Crowd4SDG
Citizen Science for the Sustainable Development Goals

Deliverable 5.2

Data usability assessment and recommendations for SDGs GEAR cycle 1

Deliverable identifier: D5.2

Due date: 30/04/2021

Justification for delay: due to exceptional activity as agreed with the Project Officer

Document release date: 24/06/2021

Nature: Report

Dissemination Level: Public

Work Package: 5

Lead Beneficiary: UNITAR

Contributing Beneficiaries: UNIGE, CSIC, POLIMI, UP

Document status: Final

Abstract:

Deliverable 5.2 assesses datasets provided by other Crowd4SDG Partners and generated with support of citizen science using the criteria framework developed in Deliverable 5.1 and reflecting the perspective of National Statistical Offices on the potential of use on non-official data for SDG reporting and the required quality of data for this purpose. One of the assessed datasets was developed as part of the Crowd4SDG implementation. Two other datasets were produced earlier.

For more information on Crowd4SDG, please check: http://www.crowd4sdg.eu/

Document history

	Name	Partner	Date
Authored by	Elena Proden	UNITAR	20/04/2021
Reviewed by	Barbara Pernici, Romain Muller	POLIMI, CERN	26/04/2021
Edited by	Elena Proden	UNITAR	14/06/2021
Approved by	François Grey, Jose Luis Fernandez-Marquez	UNIGE	24/06/2021

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

Table of Contents

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

Project Partners

	Partner name	Acronym	Country
1 (COO)	Université de Genève	UNIGE	CH
2	European Organization for Nuclear Research	CERN	CH
3	Agencia Estatal Consejo Superior de Investigaciones Científicas	CSIC	ES
4	Politecnico di Milano	POLIMI	IT
5	United Nations Institute for Training and Research	UNITAR	CH
6	Université de Paris	UP	FR

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

Crowd4SDG in Brief

The 17 Sustainable Development Goals (SDGs), launched by the UN in 2015, are underpinned by over 160 concrete targets and over 230 measurable indicators. Some of these indicators initially had no established measurement methodology. For others, many countries do not have the data collection capacity. Measuring progress towards the SDGs is thus a challenge for most national statistical offices.

The goal of the Crowd4SDG project is to research the extent to which Citizen Science (CS) can provide an essential source of non-traditional data for tracking progress towards the SDGs, as well as the ability of CS to generate social innovations that enable such progress. Based on shared expertise in crowdsourcing for disaster response, the transdisciplinary Crowd4SDG consortium of six partners is focusing on SDG 13, Climate Action, to explore new ways of applying CS for monitoring the impacts of extreme climate events and strengthening the resilience of communities to climate related disasters.

To achieve this goal, Crowd4SDG is initiating research on the applications of artificial intelligence and machine learning to enhance CS and explore the use of social media and other non-traditional data sources for more effective monitoring of SDGs by citizens. Crowd4SDG is using direct channels through consortium partner UNITAR to provide National Statistical Offices (NSOs) with recommendations on best practices for generating and exploiting CS data for tracking the SDGs.

To this end, Crowd4SDG rigorously assesses the quality of the scientific knowledge and usefulness of practical innovations occurring when teams develop new CS projects focusing on climate action. This occurs through three annual challenge-based innovation events, involving online and in-person coaching. A wide range of stakeholders, from the UN, governments, the private sector, NGOs, academia, innovation incubators and maker spaces are involved in advising the project and exploiting the scientific knowledge and technical innovations that it generates.

Crowd4SDG has six work packages. Besides Project Management (UNIGE) and Dissemination & Outreach (CERN), the project features work packages on: Enhancing CS Tools (CSIC, POLIMI) with AI and social media analysis features, to improve data quality and deliberation processes in CS; New Metrics for CS (UP), to track and improve innovation in CS project coaching events; Impact Assessment of CS (UNITAR) with a focus on the requirements of NSOs as end-users of CS data for SDG monitoring. At the core of the project is Project Deployment (UNIGE) based on a novel innovation cycle called GEAR (Gather, Evaluate, Accelerate, Refine), which runs once a year.

The GEAR cycles involve online selection and coaching of citizen-generated ideas for climate action, using the UNIGE Open Seventeen Challenge (O17). The most promising projects are accelerated during a two-week in-person Challenge-Based Innovation (CBI) course. Top projects receive further support at annual SDG conferences hosted at partner sites. GEAR cycles focus on specific aspects of Climate Action connected with other SDGs like Gender Equality.

