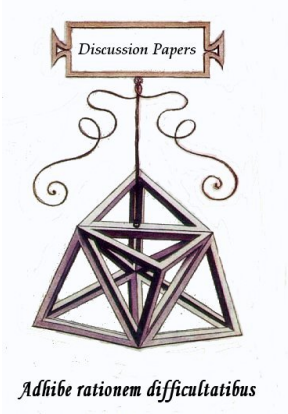
Discussion Papers
Collana di
E-papers del Dipartimento di Economia e Management – Università di Pisa



Monica Pratesi
Claudio Ceccarelli
Stefano Menghinello

Citizen-Generated Data and Official Statistics: an application to SDG indicators

Discussion Paper n. 274

2021

Discussion Paper n. 2021, presentato: Giugno 2021

Indirizzo dell'Autore:

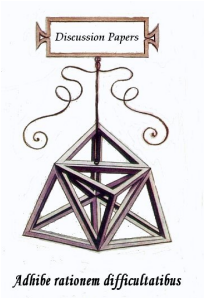
Monica Pratesi, Department of Economics and Management, University of Pisa;
monica.pratesi@unipi.it

Claudio Ceccarelli, Directorate for data collection, National Statistical Institute (ISTAT), Rome;
clceccar@istat.it

Stefano Menghinello, Directorate for data collection, National Statistical Institute (ISTAT), Rome;
menghine@istat.it

Si prega di citare così:
Pratesi M, Ceccarelli C, Menghinello S. (2021), "Citizen generated data and Official Statistics: an application to SDGs indicators", Discussion Papers del Dipartimento di Economia e Management – Università di Pisa, n. 274 (http://www.ec.unipi.it/ricerca/discussion-papers).

Discussion Paper
n. 274



Autori

Citizen-Generated Data and Official Statistics: an application to SDG indicators

Abstract

Official statistics are collected and produced by national statistical institutions (NSIs) based upon standardized questionnaire forms and a priori designed survey frame. Although the response to NSIs' surveys is mandatory for respondent units, increasing disaffection in replying to official surveys is a common trend across many advanced countries. This work explores the possibility to use Citizen-Generated Data (CGD) as a new information source for the compilation of official statistics. CGD represent a unique and still unexploited data source that share some key characteristics with Big Data, while they present some specific features in terms of information relevance and data generating process. Given the relevance of CGD to reduce the information gap between the demand and supply of new or more robust Sustainable Development Goals (SDG) indicators, the experimental setting to assess the data quality of CGD refers to different ways to integrate official statistics and CGD. Istat collects CGD within the framework of a pilot survey focused on key SDG indicators, and the appropriate methodological approach to assess data quality for official statistics is defined according to different data integration modalities.

Keywords: Citizen-Generated Data (CGD), National statistical Institutions (NSIs), Sustainable Development Goals (SDG), Official statistics (OS), Data Science, Latent variables models, civil society organizations (CSOs).
JEL: C81, C83

1. Introduction

National statistical institutions (NSIs) are plagued by human and financial resources constraints as well as by a long-term trend of increasing disaffection of people and businesses in replying to mandatory surveys. On the other side, the data gap between demand and supply of official statistics in growing fast in some specific statistical domains, including for instance Sustainable Development Goals (SDG) indicators. In addition, the use of a traditional data collection approach by NSIs has already been proved to unsustainable in terms of financial budget, human resources availability and additional statistical burden generated on the respondents.

As a result, NSIs are increasing interested in the next generation data in order to overcome the current limitations of their data collection and production frameworks. The identification, classification, data quality assessment of new data sources play a crucial role in the possibility to shift NSIs production frameworks from traditional to new data sources.

Indeed, in the last ten years official statisticians have been discussing on the impact of the Big Data in the production of Official Statistics (OS), highlighting many advantages and also disadvantages of their use. The main question was and is: "What is the future of Official Statistics in the Big Data era?". A lot has been done for the use of Big Data in OS by the International and NSI, including the Istat. For instance, quality issues and model based estimates using Big Data sources were developed by Marchetti et al (2015; 2016) and Pratesi (2017; 2018). Pratesi (2021) has also been focusing on Citizen Science as the global process of digitization is so pervasive that times are mature for studying how using and reusing Citizen Data in the production of OS.

As a matter of fact, Official Statistics have always been evolving and the term "Trusted Smart Statistics" (TSS) was put forward by Eurostat and officially adopted by the European Statistical System (ESS) in 2018 in the so-called Bucharest memorandum to signify this evolution. But using Big Data, smart statistics, citizen data and citizen science in producing OS, could it be a danger? Would OS be under attack either by discussions on trust or by competition with statistics produced with lower quality? For this, the official statisticians of the future have to be more than just data engineers (Radermacher, 2019). These data are nothing more and nothing less than Next Generation Data and following the same evolution track in data production process than NSI have always followed we will answer to the above questions.

This work is organized as follow. Section 2 describes the current limitations of traditional data collection frameworks adopted by NSIs and the need to evolve toward alternative solutions. Section 3 highlights how the evolution process for producing official Statistics and indicators based upon new data sources can generate benefit for stakeholders, respondents, NSIs. Section 4 illustrates the

challenges of Citizen-Generated Data (CGD) and Citizen Science (CS) for OS. Section 5 introduces some conceptual definitions of CGD and the key characteristics of this peculiar kind of data for OS. Section 6 illustrates the standard approach adopted by NSIs to assess the quality of data and its evolution based on the use of new data sources as input data in the OS production process. Section 7 describes a project on the use of Citizen Science and Citizen Data to estimate BES indicators (Equitable Sustainable indicators), which will be implemented by Istat, starting from the assessment of quality of this new type of data for statistical purposes. Section 8 draws some preliminary conclusions and the way forward.

