Choosing and engaging with citizen generated data

Global Partnership for Sustainable Development Data

OPEN KNOWLEDGE INTERNATIONAL

Public Data Lab

The past decades have seen the rise of many citizen-generated data (CGD) projects. A plethora of concepts and initiatives use CGD to achieve many goals, ranging from citizen science, citizen sensing and environmental monitoring to participatory mapping, community-based monitoring and community policing. In these initiatives citizens may play very different roles - from taking on the role of mere sensors, to giving them agency to shape what data gets collected. Initiatives may differ in respect to the media and technologies used to collect data, the ways stakeholders are engaged with partners from government or business and in terms of how activities are governed to align interests between these parties.

Who is this guide for?

This guide will help you understand if CGD is suitable for your proposed project as well as what type of data is appropriate for your needs. It is designed for governments, international organisations and others interested in developing, engaging with and supporting CGD initiatives. It presents a list of distinction criteria between CGD methods, highlights the benefits and pitfalls of CGD, and provides a basis for strategic engagement with CGD.

The guide draws from an analytical framework presented in the report 'Advancing sustainability together? Citizen-generated data and the Sustainable Development Goals'. The analytical framework revolves around three aspects: workflows to generate data; participation; and data's fitness for purpose. The report illustrates these nuances through several case studies and a discussion of how CGD can support implementation and monitoring of the SDGs.

The following three aspects are key when designing a CGD project:

Fitness for purpose: CGD must be 'good enough' to be useful for a specific task. Governments must first articulate their question, or clearly define a problem area they care about and have a remit to manage. Sometimes, engaging the public is needed to define the question. Once the question and data needs are defined, several ways exist to gauge their fitness for purpose.

Participation: CGD initiatives enrol citizens in various capacities. Participation can vary in breadth (how many different tasks are people enrolled in?) and depth (what kinds of participation are practised?). CGD does not always have to be about maximising the breadth of participation. Rather, governments should ask *what kind of participation* is meaningful and useful for an initiative.

Workflows: Generating data takes many shapes and CGD initiatives are often located within larger workflows. Methods differ by data type collected, protocols used to gather data and technologies involved.

The guide is structured along following questions:

1. What are your objectives, questions and data needs?
2. How can the engagement and participation of people help?
3. What resources are available to support CGD?

4. How can CGD be made public?

5. What considerations are relevant for data protection?

Each section is accompanied by CGD examples from our report ‘Advancing sustainability together? Citizen-generated data and the Sustainable Development Goals’. Our guide summarises experiences from the research feeding into this report. We also draw inspiration from existing toolkits to recommend civic technologies, as well as the many existing toolkits for participatory mapping, citizen sensing, citizen science and other data-related activities. You can find a list of the tools that inspired us at the end of this document.

Step 1: Define the goals and scope of your intervention

What is your priority when engaging people?

Before engaging with CGD, governments must define what their initiative is trying to achieve. Governments engage with CGD initiatives for different purposes. Some seek greater engagement with the public, using CGD as an educational approach or to enable participatory deliberation that will shake up tired institutional planning and pave the way for more inclusive government processes. Other times, governments might commission a community or organisation to crowdsource data in order to equip local decision-makers with data. Table 1 illustrates some of the observed activities of government to engage citizens in.[1]

Educate	Gather baseline data	Help manage services and interventions	Define policy
Engage with communities	Inform research	Monitor performance	Make regulatory decisions

Table 1: Illustrative list of purposes for CGD

Governments may want to engage with CGD to serve multiple aims, such as gathering research data while educating people about the process. The following questions shall help you think through your primary goals engaging with CGD initiatives:

1. Do you primarily need data for a government-internal action (for example, baseline research, planning, management, policy design)?
2. Are you primarily interested to engage people in data-intense projects to increase their technical competences and data literacy?
3. Are you primarily interested to enhance institutional literacy and educate people about government institutions and how they function? For example, do you want to increase critical or creative engagement with official data collection practices?

[1] This list is derived from our case studies, and inspired by existing research on the linkages between public policy and environmental monitoring.

Example 1: The Ministry of Health and Wellness in Botswana has commissioned data collections to OSM and local residents. It had a clearly defined use case for geolocated building data as part of anti-malaria interventions. But not enough detailed data was available. Recruiting and training residents was also seen as opportunity to increase technical literacy.

Example 2: Statistics Canada has tested the idea of crowdsourcing with existing OpenStreetMap communities. The intervention was not so much focused on gathering data, but rather on experimenting how the statistical community needs to design engagement strategies with existing communities.

Example 3: The Australian Bureau of Meteorology has multiplied its interaction channels with key target groups such as farmers by allowing them to submit data on extreme weather. The focus was to design channels that help forecasters understand local weather conditions.

Scoping: What is your question or problem and how could people help further define them?