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

Grant Agreement description of the deliverable

The Deliverable 5.2 "Data usability assessment and recommendations for SDGs GEAR cycle 1" is produced under Task 5.2: Assessment of CS project data from the perspective of official data producers and users. This task is led by UNITAR with contributions from UNIGE, POLIMI, CSIC, and UP.

This task focuses on the evaluation of the data generated by the selected citizen science projects created during the three cycles of the GEAR methodology. For each GEAR cycle, the generated data will be assessed using the established criteria and quality requirements from Task 5.1 regarding its potential use as a source of information for decision-making and/or as a secondary source to help inform the production of official statistics. This task complements T2.3 which validates CS data quality in terms of accuracy and precision independent of users' needs.

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

1. Purpose and scope of the deliverable

The Work Package 5 (WP5) aims to explore the perspectives of National Statistical Offices (NSOs), National Statistical Systems (NSSs) and national Governments on the potential of citizen science data for monitoring progress on SDGs. One of the key elements it aims to examine is related to the expectations of NSOs and NSSs to the quality of data that may be generated through citizen science projects and contributions.

One of the key deliverables under this working package are the data usability assessments of the data to be generated through the GEAR (Gather, Evaluate, Accelerate and Refine) cycles[1] that are to be compiled in annual data usability assessment reports.

This deliverable is linked to a task that focuses on the evaluation of the data generated by the selected citizen science projects created during the three cycles of the GEAR methodology. Since this deliverable is being produced at the end of the first GEAR cycle, we expect a time lag between the creation of prototype CS projects and their deployment as fully functional projects that generate data sets. Therefore, and as planned for in the DoA, UNIGE and POLIMI have provided three data sets described below. The first dataset on compliance with COVID-19 related social distancing measures was entirely produced in the context of the Crowd4SDG project in 2020 to test the VisualCit pipeline and was reported under Deliverable 2.1[2]. The other 2 datasets had been generated previously, and one of them – on the Albanian earthquake that had been generated under the EU-funded E2mC project that used Crowd4EMS platform[3] - was subsequently processed by CSIC under the Crowd4SDG project as described in Deliverable 2.1.

The report presents the results of the assessment of the first three data sets. The objective of this exercise is two-fold:

1. Enable dataset providers and other member of the Crowd4SDG Consortium to see if their datasets can be useful to NSOs and under what conditions/what requirements should be met, and to integrate lessons learnt into guidance that is being provided to innovation teams as part of the Crowd4SDG GEAR cycles 2 and 3 so their datasets could have potential to inform SDG monitoring;
2. Test the new criteria framework developed from the NSOs perspective under Deliverable 5.1 and briefly described in the section below.

---

[1] Please read more about GEAR cyle in the following Crowd4SDG Deliverable 3.3 "GEAR report cycle 1".
[2] Crowd4SDG Deliverable 2.1"CS tools design and early prototype available".
[3] Please read more about Crowd4EMS in A. Ravi Shankar, al. Crowd4EMS: a crowdsourcing platform for gathering and geolocating social media content in disaster response, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XLII-3/W8, 2019.

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

2. Methodology

For the usability data assessment, UNITAR has used the criteria framework developed under Deliverable 5.1 of the Crowd4SDG project. The criteria were developed drawing on a survey and a series of interviews conducted primarily with National Statistical Offices (NSOs) as entities in charge of coordination of data production and quality assurance for monitoring the SDGs at the national level. They reflect the expectations of NSOs and, more broadly, National Statistical Systems (NSSs) towards the quality of data that could be leveraged for monitoring progress on the SDGs. The scope of quality as understood by NSOs goes beyond the scientific accuracy and reflects other dimensions that make data relevant and easily usable by policy-makers. The perspectives of NSOs on quality are guided by requirements related to the UN Fundamental Principles of Official Statistics, and National Quality Assurance Frameworks for Official Data, and several NSOs used those as the point of departure to develop specific guidelines for non-official data sources. Deliverable 5.1 provides an in-depth analysis of this work and proposes a set of recommendations for National Statistical Offices and International Organizations, as well as a generic set of criteria that could be used by NSOs to assess the quality and usability of data for monitoring SDGs or incorporating non-traditional data in data production and reporting as a secondary data source.

This criteria framework is now used in Deliverable 5.2 for the first time to assess the provided data sets generated with inputs from citizen science.