Section 2 – Current and future trends in data collection by NSI

Data collection plays an essential role in the statistical production process set up and maintained by NSI. Data collection approaches adopted by NSI have evolved over time. In this respect, three different phases can be identified. Direct surveys had for a long time been the exclusive way by which the National statistical institutes (NSIs) have collected data. The need to collect information directly from respondent units was primarily motivated by the fact those data were not otherwise available, and the adoption of random sampling techniques was unanimously considered the only possible way to mitigate the unavoidable statistical burden. The possibility to collect information according to rigorous statistical definitions and to directly control the data collection process are traditionally considered the key advantages of this approach.

The emergence of administrative data as alternative sources of information for the compilation of official statistics has generated an alternative approach to data collection in line with the new framework for statistical production rooted in the business register approach. This approach manly benefit NSI in the domain of business statistics, where enterprises are already subject to a significant administrative burden other than the statistical one. Key advantages of this approach rely in the increasing granularity of information, removal of the sampling component of the TSE, strong reduction of the statistical burden, use of well structure data sources, although not always fully compliant to statistical definition in terms of units, classifications and variables.

The rise of Big Data as new sources of data for the compilation of official statistics has opened new opportunities to access a huge amount of timely information across a wide range of statistical domains. Besides IT and methodological problems on how to manage these data according to high data quality standards required by official statistics, the issue of accessibility and privacy consensus has emerged as a critical one, since Big Data are mainly collected by digital platforms with potentially ambiguous business purposes.

Two distinct drivers generate the last wave of change in data collection strategies by NSI. Firstly, an increasing disaffection by respondents, especially physical persons, both in reporting directly to NSI and to give their consensus to the unconditional use of their personal information by digital platforms. Secondly, the increasing demand of data at the sub-regional level connected to emerging phenomena and policies such as SDG, smart cities etc. In order to meet these two distinct drivers of change, NSI are not only exploring new data sources buy also rethinking about their specific role as data collector and the most appropriate strategies to successfully interact with respondent units.
The possibility to exploit Citizen-generated data (CGD) for official statistical purposes, which will be illustrated in detail in the next sections, seems to represent a very promising avenue to collect timely and relevant new data. The motivations are twofold. Firstly, nowadays new technologies and methodological approaches can enhance the accessibility and data processing of this kind of data in a way that is valuable for official statistics. Secondly, leveraging the institutional values and high sense of trust between citizen and NSI, as independent and highly regarded public institutions devoted to the production of official statistics as public goods.

Section 3 - The evolution process for producing official Statistics and Indicators

The mission of OS has always been to provide a quantitative representation of the society, economy, and environment for purposes of public interest, for policy design and evaluation and as basis for informing the public debate. The production of modern OS is based on a system of scientific methods, regulations, codes, practices, ethical principles, and institutional settings that was developed through the last two centuries at the national level in parallel to the developments of modern states (Ricciato et al, 2019).

The Figure 1 illustrates the evolution mechanism of the production process of a general OS system (engine), with its data sources (fuels) and User's information needs (accelerators). We see immediately that statistics and indicators are influenced both by fuels and accelerators. The rise of new data sources can give new fuels for Statistics and Indicators, but it can also act as a multiplier as it provokes new data information needs, becoming accelerators that stimulate further needs to be satisfied. Moreover, the statistical methods and the rules, for example, to guarantee the privacy and the trust on Statistics and Indicators produced, are obviously to adapt to the characteristics of the various data sources.

It is evident that this scheme of the evolution of the OS production process covers the current situation, but it is also valid for all the various breakthrough periods of data collection and statistical production lived by the NSIs.

