Crowd4SDG · unitar United Nations Institute for Training and Research · United Nations DESA Statistics · Convention on Biological Diversity · UN WOMEN · World Health Organization · International Institute for Applied Systems Analysis IIASA · Global Partnership for Sustainable Development Data · CITIZEN SCIENCE CENTER ZURICH

Policy Brief

Citizen science data to track SDG progress: Low-hanging fruit for Governments and National Statistical Offices

July 2022

Summary:

Citizen science data (CSD) presents untapped opportunities to track progress towards the SDGs offering multiple advantages for Governments and National Statistical Offices (NSOs): help fill in data gaps for indicators where traditional data collection instruments, such as, household surveys or administrative registers may not be well suited or may not have the sufficient coverage, improve the granularity of data and sometimes timeliness, and offer an effective and cost-efficient means for NSOs to address data needs of policy-makers in the context of limited or falling resources. Many citizen science data initiatives and datasets exist already. Why then have NSOs not been using citizen science data to any large extent to date? To begin with, NSOs require a better understanding of how such data can be useful in their specific context and requirements, as well as understanding what exactly needs to be done to leverage it. Citizen science practitionners, for their part, need to better understand policy applications and know how to ensure that the generated data meets necessary quality standards. This Policy Brief aims to bring greater understanding to these issues by drawing on a research conducted as part of the EU-funded Crowd4SDG project with contributions from several international organizations and research centers. It shows how NSOs can give value to CSD benefitting official statistics, policy-making and citizen science communities and helping ensure that no one is left behind.

Introduction

In this Policy Brief, Citizen Science Data (CSD) is defined broadly as **data produced -by citizens who voluntarily contribute their time, knowledge, skills and/or their data** to help produce evidence, strengthen accountability, or develop locally rooted solutions. This can range from the participation of citizens in data collection or data analysis/classification to defining data needs and relevant indicators (e.g., indigenous people), sharing their personal data (i.e. health and lifestyle) or donating their devices' computer power for modelling and simulations. Citizen science data can be generated under different institutional set-ups, from projects run by academic, research institutions or NSOs to those run by local civil society organizations, large international NGOs or by communities. There are several closely related terms, such as, citizen-generated data, crowdsourcing, community-based monitoring, among others. For consistency, we use citizen science as an umbrella term that covers all these terms and diverse activities The definition provided here was the one used in the report produced as part of an EU-funded Crowd4SDG project with contributions from several international organizations (IOs) and research centres. That report provides key inputs for the recommendations included in this Policy Brief together with studies and experiences of other contributing partners.

Over the past five years, perceptions of National Statistical Offices and the broader official statistics community towards, what have been coined "new" or "non-traditional" data sources, have considerably evolved. The sheer scope of SDG monitoring, including disaggregation requirements, has placed unprecedented pressures on NSOs – even in developed countries with more resources, stronger statistical capacities and administrative systems – to produce data relevant for policymakers addressing the 2030 Agenda in a comprehensive manner. A study by Fraisl, D. (Fraisl, 2020), demonstrated that CSD has the potential to contribute to the production of data that would help to monitor 76 global SDG indicators, and has already contributed to five of them. At national and local levels, more indicators can benefit from citizen science data. A number of citizen science datasets already exist but are not yet known to official data producers. Many citizen science initiatives operate effectively at local level, with potential to be scaled up to produce data relevant for national-level monitoring. CSD can also be helpful in producing disaggregated data, incl. sex disaggregated data for some of the indicators typically sourced from household surveys, which are sometimes difficult to disaggregate when household heads are the sole respondents. However, the full potential of CSD is not exploited yet.

As NSOs gain greater knowledge and experience with using non-traditional data sources, it is likely the message that these data can be extremely useful is reinforced, specifically in providing more timely statistics, potentially with higher levels of granularity, and at a lower cost. Partnerships with institutions that can generate data from non-traditional sources have already proven successful in many instances and have helped to address capacity and funding constraints, demonstrating the power of vibrant data ecosystems. CSD has been one of them.

A 2021 survey conducted as part of the Crowd4SDG project (Crowd4SDG, 2021) has shown that only around 10% of respondents from NSOs and NSSs have been involved at some point in citizen science projects. Equally, qualitative responses from those who have had no CSD exposure show that there is a lack of awareness and understanding of what CSD means and how it can be used by NSOs.