Please see the diagram below for the key criteria identified for inclusion in the criteria framework for non-official data sources for monitoring SDGs.
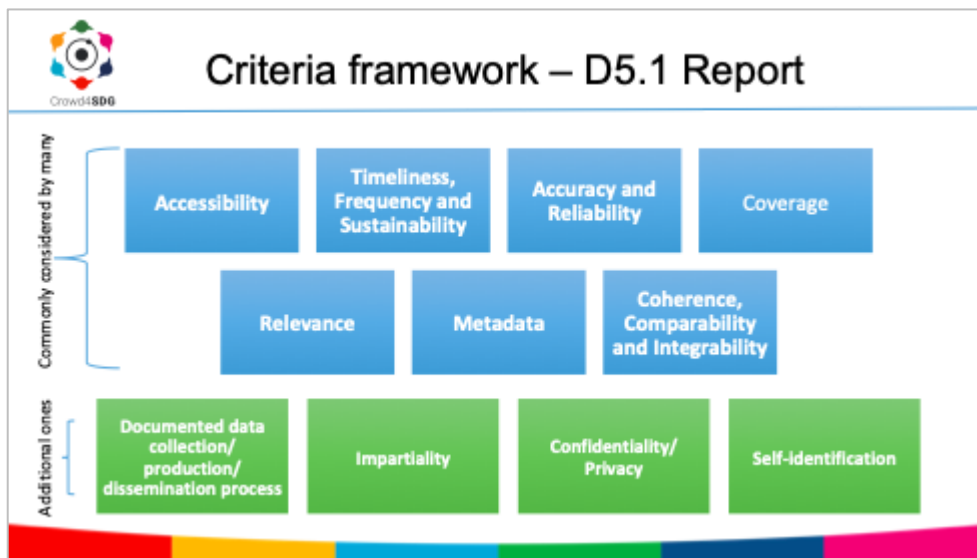

Figure 1
Source: Crowd4SDG Deliverable 5.1.

In addition, the Institute has contacted several experts – one from an international organization and two from National Statistical Offices - to obtain their feedback and assessment of the provided data sets. Their qualitative inputs were used to inform the assessment exercise against the provided criteria. The resulting assessment scoring and descriptions are a combination of these experts' feedback and additional analysis by UNITAR's team. They are provided in the subsequent section separately for each dataset using a table showing key criteria (see Figure 1), their descriptions, score, and comments. The scoring approach was developed based on the experience of UK's Office for National Statistics (ONS)

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

with 3 levels of compliance: zero compliance (0), partially compliant (1), and fully compliant (2). This work in turn was in turn inspired at the beginning by the quality assurance mechanism that was being developed by Statistics Denmark. However, for the purpose of the CrowdSDG assessments and given the difference in the combination of criteria, the scoring was adapted: datasets assessed as equal to or more than 1.5 on average are considered validated (in the UK's approach it is above 1.5).

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

3. Assessment results

This section is dedicated to the analysis of the three datasets provided by Crowd4SDG partners, UNIGE and POLIMI, using the criteria framework for non-official data developed in Deliverable 5.1. The analysis is a combination of the feedback provided by external experts and UNITAR team's additional analysis.

3.1 Crowd4SDG-VisualCit COVID-19 behavioural indicators

The first data set is called "Crowd4SDG-VisualCit COVID-19 behavioural indicators" and was generated as part of research undertaken by researchers from the University of Geneva and Politecnico di Milano (POLIMI) in 2020 under Crowd4SDG project. To produce this dataset, social media data had been analysed by the researchers on compliance with masks wearing and other social distancing measures using Twitter images sampled during 3 periods: 11-17 May, 27 July-2 August, and 17-23 August 2020. Images obtained through Twitter crawls with COVID-19 keywords were filtered with machine learning (ML) classifiers to identify those images that were photos with people in public places. Citizen scientists contributed to the classification of the pictures from the Twitter using a number of questions such as "Are people in the image wearing masks?", "What type of masks are they wearing?", "Is there more than 1 person in the image?", "Are they respecting social distancing?", "Are people in the public place?" and "What type of public place are they in?". The dataset covers 23 countries in different world continents. The analysis focused on the percentages of people wearing masks and maintaining social distances.

A link to the dataset and to the related research article was provided to UNITAR by University of Geneva, both published in open access online at https://zenodo.org/record/4539697#.YC-46JNKhE4 and respectively.