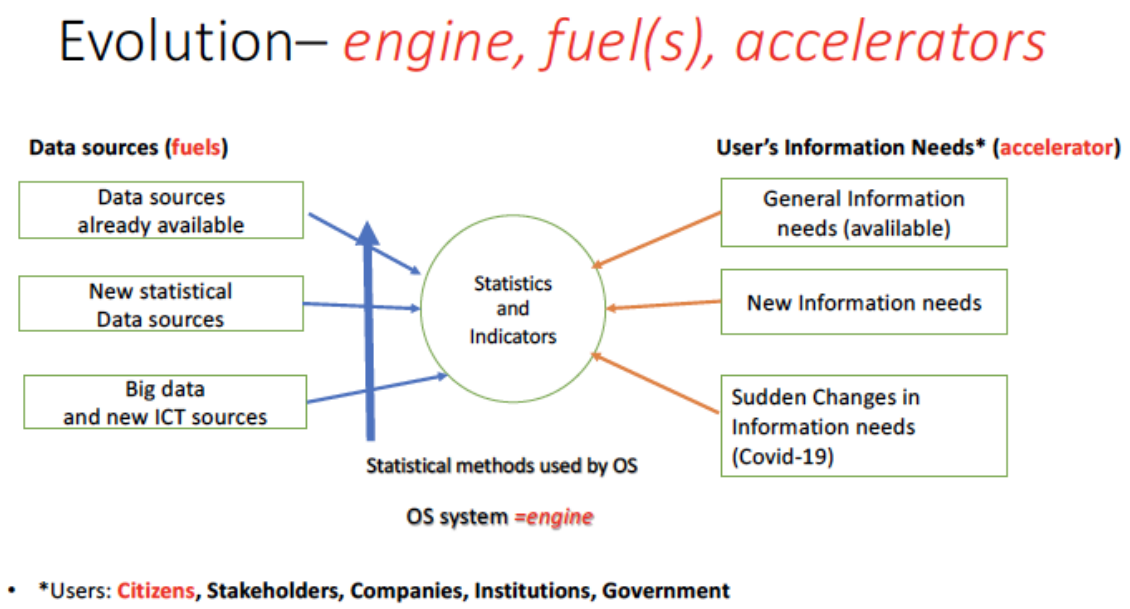


Figure 1: Evolution of the production process of a general OS system

For example, in Italy a sudden change in the information needs happened after the II world war. The government, policy makers and all the stakeholders needed new statistics to reconstruct economic situation (Marshall and Fanfani Plans) and the OS reacted designing and carrying out surveys in different domains, in particular for the construction of the National Accounts, developing new structured and standardized survey methods.

Therefore, the new scenario of data sources outlined in the introduction, moves the evolution mechanism, affecting for example the roles of the various stakeholders and their mutual relationships, to reply also to the questions by the National Recovery and Resilience Plan (NRRP). Summarizing there is a need for timely reactions, both in terms of necessary reorganizations of the NSIs and publications of the first, even provisional results of the data collection process, as Experimental Statistics.

Section 4 – The Citizen Data and the Citizen Science: a challenge for OS

As already said in the introduction, Big Data, smart statistics and citizen are inseparable: from smartphones, meters, fridges and cars to internet platforms, the data of most digital technologies is Citizen Data, that is the data of the citizens and on the citizens.

In addition to raising political and ethical issues of privacy, confidentiality and data protection, the repurposing of Big Data call for rethinking relations between the citizens and the production of

official statistics, if they are to be trusted. The future of Official Statistics does not depend only on the possibility to use new sources of data or new methods, but also on the possibilities that the new digital technologies offer to establish new relationships with the citizens. Their role is destined to evolve from that of respondents to that of collaborators and co-producers of official statistics data (Ruppert, E., et al., 2018; Ruppert, E., 2019).

First, the possibility to exploit Citizen-generated data (CGD) - that are produced by non-state actors, particularly individual or civil society organizations - for official statistical purposes seems to represent a very promising avenue to collect timely and relevant new data. Privacy issues prevent citizens to fully disclosure this kind of data, while their management and storage by privately owned digital platforms generate some remarkable concerns by citizens themselves on their correct protection. In order to fully exploit this kind of data, NSIs need to develop a better understanding on the way they are generated and how can be made accessible for official purposes (Casarez-Crageda et al 2020).

The second approach, that aims at the direct collaboration of the citizens in producing OS, following the principles of the Citizen Science (CS) involves citizens along all the phases of the so-called data value chain: planning, collection, processing, analysis and use (Nascimento et al 2018). This is an important involvement that we can also link to the Post Normal Science approach (Pratesi, 2020).

The general opportunity for OS resides in gaining a new awareness of citizens in their participation to the process of official data production. Rethinking citizen involvement along the phases of the data value chain can help counter the trust deficit between citizens and governments and consequently establish a participatory data ecosystem (Misra and Schmidt, 2020)

The use phase in the data value chain requires an uptake stage that involves three activities: connecting data to users; incentivizing users to incorporate data into the decision-making process; and influencing them to value data. The active involvement of citizens in the data production is a challenge for OS to reduce the gap between users and producers. It would also have a positive feedback on Statistical literacy, as the ability of data users to interpret and critically evaluate statistical information in a variety of contexts.

It is clear, the concept of citizen data and co-production raise practical and political questions that it is impossible to summarized here.

Moreover, CS produces data difficult to compare, the measures of precision are not clear. The challenge for OS resides in the rethinking the data collection process and of the concept of quality of the data. Traditional aspects inherent to the data production process and that are typical when NSIs conceive and govern it such as accuracy, timeliness, representativeness, completeness, etc.

should be rethought and enriched introducing also other aspects. These last are important when the NSIs are not in the position to interfere in every aspect of the production process of the data as: evaluation of self-selection bias, quality checks post-production, evaluation of potential use of the data. Issues as comparability across domains, coherence and benchmarking will be even more important than in traditional data production settings. Even if there are proposals to define the quality profile of citizen-generated data, there are not yet comprehensive and meaningful empirical studies.

To fill this gap, we need research in OS to generate many Experimental Statistics, producing results also by unusual tools such as inference from nonprobability samples, data integration and data fusion of new and traditional data, model based and model assisted estimation methods.

This is true for all the thematic areas where OS is called to produce data: from economic life, as consumption expenditure, earning and usage of disposable income to the aspects of daily life, like access to public services, life-long education, participation in social and cultural life.

Section 5 – CGD key characteristics as compared to other data sources for OS

Citizen-generated data (CGD) have recently emerged in the larger world of digitally generated data as a unique and specific data source. Given the very preliminary stage in the identification and classification CGD, different concepts and definitions of it are under development by scholars and private and public institutions engaged in the analysis and data exploitation of this new kind of data source. Among the different definitions of CGD currently available, the Partnership in Statistics for Development in the 21st Century (Cázarez-Grageda et al, 2020) - an international center of research and high-level competences on data classification and analysis with a strong focus on OS - has defined CGD as follows. “Citizen-generated data (CGD) are data produced by non-state actors under the active consent of citizens to tackle social issues explicitly”. The PARIS21 also highlights the specific features of CGD:

- they are independently produced by non-state actors, particularly individuals or civil society organizations (CSOs), based upon their specific goals and information needs.
- They require an active engagement by citizens, since they have to give access to those data for statistical purposes.
- As for administrative and Big Data, they are generated for proposes other than official statistics, so they need to be properly organized and tested for quality according to statistical standards.