Beyond engaging people for its own sake, CGD initiatives are often focused on solving a problem or answering a question. Governments and public institutions may start from very different situations. Some might have a well-established data collection process and want to benefit from gathering additional data. Here the data needs are clearly defined, and a limited range of methods may be suitable. For others, the problem area may be clear, but government may not be able to define the root causes or prioritise interventions without public engagement. You may ask yourself:

1. Do you want to consult people to develop an initiative based on their most important problems?
2. Do you have a prioritised problem area? Could early engagement with people help clarify the problem you should pay attention to? (see also step 3 on ways of engaging people)
3. Does early engagement with citizens help you prioritise the problems you should address?
4. Could early engagement with people help increase people's ownership, dedication for, or identification with the project (see also 'What motivates your audience')?
5. Do your priorities and questions resonate with the people you engage? (check Step 3).

Example 1: The United States Environmental Protection Agency already has a well-established environmental monitoring program, but cannot cover all relevant country regions. The agency also informs citizens about the types of actions 'volunteered monitoring' can support.

Example 2: Health facilities and the Ministry of Health in Mozambique are vaguely aware of service delivery issues, but do not know their root causes, which population groups are mostly affected or how performance varies between facilities. Partnership building and deliberative engagement formats such as focus groups helped surface key issues to prioritise for service management.

Data stocktaking: What information are you lacking to address the problem, and how could the public help gathering it?

Governments may have different information to address a problem. Data may already be collected by someone (a government administration, a business sector) but may not be complete, not granular enough, outdated, or otherwise unhelpful for the problem. The types of information collected also dictates governance, access to data and responsible data use.

CGD initiatives can enrich or complement official data, thereby enhancing its use value. When taking stock of data, governments should not only consider their internal use of data, but also how the public can work with it to gather new insights (see step 2 for details how the public can enrich official data). Key questions include:

1. If you have data available, can the data be made public so that people can enhance it further?
2. Do you see a need to cross-verify, complete, or otherwise complement your data with CGD?
3. What types of information do you intend to gather? Does the data collection raise ethical and legal issues (such as privacy concerns) (see more details in step 6)?
4. Can the perspectives of different public communities help detect gaps in your data collections?

Example 1: Statistics Canada gathered municipal housing data, and ingested it into the OpenStreetMap (OSM) database. OSM served as a compiling medium to host data. The data was made accessible to the OSM community and can be used for remote annotation of building footprints, or cross-verified and updated through on-site field surveys.

Example 2: Canada's provincial governments have started uploading water monitoring data onto Atlantic DataStream. This helps digitise existing data, but also gathers all data in one location so that it is accessible to communities.

Example 3: Black Sash gathered public records about service plans and planned performance targets. A first compilation of this data helped designing performance metrics that could be tested in a social audit.

Step 2: Clarify what CGD approach is useful for your purpose

How can citizens generate data?

Citizens can generate data in many ways, not only by producing new data, but also by enriching and analysing existing (official) data. This can include to compile formerly unstructured data on a database, to classify, format, annotate, mediate, translate or otherwise engage with data. Figure 1 illustrates some tasks we identified as part of our report 'Advancing Sustainability Together: Citizen-generated data and the Sustainable Development Goals'.