This dataset and the article were reviewed by UK ONS using their own scoring methodology as a test and both have met the identified criteria and could have in principle been featured as non-official data on the UK ONS SDG portal, had they been directly related to UK SDG monitoring. The below assessment table for this first data set was completed using qualitative elements from the UK ONS analysis complemented by UNITAR's analysis in some of the same criteria and in those criteria that do not feature in the UK framework.

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	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

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

	those indicators where trends over time are important.		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.
Accuracy and Reliability	Data should be produced using sound statistical procedures and methods.	2	Brief description on the main page (https://zenodo.org/record/4539697#.YHRLBHySlPZ), but there is a link to the paper which gives more detail including validation and limitations (https://re.public.polimi.it/retrieve/handle/11311/1161146/584481/ICSE_SEIS_Image_based_Social_Sensing%20%2837%29.pdf) Validation through the public and survey comparisons is statistically assessed and explained in the paper.
Coverage	Incomplete coverage may be a serious obstacle to an effective use of data. When no proper sampling techniques are applied, it may be difficult possible to establish to what extent incomplete data coverage are representative of the population or a given group. For national indicators, complete coverage for the country's population or territory is often a must. In some cases, incomplete coverage may be addressed through standard estimation techniques from household surveys.	1	The data set is global. Not all countries are covered. No discussion of sample limitations, however.
Relevance	It is important that the collected data are relevant to decision-makers to make progress on national development objectives and / or inform specific public policies. For NSOs using or wanting to use citizen data for SDG monitoring, nationally relevant/adapted SDG indicators are an important benchmark.	1	While the data is not relevant for a specific SDG global indicator, it is relevant for informing policy decisions that will affect the achievement of the SDGs, mor specifically to support policymakers and epidemiologists to understand the impact of social distance measures and regulations focusing on social distancing and the use of masks in public spaces.

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

Metadata	The provision of a proper metadata and its accessibility together with the dataset is a key condition. Without metadata, NSOs cannot judge about the compliance of the dataset with many other criteria such as coverage, relevance, accuracy and reliability, coherence, comparability and integrability. The access to metadata is also essential for data users so that the data can be used effectively and appropriately.	2	The published dataset is accompanied by a published metadata document.
Coherence, Comparability and Integrability	The coherent use of standard statistical concepts and methods enables the comparability of data across regions, over time and allows its aggregation and use in combination with other data sources. In some cases, proxy indicators can prove helpful when data for the indicator itself is not available, however NSOs can promote the coherence and comparability of the produced data through guidelines, publication indicators with their metadata and training for CSOs.	1	The questions raised and concepts used are relatively new and are therefore not part of the standard statistical concepts. However, the validation has been used against other surveys and sources using similar concepts confirming their relevance.
Documented data collection/production/dissemination process	Metadata should provide full information not only on the data set but also on the process of data collection. This is an important indicator in UK's protocol called data journey awareness.	2	The collection, processing and manipulation of the data is sufficiently described in the main paper. The data is publicly available.
Impartiality	The Organization supplying the data should be able to demonstrate that it is committed to impartiality in the data production process. While this may be less of an issue for Universities and scientific community, it may be less obvious for NGOs that combine advocacy, service provision and monitoring mandates. It is important they are committed to training personnel involved in data collection and production on statistical procedures, techniques and principles. The demonstration of impartiality can be addressed through a proper metadata showing the application of sound statistical procedures and fully transparency on the process and outputs, but it is important to ensure metadata reflect the reality.	2	The impartiality is met through the commitment of Universities undertaking research to produce accurate data independent of political influences. The analysis and classification involve the comparison of results of the classification choices by several volunteers to ensure no spoilers can negatively affect the quality of data.

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

Criteria	Description	Rank 0-2	Comments
Confidentiality/ Privacy	For datasets on citizens, the Organization supplying the dataset should demonstrate that the data collection process has involved a full consent from the respondents during the data collection process / no violation of data protection legislation if such exists. As some Organizations may be collecting this data as part of their service delivery, such data may not be anonymized in their internal databases similar to administrative data. It is important however to ensure that the respondents are fully aware and consent to data collection in this way to avoid unethical behavior, and that the published datasets comply with confidentiality/privacy.	2	Data is extracted from Twitter so it is publicly available.
Self-identification	An important principle to consider from the Human Rights based approach to data about individuals is the principle of self-identification allowing the person to define his gender, ethnic, cultural and other identities in accordance with his/per perception.	n/a	This criterion is not applicable here as citizens are not asked questions about themselves.
Average score		**1.5**	

Table 1. Assessment of the dataset "Crowd4SDG-VisualCit COVID-19 behavioural indicators"

This dataset seems to meet many requirements to a certain extent and could qualify with other NSOs. Its methodology therefore could be replicated for some of the SDG indicators where the social media analysis could be of help. This dataset serves as a good example that can be replicated for producing new datasets under Crowd4SDG and the GEAR cycle.