Therefore, in order to identify a specific data source as CGD, the following features shall be verified:

- Produced directly citizens or intermediated by non-state actors on a voluntary basis to monitor or expand their knowledge on a specific topic of relevant individual or collective interest.
- Given by citizens under consent.
- Used to monitor issues that directly affect citizens.
- Available thought digital devices or platforms with appropriate technologies and data security procedures.

As an example of CGD directly and deliberately generated by citizens, one can consider the increasing amounts of health-related data outside the healthcare official system, either intentionally through the use of health tools including fitness trackers or home monitoring devices, or passively through environmental sensors and online activity (Cook and Raza, 2018). By and large, online data forum and twitter can also be considered as a relevant source of CGD (Harris et al., 2017; Reece and Danforth, 2017), although the intentional commitment of citizens in generating CGD can be questioned. Civil society organizations (CSOs) in general, and international no profit organizations in particular, can promote or directly develop digital applications (apps, online fora, etc) to generate and collect CGD. As an example, one can consider the international project developed by Civicus (2019). Civicus is a global alliance covering 160 countries that promotes the use of new data and develop research on civil society. Civicus launched the DataShift initiative to build the capacity and confidence of civil society organizations to produce and use citizen- generated data. Citizen-generated data is data that people, or their organizations, produce to directly monitor, demand or drive change on issues that affect them. CGD were collected by the Civicus network of CSOs by using a pilot survey to test the ability and usefulness of CGD on SDG16.7.2 indicator data. In addition, Civicus develops at the Monitor, which is an online, real time assessment of civic space and the State of Civil Society Report and represents a unique forum to discuss and listen to diverse civil society actors.

In a more local context, one can mention “Ehilapp!”, an app for smartphone designed and developed by an IT company in Verona with the aim to provide, with a digital device, the same information service offered by the local desk of Caritas and to expand them to a larger target of population at risk of poverty. The user of this app can also decide to share information with its own network of friends and relatives. The CGD generated by this application, could be of relevant interest to assess the spatial patterns and drivers of poverty in that regional area. In eventually using

this kind of data for statistical purposes, both the privacy of personal data and the explicit consensus of citizens should be guaranteed.

CGD present some relevant advantages for NSI over other data sources. As for administrative and Big Data, they avoid the cost of direct reporting (financial costs and statistical burden on respondents). In addition, they present a more informative power for both citizens and local and national governments as compared to other data sources, since they are deliberately generated by citizens and, sometimes, collected by CSO. As far as potential disadvantages are considered, they need to be properly organized for statistical purposes, and in particular their quality shell be carefully assessed. This last issue will be discussed in detail in the next section. For now, some basic implications can be easily derived from these preliminary remarks on the nature and scope of CGD in order to lead NSIs to successfully exploit this specific kind of data for official reporting. In particular, NSI shall:

- Establish an institutional setting in which citizens directly or indirectly (CSOc) give their consensus to NSIs to access GDC based upon NSI high scientific and institutional reputation in accessing personal data for statistical purposes and to protect personal and confidential data.
- Design and implement appropriate IT solutions and methodological tools to access, organize and test the quality of this kind of data as input data for statistical purposes.
- Compare CGD with other official data sources already available to NSI in order to test the overall consistency of this kind of data within and across related statistical domains in order to guarantee the overall consistency of the official statistics production.

Section 6 – New data sources and the assessment of data quality in official statistics

This paragraph briefly summarizes how data quality is assessed by official statisticians, in effect survey methodologists, in connection to different types of data sources used as inputs in the statistical production process; direct reporting, administrative data and new data sources, such as Big Data.

The measurement of data quality in official statistics has its roots in the *Total Survey Error* (TSE) paradigm (Lyberg e altri, 2017). This theoretical framework aims at optimizing surveys by maximizing data quality within budgetary and respondent burden constraints, It is based upon the identification, measurement of the sources of error related to two TSE main components: sampling error and non-sampling error. The identification, measurement and minimization of the determinants of sampling error, such as sampling scheme, sample size and estimator choice, have

been the major sources of concern for official statisticians in a time dominated by direct surveys. The introduction of census-like administrative data sources as input of the statistical production process, has increased the relevance of the sources of non-sampling error, such as frame, measurement and data processing, as the key goals of error assessment and minimization by survey methodologists.

The possibility to detect and measure the different sources of non-sampling errors, in effect coverage, structural bias, selection effects and measurement errors with respect to the target population, crucially relies on the possibility to link, at the individual level, survey or administrative data with statistical registers. Statistical registers are set up and maintained by NSOs for this purpose and include a wide range of statistical units: individuals and families, enterprises (natural persons and corporations), public and not for profit institutions. The relevance of statistical registers for data quality assessment has recently been increased by the adoption of a new business model for statistical data production by NSOs, which considers basic statistical registers and extended statistical registers at the core of this new statistical production system to expand the output as well as to increase the consistency of official statistics within and across different statistical domains. While basic statistical registers include only a limited set of variables, focusing on the full coverage of all resident statistical units in a given country, extended statistical registers are built from basic statistical registers and incorporate an additional set of variables obtained through integration with multiple data sources, thus increasing the possibility to test for data quality of new data sources. In the case of Italy, Istat set up and maintained a wide range of both basic and extended statistical registers, that are increasingly integrated within and across statistical domains. In the business statistics domain, they include the business registers on enterprises, local units and enterprise groups and the extended business registers on enterprise economic accounts (Frame SBS) and the extended business registers on enterprise' local units economic accounts.