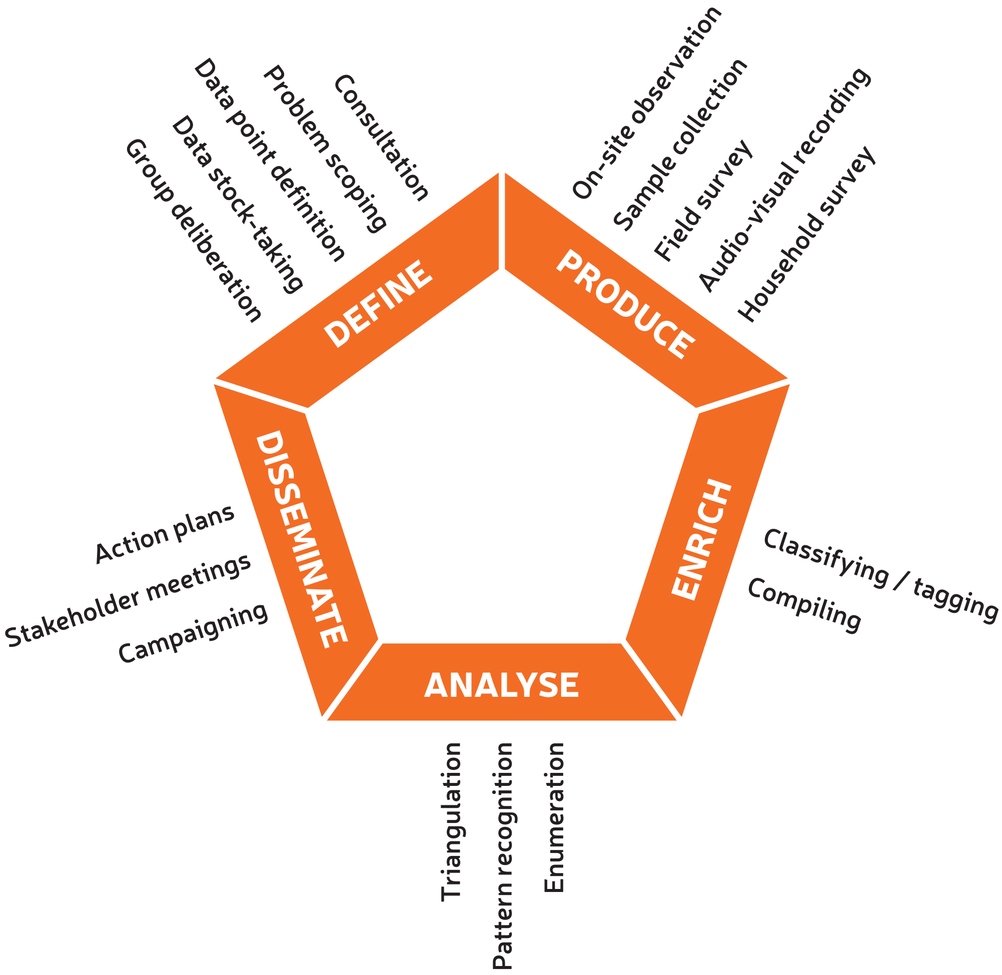

Figure 1: Illustration of tasks underpinning CGD initiatives and their workflows

We suggest to read it not as a linear value chain, or a step-by-step list of tasks to follow. CGD initiatives can start with any of the tasks outlined above, and let other tasks follow as needed. Some tasks are not discrete, either. For instance, some tasks can have the purpose to enrich or to analyse, depending on the question at hand and data involved. Instead, we wish to emphasise that CGD initiatives involve different actions, which can prompt different questions for their design.

How each task plays out **can differ depending on what data shall be collected** and **what instruments and protocols are used**. For instance, on-site observations can be done in many different ways, from randomly spotting wildlife via cameras, to documenting the status of water and sanitation infrastructure in public service facilities following a stricter sampling protocol and instructions. Likewise, tasks such as compiling can involve different infrastructures - from locally stored compilations of government data prior to a project (Black Sash) through to gathering national data collections of weather observations on a database (WOW Australia). Thinking in terms of flows may help understand how CGD initiatives organise and distribute participation among different actors. We may also study **possible dependencies** between these actions **and where CGD depends on other data infrastructures**.

Here we illustrate some of these operations as well as examples from our case studies.[2]

On-site observations

Citizens can visit and describe sites to collect new data or enhance existing information about places, physical infrastructure, environmental conditions, wildlife presence or events occurrence. This approach is useful in the absence of pre-existing data infrastructure to capture on-site data or when validation of existing information is needed. It usually requires the physical presence of people and involves using pre-defined questionnaires or survey tools. Limits are set by the number of citizens available, the accessibility of places and costs incurred such as travel allowances or staff costs needed to complete observations.

Examples from our case studies:

- Social audits (Black Sash)
- Weather condition documentation (Australia's Weather Observation Website)
- Field surveys (Humanitarian OpenStreetMap Team)

Surveys

Some initiatives collaborate with local leaders and trusted community members to interview people on questions as diverse as household welfare, accessibility to public services, or perceived issues with facilities and infrastructure. Depending on the object of study, surveys may not only capture households (to gather representative local household data), but also service users and providers in public facilities (to understand their experiences in public facilities). Surveys may also vary in their sampling approach (capturing data from, and survey design.

Examples from our case studies:

- Social audits (Black Sash)
- Local household surveys (Uganda's Bureau of Statistics)

[2] The list is based on an inductive analysis of 230 CGD cases as well as existing classifications of CGD methods.

Sample collection and measurement

Citizens follow procedures to identify, and collect samples of different objects of study. These may include soil, water, air samples and others. People may want to measure physical properties in their environment they cannot directly observe (e.g. radiation) or cannot otherwise quantify (e.g. temperature or noise).

This approach can be useful to understand health and pollution parameters. Limits are set by the difficulty of the sampling procedure, the accessibility of places, and the quality and status of sampling tools (e.g. through contamination), but also sensor quality and sensor behaviour.

Examples from our case studies:

- Water quality monitoring in Canada (Atlantic Water Network)
- Air quality monitoring in Pristina (Science for Change Movement)

Audio-visual recording

People can accompany observations via audio and video recordings, which can be collected via stationary devices (sensors and cameras), mobile devices (drones) or via people's consumer devices (mobile phones, cameras), either automatically (e.g. taking records in intervals) or manually (when people make an observation). Data can be used for follow-up analysis or other tasks such as 'classifying/tagging'. For instance, some groups have collected higher resolution aerial imagery, ingested it into OSM (compilation), and have annotated the images with digital building footprint data. In other cases, provide context to existing information (such as enhancements of location-based services like surveys, on-site descriptions and others).

Examples from our case studies:

- Weather monitoring stations installed by farmers (feeding into Australia's Weather Observation Website)

Group deliberation

Group deliberation can be useful to scope CGD projects, to collectively define data models to collect, but also to produce data directly. Approaches such as community scorecards organise group deliberation by facilitating focus group meetings with different groups of people (usually split by sex, age and other relevant criteria). The goal is to collectively define assessment criteria for public services based on people's perceptions of the most critical problems. The method is tightly linked to benchmarking.