This Policy Brief aims to raise awareness about the opportunities offered by CSD to NSOs and introduces a number of tools that can be used to leverage it.

Challenges of Citizen Science Data Use and How to Overcome Them: Evidence and Analysis

The 2021 Crowd4SDG survey and interviews with national data producers provided some interesting evidence of both the challenges faced by NSOs in introducing and using CSD and how they can be overcome.

More specifically, the survey found that the key obstacles identified by NSOs who have run citizen science data projects were:

- Limited access to data (67%);
- Legal issues with access to or use of data (60%);

- Incoherent use or lack of use of statistical concepts (53%);
- Selection bias (53%);
- Lack of information about how the data are being produced (53%).

Interestingly, those who had no experience emphasized a lack of awareness and methodological guidance, but also identified other obstacles not highlighted by experienced users, such as lack of human capacities, technical and financial support, and technological limitations.

A number of case studies exemplify the use of CSD by NSOs and IOs in practice. The examples detailed below deal with marine litter and sentiment analyses, but other examples include gender-based violence, vulnerability and disaster risk, and biodiversity. Additional examples offered by academic institutions can also be of interest to NSOs, such as the Crowd4SDG study on compliance with social distancing measures during the first year of the COVID-19.

Case study 1. Marine litter in UK (ONS UK) and Ghana (GSS and IIASA) – Goal 14

Ghana use case is an initiative where already existing citizen science data are incorporated into formal SDG monitoring and reporting processes for indicator 14.1.1.b on plastic debris density. The main achievement of the project is that Ghana became the first country to report on the SDG indicator 14.1.1.b using CSD. Additionally, these data will serve as inputs to Ghana's Ocean Plan, currently under development. The key to success was building a data partnership and a trusted environment around CSD, and creating time and space for the government, incl. Ghana Statistical Service, international organizations, Civil Society Organizations (CSOs) and volunteer groups to meet and work together towards common goals ensures ownership over the results.

Source: IIASA.

UK Office for National Statistics (ONS) is also looking into the use of the citizen science data from a charity (Marine Conservation Society) for compiling the same indicator. The data source is currently undergoing the quality evaluation under the protocol developed by UK ONS SDG data sourcing team. This collaboration has been prompted following research by the SDG data team on possible data availability from non-traditional data sources. In addition to the protocol approach and the open SDG reporting platform, another important factor of success is the sustained and targeted work of the SDG data team to identify possible data sources with advice from experts across ONS who can advise on new data sources.

Source: Interview with UK ONS.

Case study 2. New measurements: Sentiment analysis Mexico

One of the experimental projects Mexican NSO, INEGI, has been running since 2014 on the sentiment analysis involved the use of social media data and volunteers' help for classification. The social media component was managed by INEGI while citizen science collaboration was brokered through a partnership with the Universidad Tec Milenio. The sentiment data are available every day at the national level and state levels. They are published as experimental statistics which are not considered official statistics by INEGI. Students are asked to label manually tweets provided by INEGI as negative or positive. The manual classification provides basis for building large training datasets with more than 50'000 values. The machine learning algorithm developed with the support of data scientists from two research centers, INFOTEC and CentroGeo, is then used to label millions of tweets automatically. The geo-referencing allows producing sentiment analysis index not only at the national level but also at the level of states. Correlations are clearly observed for some events such as the 2017 earthquake with more of negative tweets on average while Christmas appears to be the happiest day of the year.

Source: Interview with INEGI.

Case study 3. Using air quality data produced with support of volunteers in the Netherlands

Scientists from the National Institute for Public Health and the Environment (RIVM) in the Netherlands have been leveraging publicly accessible air quality datasets produced by Sensor.Community. The government platform Measure Together uses this air quality data collected by volunteers from Sensor.Community using simple sensors. The platform is also open to other citizen scientists using similar sensor kits. These data are integrated with official sensor data and provide a geographically more detailed information on the levels of fine particulate matter such as PM2.5 and PM10. While the data from volunteers' sensors is of lower quality, it can be adjusted based on the official data and has the advantage of better geographic coverage around the country providing a useful source of information during calamities and for local authorities' decision-making. Its use by the official monitoring body increases its credibility and enables local authorities to benefit from it. The number of such sensors has more than doubled from January 2020 to March 2022 and reached around 2'500 sensors in the country of which the large majority are connected to Sensor.Community. Sensor.Community is a bottom-up citizen science initiative, and their infrastructure supports local citizen science projects in the field of environmental measurements, such as air quality and noise. It provides citizen scientists with manuals on how to build a sensor kit and brings together volunteers from over 70 countries. It can support the monitoring on SDG indicator 11.6.2 at the local level.