New data sources, such as Big Data in general, CGD in particular, have further increased the complexity to test for data quality, given their peculiar characteristics in terms of accessibility, data structure, identification and measurement of statistical units and related variables. Nevertheless, the possibility to integrate these data with basic or extended statistical registers play an essential role in order to design an appropriate empirical setting for data quality assessment and to choose the most appropriate methodological approach to measure data quality. In this respect, five different circumstances can be considered:

1. Individual data from new data sources can be linked to the statistical register held by NSI based on the type of statistical unit (population census and register, the business registers on

enterprises, local units and enterprise groups, the business register on not for profit institutions) and through a common identifier (fiscal code, VAT code).

2. Individual data from new data sources cannot be linked to the statistical register held by NSI based on the type of statistical unit because the common identifier (fiscal code, VAT code) is missing, partial or not correctly codified in CGD.
3. Variables embedded in new data sources are directly comparable with similar variables already collected by NSI throughout direct surveys or administrative data at the individual level or the aggregated level (territorial, industry, etc).
4. Variables embedded in new data sources are correlated with variables already collected by NSI throughout direct surveys or administrative data at the individual level or the aggregated level (territorial, industry, etc).
5. Variables embedded in new data sources are weakly or not correlated with variables already collected by NSI throughout direct surveys or administrative data at the individual level or the aggregated level (territorial, industry, etc).

For options 1,3 and 4 new data sources can be substantially assimilated to traditional ones (survey and administrative data). Therefore, standard methodical approaches to detect coverage errors, structural bias, selection effects, measurement errors with respect to the target population can be adopted (Eurostat, 2007). In contrast, for option 2 the finite target population represented by the statistical register cannot be identified and for option 5 variables from official data sources are not available to test for data quality of new data sources.

The wide class of latent variables models represent a possible methodological approach to cope with this kind of data specification and modeling problems (Vermunt and Magidson, 2003). Latent class (LC) modeling was initially introduced by Lazarsfeld and Henry (1968) as a way of formulating latent attitudinal variables from dichotomous survey items. In contrast to factor analysis, which posits continuous latent variables, LC models assume that the latent variable is categorical, and areas of application are more wide-ranging. This class of model in quite flexible in terms of model specification and estimation, therefore it can be applied to different topics where specific target variables are extremely difficult to be collected, including output quality problems in industry (Gertler, 1988) or data quality issues in OS (De Waal et al., 2019).

As anticipated in the previous section, CGD are a peculiar type of new data sources, very similar in terms of technology platform and data structure to Big Data, that is directly generated by citizens or collected by non-state actors, under the consent of citizens, with the purpose to monitor issues that directly affect them. In particular, CGD are deliberately generated by citizens as an independent data generating process outside an a priori defined survey design and business register framework.

Therefore, their data generating process is very likely to be affected by frame bias with respect to their relevant target population. In effect, more educated, digital-oriented and socially engaged individuals will naturally tend to generate more CGD, while other individuals within the same relevant target population, for instance elderly people, less educated ones, will be less prone or avoid generating CGD. In order to identify the most appropriate empirical setting and methodology to assess their data quality is essential to classify them according to the above mentioned five options and to design a consistent and robust experimental framework as the one described in the following section.

Section 7 – An experimental framework to test the quality of SDG indicators for official reporting based on CGD

An important area, essential for regeneration and government in this difficult moment marked by the pandemic, is that of the Sustainable Development Goals (SDGs). SDGs are objectives included in the government programs of European countries. A recent review highlighted how data collected through CS initiatives can feed an important part of indicators for monitoring the Sustainable Development Goals (Fraisl D. et al, 2020). SDG indicators are identified by the Global indicator framework as essential information to be produced as a matter of priority by NSIs in order to assess and monitor the evolution of countries worldwide for the Sustainable Development Goals and targets of the 2030. The total number of indicators listed in the global indicator framework of SDG indicators is 247. Some countries have also adopted the SDG indicators Framework to monitor sustainable development goals at the regional or local level.

The production of SDG indicators and their regular update is very demanding for (NSIs) that are struggling to balance financial constraints, the fast growing demand of new official statistics across different social, economic and environment domains with an increasing disaffection of respondent units in reporting to NSI, despite their legal obligations. Since the production of SDG indicators may address data collection from specific target populations not always included in standard statistical business registers or consider information very difficult to collect based upon large scale official surveys or administrative data, CGD clearly emerge in this area of statistical production as a unique and very promising solution.

However, the possibility for NSI to successfully use CGD data for official statistical production in general, and for the set up and maintenance of SDG indicators, has to meet the essential condition

that their quality for statistical purposes can be carefully assessed and their potential biases corrected using a consistent methodological and statistical data processing approach.

The specific types of SDG indicators and related CGD and official data considered by Istat are highlighted in Figure 1, while the process that will be adopted by Istat to set up an experimental setting to test for their quality of SDG indicators based on CGD is described in the rest of this section.