3.2. UK floods

The second data set is called UK floods and includes data related to the flooding that took place in UK in 2014. Two csv files were provided, incl. one with tweet links, volunteers' names/nicknames and geolocation. No metadata or other descriptions about the dataset were available in the package.
UK ONS has indicated that the information about the dataset was minimal so it was impossible to evaluate it and it would not have passed the criteria for publishing as non-official data source. The below table was completed using the elements of some additional feedback received from UN Office for Disaster Risk Reduction (UNDRR) experts and from Mexican NSO expert complemented by UNITAR's analysis.

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.	0	Not available online.
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.	0	It is not clear what was the actual production timeline: whether the dataset with analysis was available in the days that followed the flooding or much later.
Accuracy and Reliability	Data should be produced using sound statistical procedures and methods.	0	It is difficult to say what statistical procedures may have been applied in this analysis. No information on that was provided. It is also not clear what kind of indicators had been calculated based on this dataset. Subsequent explanations from the dataset producers indicate how accuracy is being addressed, however this is explained nowhere in the dataset.
Coverage	Incomplete coverage may be a serious obstacle to an effective use of data. When no proper sampling techniques are applied, it may be difficult possible to establish to what extent incomplete data coverage are representative of the population or a given group. For national indicators, complete coverage for the country's population or territory is often a must. In some cases, incomplete coverage may be addressed through standard estimation techniques from household surveys.	1	The data set covers several areas of UK affected by the flooding. As there is no metadata, it is not clear to what extent it covers all affected areas or not and how representative it may be of damages in various regions.
Relevance	It is important that the collected data are relevant to decision-makers to make progress on national development objectives and / or inform specific public policies. For NSOs using or wanting to use citizen data for SDG monitoring, nationally relevant/adapted SDG indicators are an important benchmark.	1	While the data may potentially be relevant for disaster loss accounting under Sendai Framework and DesInventar Database (particularly on livelihoods and effects on infrastructure/industry), it is not clear what indicators exactly are calculated just by looking at the dataset.

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

Metadata	*The provision of a proper metadata and its accessibility together with the dataset is a key condition. Without metadata, NSOs cannot judge about the compliance of the dataset with many other criteria such as coverage, relevance, accuracy and reliability, coherence, comparability and integrability. The access to metadata is also essential for data users so that the data can be used effectively and appropriately.*	0	There is no metadata accompanying the dataset.
Coherence, Comparability and Integrability	*The coherent use of standard statistical concepts and methods enables the comparability of data across regions, over time and allows its aggregation and use in combination with other data sources. In some cases, proxy indicators can prove helpful when data for the indicator itself is not available, however NSOs can promote the coherence and comparability of the produced data through guidelines, publication indicators with their metadata and training for CSOs.*	0	The concepts are not defined so it is not possible to discuss coherence, comparability or integrability. One the positive side, datasets are made available in CSV.
Documented data collection/production/dissemination process	*Metadata should provide full information not only on the data set but also on the process of data collection. This is an important indicator in UK's protocol called data journey awareness.*	1	The analysis process is not documented at all. The only part of the process known is where the raw data – pictures – are coming from.
Impartiality	*The Organization supplying the data should be able to demonstrate that it is committed to impartiality in the data production process. While this may be less of an issue for Universities and scientific community, it may be less obvious for NGOs that combine advocacy, service provision and monitoring mandates. It is important they are committed to training personnel involved in data collection and production on statistical procedures, techniques and principles. The demonstration of impartiality can be addressed through a proper metadata showing the application of sound statistical procedures and fully transparency on the process and outputs, but it is important to ensure metadata reflect the reality.*	1	Due to the lack of any description on indicators or analysis, it is impossible to conclude with certainty about impartiality. It is assumed however that through the commitment of universities undertaking research to produce accurate data independent of political influences, impartiality may not be at stake.