Source: Interview with RIVM.

Citizen science data can be used to produce indicators which cannot be compiled otherwise, used alone or combined with other data sources. Some NSOs publish indicators directly derived from CSD and labelled as "non-official" – either alone or along indicators from official producers. Others process CSD jointly or without other data sources applying standard quality assurance procedures and eventually publishing the resulting indicator as official statistics.

To address concerns regarding data quality, such as correct application of statistical concepts, awareness of possible selection bias, and poor documentation of data production protocols when statistics and indicators are published as non-official or experimental statistics, existing good practices can be applied and further developed. These can be built upon practices developed in a number of countries, such as NSOs from UK, Kenya, and Colombia, and include the development of a **Quality Assurance Framework for data collected and processed by non-official producers**. It is worth noting, however, that not all quality issues can easily be addressed through QAF. In some cases, when CSD needs to be further integrated or benchmarked with another data source, significant technical and financial resource may be needed. This is because each type of CSD would require a new set of benchmarking tools or other data sources.

However, even high-quality data are of little use if people (specifically NSO) are not aware of their existence or cannot access them. The concept of accessible data implies the existence of a data ecosystem that enables sharing and ensures a sustainable management of existing data. The first step can be an exhaustive mapping of all CSD that can be helpful for monitoring SDGs and progress towards national development priorities as was done by NSO Philippines. Having a good snapshot of CSD opportunities alone is not sufficient. More systemic changes are required to create incentives for brokering effective data collaborations between NSOs and CSD producers. Sustainability of both production and access to data is critical if NSOs are going to invest time and resources in the use of non-traditional data sources.

At the same time, legal issues may require the modernization of statistical legislation to empower NSOs to play a more central coordination role in the National Statistical System, and ensure access to administrative and other new types of non-traditional data (see MacFeely, S. & N. Barnat, 2017). Colombia's recent law has enabled the NSO – DANE – to play this role.

Finally, to ensure CSD is widely used, it is important to communicate - rather than simply disseminate the data - along with its characteristics, advantages, and limitations. On the one hand, CSD producers should create transparency around the data journey to address upfront possible concerns over accuracy or impartiality and may need to have a communication strategy to ensure key stakeholders are aware about their CSD and the data are widely used. On the other hand, the differences between CSD and other data sources (including on timeliness, production process and any possible limitations) should be openly communicated by both, CSD producers and NSOs, along with the data itself.

Policy Implications and Recommendations

Producing quality assurance guidelines for data from non-official sources in general or for citizen science data specifically is one of the key tools that can be used by NSOs to enable CSD producers to review their approaches and make their data more fit-for-purpose for SDG monitoring and reporting.

A number of criteria have been proposed based on the National Quality Assurance Framework (NQAF) promoted by UN Statistical Commission and used by many NSOs as a model in developing their national frameworks as can be seen, for example, in a recent PARIS21 guide (PARIS 21, 2022). Those often include:

- Accessibility;
- Timeliness, frequency, and sustainability;
- Accuracy and reliability;

- Coverage;
- Relevance;
- Metadata;
- Coherence, Comparability and Integrability.