Figure 2 - SDG indicators and related CGD and official data sources

Area of reference of SDG indicators	Specific type of SDG indicators to be identified and tested	Possible types of CGD sources and related technological platforms/ personal devices			Official data sources potentially related to CGD
		Web	Corporations'owned apps and digital platforms	CSOs'owned apps and digital platforms	
Goal 1 – Poverty	Resilience of poor people measured as presence /intensity of informal networks and accessibility to local services			X	EuSilc and Household Budget Survey
Goal 4 - Education	Informal education (i.e. cinema, read books, theatre), soft skills.	X	X		Labour Force Survey, educational register and others administrative sources.
Goal 2 – Food security improve nutrition	Food waste and food saving and recycling initiatives carried out by families	X	X	X	Household Budget Survey, Aspects of daily life.

Figure 2 illustrates, for each "new" SDG indicator, which is its specific knowledge goal, which types of CGD can be used to build it, and which sort of variables derived from OS can be used to directly o indirectly benchmark the quality of SDG indicators built from CGD. The possible types of CGD sources and related technological platforms/personal devices where these data are saved and stored can be broadly divided into three groups: Web, personal apps owned by corporations or CSOs. This basic classification mainly reflects different issues connected to the accessibility and personal data treatment of the data, which is generally free in the case of SDG available through the web (excluding private access web fora), while it is managed and protected by platforms and apps owners and developers in the case of apps and other data resident on personal digital devices. In this respect, apps and platforms owned by either private corporations or civil society organizations (CSOs) should be distinguished on the basis of a possibly different commercial versus more cooperative aptitude to share with NSOs their CGD data, given the fact that full protection and authorization by citizens in managing their personal data should always be guaranteed. As far as

SDG Goal 1 is concerned, Istat is already collecting trough official surveys data on severe material deprivation[1]. Additional measures on poverty, connected for instance with factors affecting the resilience of the poor people in facing day by day difficulties, such for instance the presence of a formal or informal network of assistance and help and the possibility to access local support services, can be of relevant interest to expand the measurement of SDG indicators for Goal 1. CGD can be achieved by Istat either by running an additional survey based on a specific digital application or by accessing data already available at the national or local level, as in the example of the app developed for Caritas in Verona described in the previous section. Under the conditions of deliberate consensus by the users of Caritas or Istat digital applications and correct treatment of personal data, CGD could be linked to the register of population and related surveys. In a similar vein, SDG indicators on SDG Goal 4 can be expanded by accessing or collecting CGD on informal education, cultural activities not related to formal education. Specialized online fora can be scanned, commercial apps accessed, or a specific app can be developed by Istat as a follow up of an official survey, to collect relevant information on this topic.

Information on food waste and food saving and recycling initiatives carried out by families are of remarkable interest for SDG Goal 2. These data can be collected by Istat either as a follow up of an official survey on a related issue, such as the household budget survey (HBS), or by accessing the CGD generated by a specialized app promoted by corporations and potentially CSOs initiatives.

The standard process carried out by Istat to set up and maintained new statistics and indicators build from new data sources and certified as OS includes three different stages:

- Scouting of new data sources.
- Experimental phase based on a small scale research oriented feasibility study.
- Industrialization process of new data sources as standard and continuous inputs of statistical data production processes.

The scouting of new data sources is not only limited to spot new data sources opportunity for the compilation of new official statistics and to classify them according to the Eurostat taxonomy of new data sources. It also encompasses the set up of an appropriate technological, institutional and legal setting in the case the accessibility of data is bounded by technological, commercial or legal barriers. As an example, Istat will obtain CGD on the above mentioned issues either by establishing

[1] The indicator of severe material deprivation is given by the percentage of people living in families who experience at least four of the following nine symptoms of distress: 1. Not being able to adequately heat the house; 2. Not being able to sustain an unexpected expense (the amount of which, in a given year, is equal to 1/12 of the value of the poverty threshold recorded in the previous two years). 3. Not being able to afford a protein meal (meat, fish or vegetarian equivalent) at least once every two days.4. Not being able to afford a week's vacation a year away from home. 5. Not being able to afford a color TV.6. Not being able to afford a washing machine.7. Not being able to afford a car. 8. Not being able to afford a phone. 9. Being overdue on paying bills, rent, mortgage or other type of loan.

formal agreements with cooperations or civil society organizations (CSOs) to have access to the data included in their platforms (including formal consensus form citizens to use their personal data) or by accessing CGD freely available on the web or by running pilot surveys with voluntary reply from citizens. Given the early stage of development of platforms and digital devices developed by CSOs in Italy and the time required to set up the appropriate legal and institutional setting, the second (web) and third (direct survey) solutions will be adopted by Istat to carry out the data quality assessment of CGD according to the above described empirical setting and methodological approach.

The experimental phase based on a small scale research oriented feasibility study will be carried out by Istat in cooperation with scholars and external stakeholders of this projects. This phase aims to assess the data quality for statistical purposes and the concrete feasibility of using CGD to build new SDG indicators. This phase will encompass:

- Analysis of CGDs characteristics (target population, data reference period, periodicity, timeliness, availability of sources over time) to set up and maintain the "new" SDG indicator.
- Choice of the most appropriate technological solution to access CGD (web scraping or access to platforms or app servers).
- Choice of the best methodological solution to reshape and codified CGD to make them usable for statistical purposes depending on their specific data sources (including text mining or other data mining methodological solutions if CGD are highly unstructured).
- Matching of CGD with OS data sources based on the classification scheme illustrated in the previous section, in order to define the most appropriate empirical setting to test for the data quality of CGD for OS purposes.
- Analysis of potential sources of bias in the quality of CGD and their measurement by using the most appropriate model specification and estimation approach as illustrated in the previous section.