Examples from our case studies:

- Community scorecards (Black Sash in South Africa / Citizen Engagement Programme in Mozambique)

Classifying / tagging

Citizens can classify existing data sources (see other steps) such as images, sounds, video and other data, in order to extract meaning and add semantic information from data. Some projects like those involving the Humanitarian OpenStreetMap Team combine an easy-to-use interface, task instructions, in combination with an accreditation system for contributors, and a peer-reviewed validation system to coordinate who classifies data and who validates it. Usually done remotely via online interfaces, classifying can gather vast amounts of data with only few people involved. Limits are set by the difficulty of extracting information, which may depend on the quality of existing information used.

Examples from our case studies:

- Remote mapping via web editors (Humanitarian OpenStreetMap Team)

Compiling

Many CGD initiatives include the compilation of information at some point in their work. Some initiatives may search, request access to, or extract information from existing documents. This may be part of an initial research process to define the scope of a CGD project (see data stocktaking). In other cases, groups or organisations may be primarily dedicated to compile data in a central access point, for example by providing a database, or an API. This approach is useful to increase the findability of data and to facilitate the extraction of meaning and insights from unstructured and structured data. Compiling is often a necessary step towards other analytical tasks that are not possible with individual datasets, be it data definition in the beginning of a project, pattern recognition, cross-verification or others.

Examples from our case studies:

- Compilation of public service records to define scope of project (Black Sash)
- Compilation of water monitoring data from government and monitoring groups (Atlantic DataStream)
- Compilation of government maps (OpenStreetMap)

Triangulation

Data that is gathered, and put into relationships through compiling or otherwise may be cross-verified with other data. This can have several purposes. For instance, citizen groups could want to ensure the reliability and accuracy of their data by comparing it against official data collections or prediction models. Likewise, government may use citizen data as a control value to test the accuracy of its existing data and predictive models. In some cases, CGD has the main purpose to provide comparative data and first baselines that governments later verify by conducting their own data collections. These triangulation practices show that CGD is in a relationship to other types of data, and adds meaning to other data.

Examples from our case studies:

- Air pollution monitoring in Kosovo (Science for Change Movement)

Pattern recognition

Many CGD initiatives put data points into new relationships, giving it different types of value. Thereby, CGD initiatives may discover spatial distributions (Where are buildings with higher exposure to disasters in cities? How many households can reach public services?). In other cases, citizens may discover temporal distributions such as pollution spikes at certain points in time, or continuously high air pollution values. There may be different criteria to assess the validity of these patterns. In some cases CGD initiatives have argued that just a sufficient amount of individual points is needed to detect repeating patterns, for example commonly encountered problems in health facilities.

Examples from our case studies:

- Commonly encountered service issues across health facilities (Citizen Engagement Programme in Mozambique)
- Air pollution concentrations in cities (Science for Change Movement)
- Geographic distribution of households with little access to public services (Data Zetu and Humanitarian OpenStreetMap Team)

What questions could CGD methods help answer?

Are you concerned about how local public facilities are used? Would you like to learn about the living conditions of people in your community? Maybe you would like to measure if public services perform as they are supposed to? Table 2 lists CGD methods and what data types they lend themselves to generate.

Example question	Information type	Suitable methods	Use case
How do people perceive the quality of public facilities?	People's perceptions	Client surveys (in facilities) Collective deliberation (focus groups) discussing common issues (e.g. community scorecards)	Community scorecards can be used to detect problems in public facilities that are commonly encountered and agreed upon by groups of people. This helps foregrounding the problems that matter most to people, as well as reasons why people may not use services at all (e.g. health services).

Example question	Information type	Suitable methods	Use case
How do people perceive the security on the street?	People's perceptions	Location-based reporting apps Surveys	Apps can aggregate anonymised reports of locations perceived as unsafe. Local surveys can be conducted with households or in public places/ facilities to understand how safe people feel. Places can be visited, or documented via photos, and sources of safety perceptions be detected (e.g. missing street lights).
What is people's economic status? How many people live below the poverty line in my community/ city?	Socio-economic information	Local household survey (sampling area is lowest administrative zone)	Local household surveys can provide more granular poverty distributions in Uganda.
Where in my city are problems with infrastructure? Are these solved yet?	Government performance	Location-based reporting apps (e.g. FixMyStreet)	Geo-referenced reporting apps can document infrastructural issues and problems in facilities. Problems are pre-defined by app to be ingested into real-time database of government. Apps can be integrated with public works department, to allow for real-time feedback and performance assessment.