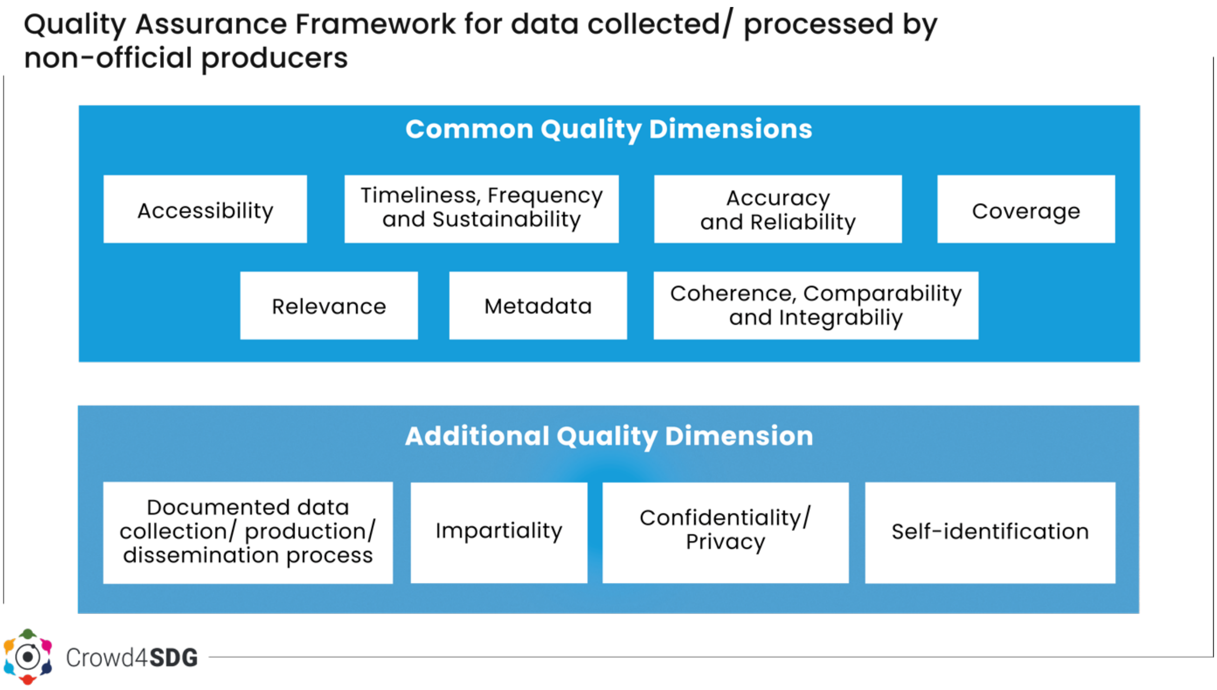
A more comprehensive approach may be required to account for other aspects that are critical for NSOs. The Crowd4SDG report features a mapping of both NQAF criteria and UN Fundamental Principles of Official Statistics (UNFPOS), and has resulted in the addition of the following criteria:

- Impartiality;
- Confidentiality/Privacy.

Furthermore, two additional criteria have been proposed; one based on the UK's experience to address a common challenge, and the other to reflect an important principle from a human rights-based approach to data, which is not included in the UNFPOS:

- Documented data collection;
- Self-identification.

Figure 1. Proposed QAF criteria to be promulgated by NSOs for citizen science data


Quality Assurance Framework for data collected/ processed by non-official producers
Common Quality Dimensions
Accessibility
Timeliness, Frequency and Sustainability
Accuracy and Reliability
Coverage
Relevance
Metadata
Coherence, Comparability and Integrabiliy
Additional Quality Dimension
Documented data collection/ production/ dissemination process
Impartiality
Confidentiality/ Privacy
Self-identification
Crowd4SDG

Source: Based on a diagram from Crowd4SDG report, 2021.

One emerging trend among NSOs is to issue QAF guidelines to improve the quality of CSD for the use in SDG monitoring and reporting. This approach can work where there is a vibrant citizen science community, such as active CSOs and academia who are engaged in CSD projects relevant to NSO operations and are willing to adopt these standards in order to make their data useful to NSOs. This can be complemented by establishing protocols to include non-traditional sources of data that meet the criteria for SDG reporting, as was done by the ONS in the UK, and the mapping of available data to see how it can inform or complement SDG monitoring by helping to close gaps or be used as complementary sources. This approach can be termed as passive or CSD Light, meaning NSOs have no control over the data production process but seize opportunities where the resultant data or statistics meet published quality standards.