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

Confidentiality/ Privacy	*For datasets on citizens, the Organization supplying the dataset should demonstrate that the data collection process has involved a full consent from the respondents during the data collection process / no violation of data protection legislation if such exists. As some Organizations may be collecting this data as part of their service delivery, such data may not be anonymized in their internal databases similar to administrative data. It is important however to ensure that the respondents are fully aware and consent to data collection in this way to avoid unethical behavior, and that the published datasets comply with confidentiality/privacy.*	0	Data is extracted from Twitter, so it is publicly available. There are names of persons, however, in one of the excel sheets. These are likely citizens acting as classifiers. It is important to verify whether these are actual names or nicknames and what are the confidentiality procedures applied in such cases. If there was no explicit consent from citizens on having their names or nicknames published, the dataset should undergo anonymization procedure before it is shared/published.
Self-identification	*An important principle to consider from the Human Rights based approach to data about individuals is the principle of self-identification allowing the person to define his gender, ethnic, cultural and other identities in accordance with his/per perception.*	n/a	This criterion is not applicable here as citizens are not asked questions about themselves.
Average score		**0.4**	

Table 2. Assessment of the dataset "UK floods"

This dataset does not meet minimum requirements for being used as a non-official data source. It needs to be complemented with a) thorough metadata, b) the detailed documentation on procedures, c) be anonymized and d) be made available publicly (for free) before it can be considered for being examined as a non-official data source. Those 4 aspects should be considered as important lessons learnt for the datasets to be produced under Crowd4SDG and the GEAR cycle.

3.3. Albania earthquake

The third data set is called Albania earthquake and includes data related to the earthquake that took place in Albania in 2019. The shared folder with files provided by the University of Geneva to UNITAR included: CSV files showing the details of tagging by volunteers and experts, as well as the tweets, geolocation, and ID of image. It has also included links to the libraries of the tagged pictures.

UK ONS has indicated that the information about the dataset was minimal so it was impossible to evaluate it and it would not have passed the criteria for publishing as a non-official data source due to that. The below assessment table was completed using the elements of some additional feedback received from UN Disaster Risk Reduction experts and from Mexican NSO expert complemented by UNITAR's analysis.

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.*	0	Not available online.
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	It is not clear what was the actual production timeline, but one could assume that the event took place on 27 November 2019 while the date of the analysis entered in the dataset was 6 December 2019. This seems to meet the timeliness requirement; however, the dataset was not made available publicly to be used on time. It needs to be clarified if it was shared with potential users. Such data do not need to meet the frequency requirement as it is linked to the specific event.
Accuracy and Reliability	*Data should be produced using sound statistical procedures and methods.*	1	The dataset shows that there was a validation procedure with each tweet being analyzed by at least 5 volunteers and cross-checked by experts. It is however difficult to say what other statistical procedures may have been applied in this analysis to ensure the accuracy of data. Subsequent explanations by POLIMI discussed in conclusions potentially shed light on some of them but have not been included in this dataset.
Coverage	*Incomplete coverage may be a serious obstacle to an effective use of data. When no proper sampling techniques are applied, it may be difficult possible to establish to what extent incomplete data coverage are representative of the population or a given group. For national indicators,*	1	The data set covers several areas of Albania affected by the earthquake. It is not clear however to what extent it covers all affected areas or not and how

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

	complete coverage for the country's population or territory is often a must. In some cases, incomplete coverage may be addressed through standard estimation techniques from household surveys.		representative it may be of damages in various regions.
Relevance	*It is important that the collected data are relevant to decision-makers to make progress on national development objectives and / or inform specific public policies. For NSOs using or wanting to use citizen data for SDG monitoring, nationally relevant/adapted SDG indicators are an important benchmark.*	1	The data may be relevant for disaster loss accounting under Sendai Framework and DesInventar database, but it is difficult to say to what extent it can inform specific indicators (please see the point on Coherence, Comparability and Integrability).
Metadata	*The provision of a proper metadata and its accessibility together with the dataset is a key condition. Without metadata, NSOs cannot judge about the compliance of the dataset with many other criteria such as coverage, relevance, accuracy and reliability, coherence, comparability and integrability. The access to metadata is also essential for data users so that the data can be used effectively and appropriately.*	0	There is no metadata accompanying the dataset.
Coherence, Comparability and Integrability	*The coherent use of standard statistical concepts and methods enables the comparability of data across regions, over time and allows its aggregation and use in combination with other data sources. In some cases, proxy indicators can prove helpful when data for the indicator itself is not available, however NSOs can promote the coherence and comparability of the produced data through guidelines, publication indicators with their metadata and training for CSOs.*	0	The concepts are not clearly defined in more specific terms (what is meant by severe damage, moderate damage, no damage) although examples of the consensus pictures classified as one of the 3 or irrelevant are provided (it is not clear whether volunteers were given more specific instructions or just examples of pictures). It is therefore difficult to discuss coherence, comparability or integrability. Should the descriptions be provided, it might be possible to compare to the concepts used in the UNDRR damage loss accounting. On the integrability, a positive side of the dataset is that it was shared in the CVS format.