Example question	Information type	Suitable methods	Use case
Does the money I have spent reach beneficiaries and services on the ground?	Fiscal efficiency	Compiling of government records On-site observations and surveys ('social auditing')	Social auditing helps communities understand what services they are endowed with, and to evaluate if actual service quality meets planned targets. This can help government auditors and line ministries to understand mismanagement and to surface missing reporting chains in government.
Where are villages, households, infrastructure located?	Geographic features	OpenStreetMap	Granular location data can help frontline workers planning to allocate resources, to distribute medical aid, or simply to understand the number of households and other facilities/infrastructure on lowest administrative levels.
What is the distance between households and public facilities?	Geographic features	OpenStreetMap	Geographic positions of health facilities can be used to understand physical accessibility to basic services and infrastructure.
What is the physical condition of houses and infrastructure ?	Geographic features	OpenStreetMap	Mapping the physical conditions of infrastructure helps model disaster risks. It can help understand where maintenance and infrastructure investments are needed.
How polluted are local watersheds and what does this mean for their use?	Pollution	Participatory water pollution monitoring	Water monitoring can provide data from remote areas or upstream locations (e.g. rivers) to understand actual pollution levels.

Example question	Information type	Suitable methods	Use case
Are existing air pollution predictions accurate in all areas of a city? If not, what could be sources of deviations?	Pollution	Air quality monitoring	Collecting air pollution levels from accredited, distributed sensor technology may help gather sufficiently accurate data to identify possible pollution hotspot patterns in cities. This information can be cross-verified against official pollution predictions.

Table 2: Illustrative list of questions and CGD methods to address these

When is CGD fit for your purpose?

Citizen-generated data must be fit for purpose. Fitness for purpose means data is relevant and usable enough to provide answers to a particular problem (revisit step 1).

An increasing amount of literature **rejects essentialist notions of data quality**. Instead, data can have many 'qualities' (see table 1) which add up to make it a **sufficiently useful dataset**.[3] Policy frameworks, such as the Fundamental Principles of Official Statistics, emphasises **practical utility** (Principle 1) and states the suitability of statistics from different available sources if the quality, timeliness, costs and the burden on respondents justify their use (Principle 5).

In addition to these principles, multiple additional indicators can apply to evaluate the quality of CGD initiatives. This depends on the types of data collected, their intended purpose and the methods used to collect them. For some data there is strong **scientific agreement**, and (scientifically) agreed processes protocols and data schemas exist to account for a phenomenon. Here we provide an illustrative list of quality parameters, and what governments should consider. Governments should make sure to define quality targets and thresholds (minimum useful data). This serves to not only define what data counts as accurate or to pre-define sampling approaches and protocols, but also to define when data is complete enough.

[3] Wang, R. Y.; Strong. D. M. (1996): Beyond Accuracy: What Data Quality Means to Data Consumers. Available at: http://mitiq.mit.edu/Documents/Publications/TDQMpub/14_Beyond_Accuracy.pdf

Intrinsic quality attributes	Quality achieved	Possible quality issues	Steps to assure quality
Accuracy	Research finds that different CGD methods can achieve accuracy comparable to professional datasets, given quality assurance steps are followed (see column to the right).	Data may include errors and not adequately represent the phenomenon. Data collection protocols (e.g. sampling) might not be followed appropriately Tools such as digital sensors might be inherently inaccurate, or not correctly calibrated.	Identify existing inaccuracies. Provide technical support and provide accredited equipment. Ensure sufficient training. Ensure that task difficulty is not too high for your audiences. Iterate data collections to identify error sources and provide follow-up trainings.

Intrinsic quality attributes	Quality achieved	Possible quality issues	Steps to assure quality
Representativity	CGD may be deliberately designed to identify new patterns and distributions and being sufficiently indicative of a problem. Thereby CGD can identify possible issues and trends that are unnoticed by representative data collections. Some CGD methods choose to use new sampling approaches in smaller sampling areas, which can provide new insights beyond national or regional averages.	Citizen-generated data may be self-selected, so that the times and locations of samples are not subject to statistical design. Data collections may be incomplete. Data may be biased towards popular regions (spatial bias), or show engagement spikes and drops (temporal bias). If open participation is chosen, some populations might use the tools more strongly than others. This can bias data towards certain groups or certain problems they express.	Identify and interpret gaps in data. Plan targeted outreach to communities in specific regions. Identify target areas, and data collections with appropriate sampling size together with communities.

Intrinsic quality attributes	Quality achieved	Possible quality issues	Steps to assure quality
Reliability	CGD initiatives may achieve reliable results if quality assurance processes are followed (see column to the right). Different data gathering methods require different safeguards to reliability.	Two people may not describe the same phenomenon in similar terms. This may stem from the fact that people have different degrees of training, but also because they may perceive a phenomenon differently, or even because of group dynamics in case of focus group sessions.	Person-based assessment: Compare data collected based on people's experience and use data from more experienced people as validators. Ensure that people are trained and adhere to well-defined protocols. When dealing with deliberative models and group discussions (e.g. focus groups) ensure facilitators are well trained and detect group dynamics and other factors that can influence people's answers.

Intrinsic quality attributes	Quality achieved	Possible quality issues	Steps to assure quality
Trustworthiness	Participatory data collections and partnerships with trusted organisations and community members can build trust between citizens and governments. This may be helpful to enable deliberation around the results.	CGD may be collected from everyone making it hard for you to identify the data source. Governments may have concerns that people have certain agendas which influence how data is collected.	Person-based approach: identify data submitters if possible, and assess training and degree of experience. Early partnerships with organisations and communities help build trust and can be used to train people, agree on a methodology, or even accompany the community while collecting data. Data-based approach: gather several data samples about the same phenomenon (a location, a problem, something else), compare data across one another. This may include to check the time of creation (for time-sensitive data).

