutional environment in which the statistical organisation operates since it affects the way in which processes are built and products are produced and disseminated [7]. Nowadays, NSIs are changing rapidly their approach to non-official data sources: reducing costs and response burden, and increasing timeliness are the drivers for change. Quality frameworks are evolving as well as new data sources are used to complement, integrate or substitute direct surveys. Nevertheless, they still are at different level of maturity: i) fully developed for direct surveys; ii) well developed but still not fully consolidated and harmonised for administrative data sources, and iii) just at an initial development phase for Big Data [8].

The Data Revolution and the construction of Data Ecosystems require a new shift in quality. A holistic view is necessary based on the interrelations among different disciplines and fostering the dialogue among different experts such as the statistician, the information architect, the IT expert and the (expert and non-expert) users who are becoming also data producers.

4.2 Data quality at the local level

The exploitation of new data sources, particularly the grass root locally generated data lead to reflect on the meaning of data quality at the local level. New questions arise: Is it still convenient to apply the same quality framework to: a) all statistical outputs, and b) all geographical levels, namely global, national and local? Recognising the specificities of the local level is the starting point to identify the quality requirements that really matter to users. Probably, the answer is not developing a new quality framework, or adapting the existing ones to the local level. Maybe, what is needed is a new paradigm, where it is recognised that some dimensions have less weight at the local level compared to the national or global one. A lot of useful information at the local level, e.g. to make analyses on environmental sustainability or on resilience of local communities (or more in general on topics related to the GDP & Beyond debate), might benefit from data at the local level produced by official statistics but not disseminated because they do not meet overall quality requirements (incomplete data sets, not validated data, not comparable series,…). Nevertheless, users request such data even though they are aware of their limitations. In a sense, the problem for NSIs is no more to manage trade-offs between quality dimensions, but to allow access to statistical data and more and more often to micro-data (safeguarding the confidentiality) correctly informing the users on limitations and cautions in the use of data. This implies a shift to conscious (implying statistically educated) users.

The debate recently launched by Eurostat on quality labelling might provide an answer to these problems, pushing towards the release of information that does not meet all the quality requirements that official statistics usually meet in compliance with the ES CoP, under the label “Experimental statistics” or similar labels to clearly identify them.

5. References

[1] Independent Expert Advisory Group on a Data Revolution for Sustainable Development (IEAG) - chaired by E. Giovannini (2014), A world that counts. Available at: www.undatarevolution.org

[2] Eurostat European Commission (2015), NTTS 2015 Reliable Evidence for a Society in Transition.http://www.cros-portal.eu/sites/default/files//NTTS2015%20proceedings.pdf

[3] Heimstädt M., Saunderson F., Heath T. (2014), Conceptualizing Open Data Ecosystems: A timeline analysis of Open Data development in the UK, Freie Universitat Berlin, School of Business & Economics, Discussion Paper, Management, 2014/12. Available at: http://edocs.fu-berlin.de/docs/servlets/MCRFileNodeServlet/FUDOCS_derivate_000000003562/discpaper2014_12-2.pdf

[4] European Commission (2015), FP7 Web-COSI project-Web Communities for Statistic for Social Innovation- coordinated by D. Fazio, ISTAT. See www.webcosi.eu

[5] Hansen, M.H. Hurwitz, W.N., Bershad, M.A. (1961), Measurement errors in censuses and surveys. Bulletin of the International Statistical Institute, 38, Part II, 359-374.

[6] Lyberg L., Bergdahl M., Blanc M., Booleman M., Grünewald W., Haworth M., Japec L., Jones T., Körner T., Lindén H., Lundholm G. Madaleno M., Radermacher W., Signore M., Zilhão M.J., Tzougas I., van Brakel R., (2001), Summary Report from the Leadership Group (LEG) on Quality”, Proceedings of the International Conference on Quality in Official Statistics, Stoccolma 14-15 May 2001, CD-ROM.

[7] European Commission (2011), European Statistics Code of Practice - revised edition 2011 http://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice

[8] UNECE, (2015) “A Suggested Framework for National Statistical Offices for assessing the Quality of Big Data”, NTTS 2015 (New Techniques and Technologies for Statistics) Conference, http://www.cros-portal.eu/sites/default/files//NTTS2015%20proceedings.pdf