In-Focus. Examples of approaches to QAF to leverage data collected and processed by non-official producers for monitoring the SDGs

UK's Office for National Statistics has an open SDG National Reporting Platform with most of the data coming from official data sources. To address gaps on some of the indicators, the Office for National Statistics SDG team was open to leveraging new data sources and decided to put in place a protocol for non-official data specifically for monitoring the SDGs. The protocol consists of an ethical gateway and a scoring matrix. The Ethical Gateway serves as a pass/fail mechanism with three criteria all of which have to be met on Ethics and Privacy, Transparency and Accountability, and Need. Once this condition is met, a scoring matrix is applied to evaluate the given data set using additional criteria: Relevance, Methods, Coverage, Timeliness, and Data Quality. The average score is then calculated and 1.5 points are used as a threshold to decide on whether or not the source could be used for the SDG platform. The non-official data protocol is aligned with the UK Statistics Authority Code of Practice and its voluntary application procedure applied to non-official data sources. Data from non-official sources will be labelled as such and made available in addition to the official source data where available.

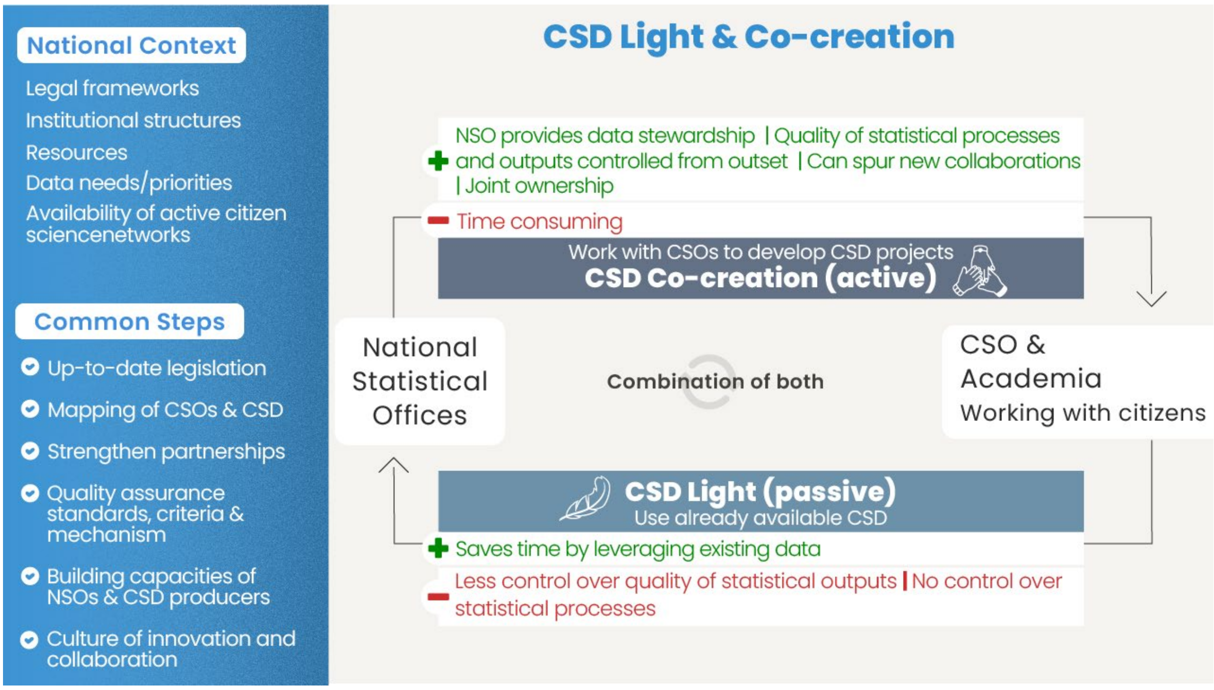
Colombia's NSO used a different approach. It does not have a quality assurance mechanism for data processed by non-official producers per se but has developed quality assurance guidelines for experimental statistics by adapting its quality assurance framework for official statistics. The experimental statistics workstream was launched in 2020 as a flagship initiative guided by the dedicated Technical Committee chaired by DANE's Director and Chief Statistician in Colombia and composed of all technical directors and some advisors. At the moment all data created by DANE are official statistics. Experimental statistics are considered official statistics in Colombia according to Decree 2404 from 2019. They offer new ways of quantifying phenomena relevant for sustainable development and can be helpful in ensuring disaggregated information, address data gaps, combine traditional data sources such as censuses and surveys with new data sources, develop in-house capacities to run experimental projects. Experimental data need to meet a defined set of quality assurance criteria such as, Relevance, Accessibility, Interpretability, Transparency, Coherence and Timeliness.