This final step can be split into sub-phases to better specify the process needed to implement to evaluate the quality of the CGD sources. The first element concerns the evaluation of a possible self selection bias. As described in the previous section, this bias could naturally be inherent in this type of source and strictly depend on the theme and the argument treated from the CGDs chosen. For this purpose, it is necessary to define an experimental design that allows to carry out analyses with the same other confounding factors in order to quantify and evaluate the self selection bias. Istat has the possibility, as shown in Figure 2, to use statistical register in order to evaluate the quality of the CGD indicators. If on the one hand the exhaustiveness of the statistical register can be fully

exploited, on the other it is customary that the indicators necessary for the SDG Goals are not directly identifiable in the available statistical registers. After linkage, using latent model where the variable of interest is latent and is determined by explanatory variables known both in the registers and in the CGD considered, then it is possible to construct a series of indicators to evaluate the self selection bias, the measurement error and consequently the accuracy declined for the subpopulations of interest and "covered" by both the registers and the CGDs.

As previously mentioned, the problem of periodicity, timeliness, availability of sources over time represents important elements that will affect the cost-benefit analysis that must be done to fully evaluate the possible introduction of CGD indicators for the evaluation of the SDG goals.

The phases just described illustrate a repeatable process regardless of the indicator chosen for experimentation. The previous scheme can simply be re-proposed to each of the Goals proposed in Figure 2. Once the inputs and methodologies have been defined, the phases for carrying out the experimentation are essentially the same ones.

This experimental phase is usually concluded with the publication of research reports and papers as well as by the eventual publication by Istat of new SDG indicators as experimental statistics.

The industrialization process of new data sources aims at scaling up the experimental phase to the right organizational, institutional and technological dimensions in order to carry out data collection of CGD and the statistical production of SDG new indicators according to high quality standards and pre-defined data dissemination calendar.

Section 8 – Conclusion

Although the use of CGD for OS is still at a very preliminary development stage in many countries, the design and implementation of a robust and reliable framework to test for the quality of CGD for OS purposes is of foremost relevance. Indeed, the recent experience in the use of Big Data by OS has already shown a project life cycle that has shifted from an enthusiastic approach in the first stage to a more consistent ad wise approach in the recent period that well balance straights and limitations of these data for OS.

In order to successfully consider CGD for OS three different pillars should be set up and integrated. The first pillar concerns the capability to establish an active cooperation between NSIs and citizens, including their intermediate organizations, and units generating CGD. In this respect, the high institutional reputation and scientific independence of NSI, including their mission to produce official statistics of valuable interest for policy makers, business and citizens at the national and local levels, represents a very valuable asset. In addition, citizens and civil society organizations

(CSOs) may be very interested in leveraging the value of their data by using the data processing capability of NSI and their capacity to generate official statistical figures from input raw data, thus increasing their relevance and data quality to support information and decision making at national and local levels. For instance, the United Kingdom Office for National Statistics (ONS) has recently joined a global initiative, along with international organizations and international no profit organizations (CSOs), to exploit CGD to develop SDG indicators.

The second pillar reflects the need to design and implement an experimental setting, in the first stage, an a statistical production framework, in the following stages, that will support NSIs in introducing CGD in OS. The possibility to assess the quality of CGD for statistical purposes represents a key issue of this pillar.

The third pillar concerns the need to respect national regulations, and in particular data privacy rules, in all stages of the project life cycle of CGD for OS.

This paper aimed to design an appropriate framework to systematically and consistently assess the data quality of CGD for OS purposes. The focus on SDG indicators and the performance of pilot surveys linked to official surveys is finalized to empirically test data quality and to provide relevant feedbacks to improve both the theoretical setting and the fine tune of the methodological tools used to assess the quality of CGD.

References

Bartholomew D.J., Knott M. (1999): Latent Variable Models and Factor Analysis. Kendall's Library of Statistics 7. Oxford University Press, New York.

Cázarez-Grageda, K., Schmidt, J., Ranjan, R. (2020) Reusing Citizen-Generated Data For Official Reporting A quality framework for national statistical office-civil society organisation engagement PARIS21 Working Paper

Civicus (2019) CITIZEN GENERATED DATA (CGD) FOR SDG, Inclusive & Responsive Decision-Making, report available on line, at the following web site: https://civicus.org/thedatashift/documents/sdg16-report_25-april-2019.pdf

Clogg C.C. (1981): New Developments in Latent Structure Analysis. In D.J. Jackson and E.F. Borgotta (eds.), Factor Analysis and Mea

Cook, Sarah and Raza, Sobia (2018) What is citizen generated data?, Briefing of the phgfoundation.org, University of Cambridge.

De Menezes L.M. (1999): On Fitting Latent Class Models for Binary Data: the Estimation of Standard Errors. British Journal of Mathematical and Statistical Psychology, 52, 149-168.

De Waal, Ton, van Delden, Arnout, and Scholtus, Sander (2019) 'Quality Measures for Multisource Statistics' Statistical Journal of the IAOS: vol. 35, no. 2, pp. 179-192.

Di Zio M., Guarnera U., Rocci R. (2007): A Mixture of Mixture Models for a Classification Problem: The Unity Measure Error. Computational Statistics & Data Analysis, 51, 2573-2585.