Contextual quality attributes	Quality achieved	Possible quality issues	Steps to assure quality
Relevance	CGD projects may be designed to directly address a governmental issue. Governments may commission CGD projects to gather data. In other cases, CGD may aim to increase the efficiency of institutions and services.	CGD initiatives may be driven by goals different from your purposes. For instance, CGD may experiment with new measures that challenge your ways of measuring data. There communities may care about data that are less useful for government operations	Defining agreed data models that are mutually beneficial for government and CGD initiatives can ensure relevance from the start. Organise dialogue and involvement of communities.
Completeness	CGD can collect data adhering to how official data is collected. This may be helpful when governments have no remit to collect data on hyperlocal levels, or if gathering data from remote locations would be prohibitively expensive.	Methods relying on people's self-selection can suffer from bias towards popular or populated areas. CGD efforts may require longer term data collection to be sufficiently complete, and engagement may spike and stall later, so that data collection is not sustainable over time.	Instead of aiming for completeness, ask when a dataset is complete enough. Define a benchmark value when a data collection counts as complete and compare ongoing data collection against it. Engage tactically and proactively with communities in your target regions. Scope out your target audience and get them engaged from the beginning.

Contextual quality attributes	Quality achieved	Possible quality issues	Steps to assure quality
Granularity	CGD may increase the resolution of existing datasets and thereby open up new ways of analysing data.	Granularity may come with incomparability to other datasets that are collected on similar spatial scales.	Use metadata or gather data attributes that are used by other initiatives, in order to join up data (e.g. by using common descriptions for facilities). Ensure to adjust your sampling approach for data relying on statistical sampling (e.g. local household surveys).

Contextual quality attributes	Quality achieved	Possible quality issues	Steps to assure quality
Timeliness	Some CGD methods may provide data as immediate responses to an event (e.g. the reporting of harassment cases). Depending on the data type, CGD can provide data at faster rates than official data collections. This is the case for participatory mapping, for example. CGD can also update existing data collections with new data. Social audits for instance provide a snapshot of service performance in a given point in time.	Some CGD initiatives may require a significant amount of time to be set up, to build partnerships, and to prepare data collection (e.g. doing local household surveys or social audits).	Ensure that people are available to collect data in certain time intervals. If applicable, consider providing sensors to people which can continuously capture data.

Representational quality attributes	Quality achieved	Possible quality issues	Steps to assure quality
Interoperability	Many CGD projects already abide by governmental data standards, use government terminology, or data collection methods accredited by government. You can benefit from CGD especially, when your own operations are the object of study, or help CGD projects to collect data (e.g. social audits, environmental monitoring).	Some initiatives may actively dispute governmental ways of classifying for being not representative of what your community wants to measure.	Explore ongoing work around data standards for CGD. Provide training to help people using government standards, if this is aligned with their interests. Be mindful that government standards may not reflect what people want to express. Engagement, management of expectations, and alignment of goals is key.

Representational quality attributes	Quality achieved	Possible quality issues	Steps to assure quality
Representational consistency	Representational consistency is achieved when values of a similar kind are described in the same terms. (e.g. 'schools' or 'educational facilities'). Some CGD projects may have well defined data models (e.g. HOT) thereby coordinating how data can be documented.	Some projects may use tools that do not pre-structure data collection. This can be the case when people use open questions to collect data in surveys. Whilst initiatives may collect consistent data, this may not be the case across initiatives, for example in the case of household surveys (survey items may differ) or social audits (the same information can be coded differently, leading to incompatibility).	Provide clear and easy guidance on how to use your CGD gathering tool. Ensure coordination across initiatives if possible. This can be achieved by providing guidance material, and standardise collection tools, and may require more or less outreach, depending on the nature of the community and your existing engagement with them.

Table 3: Illustrative data quality matrix

When data is good enough is a moving target and depends on what the data shall be used for (revisit the definition of scope). Table 4 shows different approaches that can be chosen to ensure data quality.

Quality assurance approach	Explanation
People-based	Identify the contributor's level of experience. Engage more experienced people as validators.
Process-based	Mandate the adoption of data quality management plans. Require data quality / data control measures from your audiences.

Quality assurance approach	Explanation
Data-based	Gather comparative data sets as control values. Increase sample size (can be useful for comparison across subjective information such as perceptions). Use data of more experienced contributors as comparative value.
Tech-aided	Use data models with standardised keys. Gather contextual metadata automatically (location, timestamps) for comparison. Provide trainings on how to use questionnaire.