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

Documented data collection/production/dissemination process	*Metadata should provide full information not only on the data set but also on the process of data collection. This is an important indicator in UK's protocol called data journey awareness.*	1	The analysis process is documented but not fully. The data collection process is documented to a certain extent (sources of pictures), but how these pictures were selected and the coverage aspects, for example, are not addressed.
Impartiality	*The Organization supplying the data should be able to demonstrate that it is committed to impartiality in the data production process. While this may be less of an issue for Universities and scientific community, it may be less obvious for NGOs that combine advocacy, service provision and monitoring mandates. It is important they are committed to training personnel involved in data collection and production on statistical procedures, techniques, and principles. The demonstration of impartiality can be addressed through a proper metadata showing the application of sound statistical procedures and fully transparency on the process and outputs, but it is important to ensure metadata reflect the reality.*	1	Due to the lack of sufficient description on the applied procedures, it is impossible to conclude with certainty about impartiality. It is assumed however that through the commitment of universities undertaking research to produce accurate data independent of political influences, impartiality may not be at stake.
Confidentiality/ Privacy	*For datasets on citizens, the Organization supplying the dataset should demonstrate that the data collection process has involved a full consent from the respondents during the data collection process / no violation of data protection legislation if such exists. As some Organizations may be collecting this data as part of their service delivery, such data may not be anonymized in their internal databases similar to administrative data. It is important however to ensure that the respondents are fully aware and consent to data collection in this way to avoid unethical behavior, and that the published datasets comply with confidentiality/privacy.*	2	Data is extracted from Twitter, so it is publicly available. The names of the volunteers are replaced by IDs and are therefore assumed to be anonymized.
Self-identification	*An important principle to consider from the Human Rights based approach to data about individuals is the principle of self-identification allowing the person to define his gender, ethnic, cultural, and other identities in accordance with his/per perception.*	n/a	This criterion is not applicable here as citizens are not asked questions about themselves.
Average score		**0.8**	

Table 3. Assessment of the dataset "Albania Earthquake"

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

This dataset as was shared with evaluators does not meet the overall minimum requirements for being used as a non-official data source, however, it meets more of criteria than the UK floods dataset and has been an improvement on this 2014 data. The Albania earthquake dataset needs to be complemented with a) detailed metadata, b) a much more detailed documentation on procedures and c) be made publicly available before it can be considered for being examined as a non-official data source. These are important lessons learnt for the future datasets to be generated by the Crowd4SDG project, incl. under the GEAR cycle. UNIGE has confirmed that the Work Package 2 of the Crowd4SDG has processed the dataset and is in the process of documenting and publishing it soon in open access.

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

4. Conclusions, recommendations, and outlook

Out of the 3 datasets shared with UNITAR by two other Crowd4SDG partners, one – the Covid-19 dataset dedicated to the social distancing measures and generated in the context of the Crowd4SDG project - would meet at present the criteria defined in the Deliverable 5.1 related to the potential use of the dataset by NSO as a non-official data source.

One of the key drawbacks that needs to be addressed in the 2 other datasets related to UK floods and Albanian earthquake is about the provision of a detailed metadata that also describes the process and statistical procedures. Indicators that these datasets aim to inform need to be clearly identified and summary information presented. The Covid-19 dataset followed the guidelines on FAIR data management https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf and was published on ZENODO. This was not the case for the two other datasets. The UK flood dataset was not published, and as a result was not documented to make it public. Under Work Package 2 of the Crowd4SDG project, the Albanian earthquake dataset had been processed and UNIGE is in the process of preparing a detailed metadata and publishing the dataset on ZENODO.

Furthermore, the datasets and the metadata have to be made publicly available for free as soon as they become available to meet the quality criteria defined in Deliverable 5.1 from the perspective of NSOs. These criteria define the data fit-for-purpose to be used by national or local authorities and other stakeholders as timely evidence for informing their decision-making processes.