Eurostat. 2007. Handbook on Data Quality Assessment Methods and Tools, eds. Manfred Ehling and Thomas Korner, publication available for download at the Eurostat website.

Formann A.K. (1992): Linear Logistic Latent Class Analysis for Polytomous Data. Journal of the American Statistical Association, 87, 476-486.

Fraisl, D., Campbell, J., See, L., When, U., Wardlaw, J., Gold, M., Moorthy, I., Arias, R., Piera, J., Oliver, J.L., Masò, J., Penker, M., Fritz, S. (2020), Mapping citizen science contributions to the UN sustainable development goals, Sustainable Science, 15, pp. 1735-1751

Fortini M., Liseo B., Nuccitelli A., Scanu M. (2001). "On Bayesian record linkage". Research in Official Statistics, 4, 185-198. Pubblicato anche in E. George (editore), Monographs of Official Statistics, Bayesian Methods, EUROSTAT, 155-164.

Gertler, P. (1998) A Latent-Variable Model of Quality Determination, Journal of Business & Economic Statistics Vol. 6, No. 1 (Jan., 1988), pp. 97-104 (8 pages) Published By: Taylor & Francis, Ltd.

Harris J K, Hawkins J B, Nguyen L et al. (2017) Using Twitter to Identify and Respond to Food Poisoning: The Food Safety STL Project. Journal of Public Health Management and Practice. 23(6): 577–580. 3.

Larsen M.D., Rubin D.B. (2001). "Iterative automated record linkage using mixture models". Journal of the American Statistical Association, 96, 32-41.

Lazarsfeld, P.F., and Henry, N.W. (1968). Latent Structure Analysis. Boston: Houghton Mill.

Lyberg, L. E., & Stukel, D. M. (2017). The roots and evolution of the total survey error concept. in Total Survey Error in Practice, 1-22.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinizivillo, S., Pappalardo, L., and Gabrielli, L. (2015), Small area model-based estimation using Big Data sources, in Journal of Official Statistics, 31, pp. 263-281

Marchetti, S., Giusti C., Pratesi M (2016), The use of Twitter data to improve small area estimates of households' share of food consumption expenditure in Italy, in AStA Wirtschafts- und Sozialstatistisches Archiv: 57 10 (2-3) 60 61 July 2016

Meng X.L., Rubin D.B. (1993). "Maximum likelihood via the ECM algorithm: a general framework". Biometrika, 80, 267-278

Misra, A. and Schmidt, J. (2020), Enhancing trust in data – participatory data ecosystems for the post-COVID society, in Shaping The Covid-19 Recovery: Ideas From Oecd's Generation Y And Z © OECD 2020

Nascimento, S., Iglesias, J.M.R., Owen, R., Schade, S., Shanley, L. (2018), Citizen Science for policy formulation and implementation, chapter 16 in: Hecher, S,. Haklay, M., Bowser, A., Makuch, Z., Vogel, J., Bonn, A. (2018), Citizen Science: innovation in Open Science, Society and Policy, UCI Press London

Pratesi, M., (2017), Big Data: the point of view of a Statistician, Etica e Economia, 12/4

Pratesi, M. (2018), Statistica: linguaggio sovradisciplinare per comprendere e dare valore ai dati talk in the Conference on "Data to Change" held on January 15, 2018 at the Italian House of Representatives, in Statistica&Società, 2018

Pratesi, M. (2020), Parlare chiaro: statistica, dati e modelli, talk in "Parlare chiaro, i rischi della confusione dei numeri", online workshop, 30 aprile 2020, Università Politecnica delle Marche

Pratesi, M., (2021), Official Statistics and Citizen Science, Seminar held March, 18, 2021, http://www.centrodagum.it/en/seminario-scuola-dei-dottorati-delle-scienze-sociali-universita-di-firenze/

Pratesi M, Ceccarelli C, Menghinello S. (2021), Citizen generated data and Official Statistics: an application to SDGs indicators, Discussion paper n 274, Department of Economics and Management, University of Pisa.

Radermacher W. (2019), Governing-by-the-numbers/Statistical governance: Reflections on the future of official statistics in a digital and globalized society, Statistical Journal of the IAOS,

Reece A G, Danforth C M. (2017) Instagram photos reveal predictive markers of depression. EPJ Data Science. 2017; 6(1).

Ricciato, F., Wirthmann, A., Giannakouris, K., Reis, F., Skaliotis, M. (2019), Trusted smart statistics: Motivations and principles, Statistical Journal of the IAOS, 35, pp.589-603

Ruppert, E., Grommé, F., Ustek-Spilda, F., Cakici, B. (2018), Citizen Data and Trust in Official Statistics, Economie et Statistique/Economics and Statistics, N° 505-506, pp179-193

Ruppert E. (2019), Different data futures: An experiment in citizen data, Statistical Journal of the IAOS, 35, pp. 633-641

Trivellato U. (1990): Modelli di Comportamento e Problemi di Misura nelle Scienze Sociali: Alcune Riflessioni. In Società Italiana di Statistica, Atti della XXXV Riunione Scientifica, 1, Cedam, Padova, 11-34.

Vermunt J.K., Magidson J. (2003): Latent Class Models for Classification. Computational Statistics & Data Analysis, 41, 531 – 537.

Discussion Papers
Collana del Dipartimento di Economia e Management, Università di Pisa

Email della redazione: lfanti@ec.unipi.it