Table 4: Overview of selected data quality assurance approaches

Resources:

A taxonomy of quality assessment methods for volunteered and crowdsourced geographic information. Available at: https://onlinelibrary.wiley.com/doi/full/10.1111/tgis.12329

A review of data quality achieved by citizen science projects. It emphasises that data quality of citizen science projects may be similar to 'professional collections', in particular concerning accuracy of data. Task difficulty and sufficient trainings are key factors influencing data quality: https://esajournals.onlinelibrary.wiley.com/doi/10.1002/fee.1436

A study assessing the fitness for purpose of citizen science projects: http://edis.ifas.ufl.edu/pdffiles/FR/FR35900.pdf

A report by Statistics Netherlands describing ways to assess statistical fitness for purpose: https://www.cbs.nl/nl-nl/achtergrond/2013/21/quality-reporting-and-sufficient-quality

A report laying out a data quality framework for consumer-centric data: http://mitiq.mit.edu/Documents/Publications/TDQMpub/14_Beyond_Accuracy.pdf

Governmental quality assurance processes for citizen-generated data: the United States Environmental Protection Agency has developed an Integrated Reporting Guidance, outlining the rationale for governments to select unofficial statistics for different use cases: https://www.epa.gov/tmdl/integrated-reporting-guidance-under-cwa-sections-303d-305b-and-314

How much data is needed to generate sufficient insights?

Not all questions require large-scale interventions or long-term data collections. Maybe you want to fill specific regional gaps in your data? Maybe you want to generate first baseline data to test if a problem is important enough to require follow-up measurements or larger interventions?

How much data is needed will also depend on the use purpose, and the intended data user. For example, a regional ministry might require comparative data gathered in different locations. You may ask yourself following questions:

1. What spatial expansions should your dataset cover? Do you want to focus on a particular city, a neighbourhood, a facility?
2. At what administrative level should governments make use of the data? How does this influence the required scale of data?
3. Over what timescale should data be collected? Do you need to collect data repeatedly, or in particular time intervals to make them useful? For example, does the phenomenon you look at only become meaningful if longitudinal data is collected?
4. Does one-off data collection suffice? For example, do you plan to collect information about immobile infrastructure, buildings, and other phenomena that do not require constantly updated data? Do you need real-time data?

Example 1: Mapping houses for a malaria intervention requires to collect data in regions that are particularly affected. The Humanitarian OpenStreetMap Team and Clinton Health Access Initiative worked with Botswana's Ministry of Health and Wellness to identify the regional dimensions to be covered, and recruited community members from these places to collect the data. The data collection could be done once to provide updated and granular information on buildings.

Example 2: Black Sash has used community scorecards in South Africa to identify and agree on action plans that address the most pressing service delivery issues. To understand whether community scorecards change public service delivery, it is important to continue monitoring the implementation of action plans.

Example 3: Mapping air pollution hotspots in Pristina required to monitor PM2.5 pollution levels in pre-defined time intervals over a longer period of time. The Science for Change Movement identified priority locations in a pilot phase and then repeatedly collected data in a fixed spot.

Example 4: The Citizen Engagement Programme developed a standardised taxonomy of community scorecards in Mozambique with the goal to identify the most often shared problems in public service facilities. It is currently planned that the national Ministry of Health bases funding decisions on comparative information coming, among others, from community scorecards.

How conducive are CGD approaches to scaling?

Some CGD approaches allow data to primarily be an online process, with people contributing from anywhere. Other methods require people to manually collect data on the ground. In some, data collection can (at least partly) be delegated to machines. Table 5 shows examples of CGD and how they enable scaling differently according to how the **organise and distribute labour**. As the ways of generating data show, these data production types can be embedded into other tasks and infrastructures. For instance, many local water monitoring samples may be compiled later, further increasing the size of datasets centrally accessible. In this question, we focus on how the production of data itself can be scaled:

CGD examples	Type of method	Scaling enabled
Classifying/annotating: HOT web editor	Web-based method	Small groups may identify many data points by classifying image content. Contributors from around the world can contribute (large-scale) datasets, that are already produced, and can be further analysed.
On-site observation: Social auditing HOT field survey **Sample collection:** Water sample collection	Field-based method	Small or large groups of people are required, depending on the size of the territory.
Automated, stationary sensing: Weather observation stations	Tech-aided method	Real-time and longitudinal sensing in different intervals possible in fixed location.
Automated, mobile sensing: Sensor technologies implemented in cars, and other consumer devices	Tech-aided method	Real-time and longitudinal sensing in different intervals possible in location where people use consumer device.

Table 5: Scalability of CGD methods

Web-based methods may structure, enhance, or compare existing data. In this case, data is not necessarily produced anew, but rather derived from existing data. Web-based data collections may enable contributions from everywhere worldwide, with few people generating larger amounts of data or metadata, as in the case of web-based classifying and tagging.

Field-based methods may involve households surveys, sample and specimen collections, or on-site observations. In each case people collect (often new) information. Increasing the amount of data collected usually requires to increase the group size collecting data, to expand the time to collect data, and to collect across locations, or in repetitions. Field-based data collections require to produce data from scratch. Challenges to scale include physical barriers such as travels, or the accessibility of locations. Security questions (how dangerous is an area) or questions of access to territories or communities to collect data may also play a role in how much data can be scaled.