Source: UK ONS, DANE.

The passive approach may not be sufficient for cases where no robust citizen science data work has been done. NSOs wishing to engage in this context will need to take a more proactive stance by initiating dialogue with potential stakeholders from CSOs and academia, discussing areas where data needs to be produced and opportunities for collaboration on a joint project. This is an **active approach** or **co-creation** as an NSO can provide stewardship on data quality from the outset and have more control over the data production process. For example, in the case of the Litter Intelligence program led by Sustainable Coastlines in New Zealand, citizen science practitioners have collaborated with the Stats New Zealand and the Ministry for the Environment since the design phase of the program to ensure that the data produced by citizens on marine litter meet their quality standards for use in official environmental monitoring, including the SDGs. It can help scale up local citizen science initiatives centred around local concerns and priorities but with strong potential for being useful at national-level reporting. Other advantages of co-creation include its participatory nature. It builds ownership by both the NSO and the citizens and can strengthen the entire data value chain beyond just production.

In many instances, NSOs may need to use a mix of active and passive approaches taking into account their specific situation (i.e., legal frameworks and institutional structures, resources, availability of active citizen science networks and data needs/priorities). Measures may range from making the NSS 'fit-for-purpose' from a legal perspective and helping create a more enabling political and institutional environment, mapping indicators already produced by citizen science /key stakeholders / data ecosystems (supply side) as well as critical NSO data gaps and needs (demand side). For some of the indicators the approach may be passive and for others it may be active, depending on whether the institutional environment allows for both.

Figure 2. Active (Co-creation) and passive (CSD Light) approaches by NSOs to citizen science data


National Context
Legal frameworks
Institutional structures
Resources
Data needs/priorities
Availability of active citizen sciencenetworks
Common Steps
Up-to-date legislation
Mapping of CSOs & CSD
Strengthen partnerships
Quality assurance standards, criteria & mechanism
Building capacities of NSOs & CSD producers
Culture of innovation and collaboration
CSD Light & Co-creation
NSO provides data stewardship | Quality of statistical processes and outputs controlled from outset | Can spur new collaborations | Joint ownership
Time consuming
Work with CSOs to develop CSD projects
CSD Co-creation (active)
National Statistical Offices
Combination of both
CSO & Academia
Working with citizens
CSD Light (passive)
Use already available CSD
Saves time by leveraging existing data
Less control over quality of statistical outputs | No control over statistical processes

Sustainability and Legacy

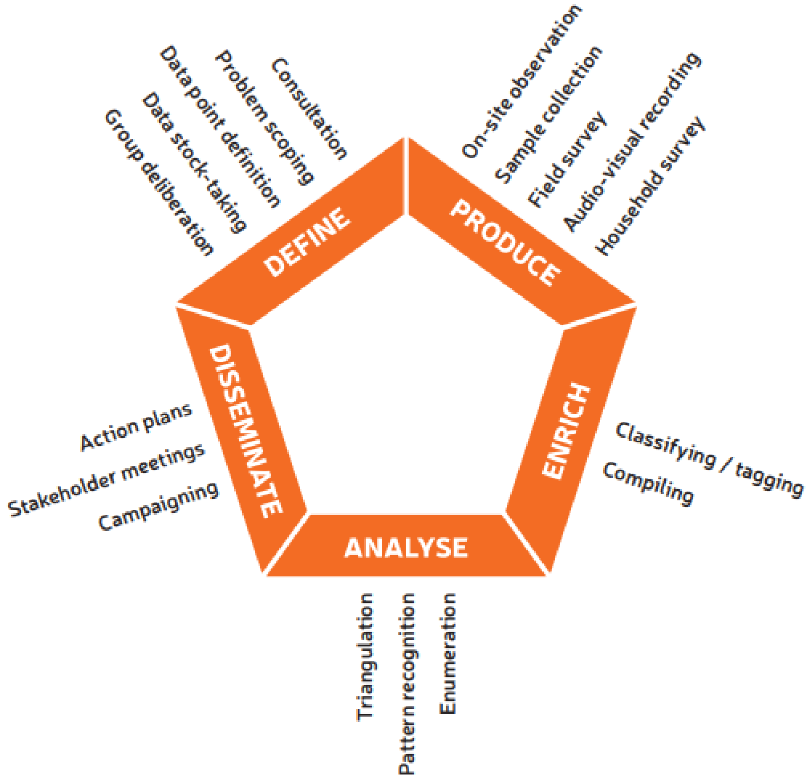
To operationalize and ensure the sustainability of the approaches outlined above, and ensure that NSOs can leverage CSD, the following, more specific, recommendations for NSOs are suggested:

1. Undertake mapping of CSOs and Academia with potential to contribute citizen science data, incl. data they already produce, and their ability/interest to produce new/additional CSD;
2. Update, if necessary, national statistical or other legislation to encourage NSOs to engage with CSOs, Academia and communities;
3. Strengthen partnerships with CSOs, Academia and communities who may potentially contribute to data production. This may include entities that already produce data (CSD Light or passive approach) and where the establishment of an MoU could be considered, similarly to those signed with providers of administrative entities. Partnerships can be also beneficial with those with whom NSOs could engage with on

collaborative projects. Working with data user communities, such as Ministries, local authorities and some other stakeholders, can help identify organizations already producing data and also those who may be interested in collaborating on new projects (relevant for point 1);

4. Define clearly the purposes to which NSO plans to use CSD recommended by PARIS 21 (PARIS 21, 2021). This will also be relevant for collaborative projects and would typically be about producing given SDG or other indicators, disaggregated data, etc. The diagram developed by GPSDD (Figure 3) showing the different roles citizens can play will be extremely useful;

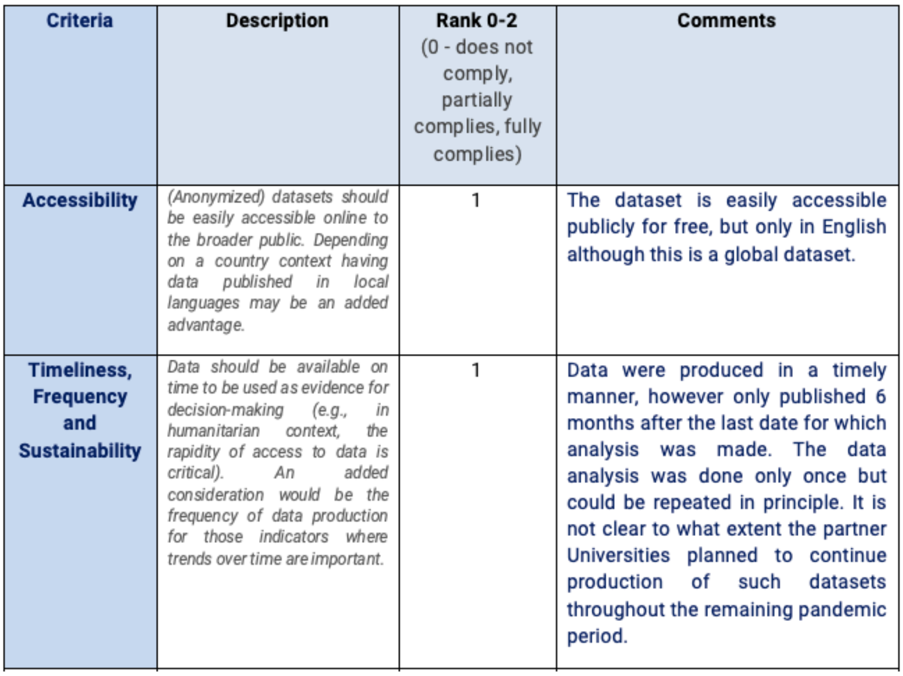
Figure 3. Illustration of tasks underpinning CGD (CSD) initiatives



Source: GPSDD, 2018.

5. Define quality standards and criteria for data quality and management;
6. Introduce quality assurance mechanisms. This can be a scoring matrix with threshold values (Figure 4);
7. Promote a culture of innovation and collaboration with new data actors. Leadership can play a decisive role on this matter;
8. Provide training and capacity development for stakeholders involved in CSD production to enhance statistical literacy (for CSOs), improve knowledge of the principles of official statistics (also important for Academia) and the awareness about the needs and the work of NSOs.