For these last two datasets on natural disasters, UNIGE and POLIMI have provided some additional information related to how their accuracy was being evaluated. Two possible sources of bias were identified through preceding work: the coverage of social media used and the choice of search keywords. The first potential bias can be addressed through estimation. The second was addressed by focusing more on visual evidence in the social media posts rather than analyzing the text. The images were analyzed either using well tested models for which evaluations can be found or through building new models for which the researchers can have control in assessing possible biases. All these explanations need to be reflected in the accompanying metadata and process description when the datasets are being published/made available.

Another positive finding related to all three datasets is the relevance of the data that have become available (Covid-19 dataset) or could have been potentially available to policy-makers (DRR datasets). The two datasets on earthquake and flooding share some common characteristics and could be particularly helpful in disaster loss accounting. UNDRR has developed a database titled DesInventar available at: https://www.desinventar.net/ The citizen science data could be used in combination with other data sources to inform some of the variables in this database. To do so, however, it would be important to examine the definitions provided in the DesInventar database for various indicators and ensure that the definitions applied throughout the citizen science study are compatible with them and could facilitate an easy integration with other data sources. In some cases, citizen science data would need to be combined thereafter with other data sources to inform the database when it is possible to geolocate the dwelling with precision. For example, several variables are related to the effects of a disaster on livelihoods, more specifically, dwelling. The DesInventar database questions would include specific questions on households living in affected dwellings: how many people lived in the household, were there people with disabilities, etc. Such information on the household cannot be obtained through social media analysis but could be available to NSO/NSS through administrative data source or through census.

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

Overall, the main advantage of the citizen science methodology used in producing the three datasets is that it allows producing close-to-real-time data for important societal indicators leveraging social media data, volunteers' contributions, and machine learning to help process huge volumes of data quickly. The research institutions leading this work develop and use open-source tools to enable volunteers, including those with no data science background, to easily contribute to data analysis. Another advantage is that data on locations can be aggregated using the hierarchical structure of OpenStreetMap.

For the subsequent data usability assessment reports, two recommendations could be made: to analyze the datasets that could come out of the GEAR cycles 2 and 3, and to analyse some additional datasets that could be generated as a result of citizen-science pilots with NSOs if such opportunities come up.

The criteria and assessment methodology would need to be further refined for the next cycle report to reflect the specificity of the new topics that may be covered, incl. gender and climate change that may come out of the GEAR cycle 2.

Finally, the lessons learnt from this study should be used for producing new datasets under Crowd4SDG project and a part of the GEAR cycle.

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

5. References

A. Ravi Shankar, A., al. Crowd4EMS: a crowdsourcing platform for gathering and geolocating social media content in disaster response, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. XLII-3/W8, 2019.

Crowd4SDG Deliverable 2.1. "CS tools design and early prototype available", draft consulted in June 2021.

Crowd4SDG Deliverable 3.3. "GEAR report cycle 1", draft consulted in June 2021.

Crowd4SDG Deliverable 5.1. "The perspectives of the NSOs on the use of citizen science and citizen generated data", submitted in June 2021.

V. Negri, D. Scuratti, S. Agresti, D. Rooein, G. Scalia, J.L. Fernandez Marquez, A. Ravi Shankar, M. Carman and B. Pernici, Image-based Social Sensing: Combining AI and the Crowd to Mine Policy-Adherence Indicators from Twitter, ICSE - Track Software Engineering in Society, May 2021
https://re.public.polimi.it/retrieve/handle/11311/1161146/584481/ICSE_SEIS_Image_based_Social_Sensing%20%2837%29.pdf

Dataset « Crowd4SDG-VisualCit COVID-19 behavioral indicators", Barbara Pernici, https://zenodo.org/record/4539697#.YIrXrLUzaUn

DesInventar database available at: https://www.desinventar.net/

D.5.2 - Data usability assessment and recommendations for SDGs GEAR cycle 1

Annex: list of abbreviations

Abbreviation	Description
CCL	Citizen Cyberlab
CS	Citizen Science
CSSK	Citizen Science Solution Kit
DoA	Description of the Action
GEAR	Gather, Evaluate, Accelerate, Refine
NQAFs	National Quality Assurance Framework for Official Statistics
NSO	National Statistical Office
NSS	National Statistical System
O17	Open Seventeen Challenge (online coaching)
ONS	Office for National Statistics (UK)
SDG	Sustainable Development Goal
UNDRR	UN Office for Disaster Risk Reduction
UNFPOS	UN Fundamental Principles of Official Statistics