Figure 4. Snapshot of the upper part of a scoring matrix from Crowd4SDG report



Criteria	Description	Rank 0-2 (0 - does not comply, partially complies, fully complies)	Comments
Accessibility	*(Anonymized) datasets should be easily accessible online to the broader public. Depending on a country context having data published in local languages may be an added advantage.*	1	The dataset is easily accessible publicly for free, but only in English although this is a global dataset.
Timeliness, Frequency and Sustainability	*Data should be available on time to be used as evidence for decision-making (e.g., in humanitarian context, the rapidity of access to data is critical). An added consideration would be the frequency of data production for those indicators where trends over time are important.*	1	Data were produced in a timely manner, however only published 6 months after the last date for which analysis was made. The data analysis was done only once but could be repeated in principle. It is not clear to what extent the partner Universities planned to continue production of such datasets throughout the remaining pandemic period.

Source: Crowd4SDG, 2021.

Finally, the international official statistics community could play an important role. In particular where benchmarking is required, a set of benchmarking tools or related best practice put at the disposal of NSOs could enable them to wrestle with more difficult quality issues and reduce financial cost and time investment NSOs would have had to incur had they have to deal with this on their own. Work has already been initiated in this area by the global statistical community led by UN Statistics Division.

References

Crowd4SDG Deliverable 5.1 “Initial report on relevance and quality-related considerations of citizen-science generated data”, June 2021.

Crowd4SDG Deliverable 5.2 “Data usability assessment and recommendations for SDGs for GEAR cycle 1”, June 2021.

DANE. Estadísticas experimentales, 2022.

Data for Change. Strengthening measurement of marine litter in Ghana, 2021.

Fraisl, D., et al., Mapping citizen science contributions to the UN sustainable development goals, 2020.

General Assembly Resolution 68/261. UN Fundamental Principles of Official Statistics. 29 January 2014.

GPSDD, Choosing and engaging with Citizen-Generated Data. Guide, GPSDD, Open Knowledge International and Public Data Lab, 2018.

INEGI Estado de ánimo de los tuiteros en México, 2022.

MacFeely, S. & N. Barnat (2017). Statistical Capacity Building for Sustainable Development: Developing the fundamental pillars necessary for modern national statistical systems. *Statistical Journal of the International Association of Official Statistics*, Vol.33, No. 4, pp. 895 – 909.

Negri V., Scuratti D., Agresti S., Rooein D., Scalia G., Fernandez Marquez J.L., Ravi Shankar A., Carman M. and Pernici B. (2021). Image-based Social Sensing: Combining AI and the Crowd to Mine Policy-Adherence Indicators from Twitter, ICSE - Track Software Engineering in Society, May 2021.

PARIS 21, Reusing citizen-generated data for official reporting. A quality framework for national statistical office-civil society organisation engagement, PARIS 21, February 2021.

PARIS 21, PSA & PSRTI, Use of CGD for SDG reporting in the Philippines: A case study. 2020.
Rijksinstituut voor Volksgezondheid en Milieu, Measure together, Sensor.Community, accessed on 6 April 2022.
UK ONS & Marine Conservation Society. Charity for compiling the indicator on ocean litter pollution: SDG indicator 14.1.1 part (b) on plastic debris density.
UK Sustainable Development Goals: use of non-official sources, 2022.
United Nations. A Human Rights-Based Approach to Data. Leaving No One Behind in the 2030 Agenda for Sustainable Development. OHCHR Guidance Note to Data Collection and Disaggregation, 2018.
United Nations. Handbook on Management and Organization of National Statistical Systems, 2021.
UN Statistics Division. UN National Quality Assurance Framework Manual, 2019.

Contributing authors

Elena Proden (UNITAR, coordinator), Karen Bett (GPSDD), Haoyi Chen (UNSD), Sara Duerto Valero (UN Women), Dilek Fraisl (IIASA), Gabriel Gamez (UNSD), Stephen MacFeely (WHO), Rosy Mondardini (CSCZ), Linda See (IIASA), and Yongyi Min (UNSD).