Tech-aided methods is based on immobile or mobile tools such as stationary or portable cameras and sensors technologies. Tech-aided methods can help collect data which can otherwise not be documented, can hardly be accessed, or which require long-term data collections, collections in real-time, or in well-controlled time intervals. Data collections may be scaled in time (collecting data in different intervals), or across space (depending on whether methods use immobile or mobile collections this may depend on the number of contributors or investments in distributing immobile technologies).

Some ways of scaling data further once it is produced:

Example 1: Australia's Weather Observation Website (WOW) allows farmers to integrate data from their weather stations onto its website. WOW also functions as an aggregator of weather station data.

Example 2: Mozambique's Citizen Engagement Programme applied a standardised taxonomy to make data from community scorecards comparable. This way data gathered in different locations could be spatially aggregated in order to commonly encountered public service delivery problems.

Step 3: Clarify how the participation of people will help

Citizen-generated data is as much a process of creating data, as it is a way of engaging with groups outside of government. Governments have developed different pathways to engage with citizen-generated data. **The participation of citizens can vary in breadth and depth**, depending on purpose of the initiative (revisit step 1 on goals and scope), and how the engagement of citizens would make a difference. Here we propose several questions for governments to identify groups, to organise reach-out, and to choose adequate participation formats.

Who are the target audiences of your project?

Some governments may have a clear goal for their project, and an established group of people to engage with. In other cases governments may want to reach out to new communities they have not engaged with, such as civic technology communities, local community networks, or others.

Some communities are well-defined and have many things in common, even if their interest in your project focuses on a specific issue, question or concern. Others may be a disconnected group of people who share a common interest, concern or hobby.

1. Which groups of people do you plan to engage with?
2. How are they connected already and in what sense? For example, do they share similar concerns? Are they part of an organisation or group of people?
3. What demographic features do these people have? Do you address people of a certain age? Or people with a specific educational degree, level of expertise, or access to resources?

What is at stake if you do not define your target audience to engage with?

“Build it and they will come” is unfortunately a common misconception when designing participatory, and civic technologies. You may not only risk to design projects that do not motivate people, but can also exclude the perspectives of the beneficiaries you try to address.

Some CGD projects are open for everyone to contribute without moderation who contributes, which can lead to strongly self-selective participation. While in some cases this can be a strength for the method, it may result in other cases in unintended outcomes. A study of issue reporting apps in the US showed that primarily wealthier people used issue reporting whilst poorer neighbourhoods could benefit less from public works.

How do you reach out to people?

To mobilise people, several strategies can be employed. Consider which ones are best suited to engage your community and your project. Find the best platforms for reaching your community. For example, some online groups share information about specific diseases, while in-person groups may deal with local issues such as air pollution or environmental justice:

1. What media do people usually consume, and what can you learn for your engagement strategies?
2. Is there a group of people with established communication channels you could join?
3. Can you collaborate with people from within the community, or do you otherwise need to establish connections and trust with a community?
4. What style of language and tone are most preferable to speak to each community?
5. How do you plan to disseminate the data? What channels and media will most effectively reach your audiences?

Use mixed channels for PR and engage with existing communities: Statistics Canada has used a mix of outreach strategies, including online advertisement, features in newspapers via PR, and engagement with established OpenStreetMaps (OSM) communities.

What is the adequate depth of participation?

CGD initiatives may have different participatory ambitions. You might want to engage citizens only in the data collection phase. Maybe you want to use CGD as part of training and educational programs. Maybe you plan to engage groups of people more deeply in how a problem could be measured, and want to help people understand not only data collection and manipulation techniques, but also how data is embedded in institutions. You may ask yourself:

1. What could citizens contribute to the definition of a project? What could you learn from citizens, and how could your problem-definition benefit from this?
2. In what ways should citizens be engaged along a project? Do you want to consult people for their opinion, or do you want to establish more substantive dialogue channels?

Example 1: The community scorecard method includes action plans, which are developed by community members and public workers. Action plans help making sense of the results of the scorecard process, and to agree upon tasks and responsibilities across government and communities. Problem ownership is ideally transferred to government and communities.

Example 2: The Science for Change Movement emphasises that data must be actionable and 'campaignable' for the communities producing them. The group has educational and campaigning committees which help the people involved in air pollution monitoring to make sense of the data, to understand the implications on people's health, but also to inform them about their rights, and the government institutions who are responsible to manage air pollution.

Use active participatory formats: such as discussions, stakeholder meetings, and others. The Humanitarian OpenStreetMap Team organises mapathons, and community meetings. Social audits are usually built around multi-stakeholder meetings from the get-go, bringing together government and civil society. This helps increasing trust between communities and other interest groups.

Let people know how they can engage with your organisation. The United States Environmental Protection Agency provides a series of toolkits, quality assurance protocols, trainings, and engagement channels for volunteers to monitor environmental pollution.

How else participation could be designed. Action research, advisory committees, citizens' juries, community reference groups, retreats, drama workshops, learning circles, design workshops, focus groups, participatory editing, policy action teams, citizens' panels, deliberative polling, summits, World Cafes, community visioning and community cultural development.

What motivates your audience?

It is important to know whether a given citizen-generated data project will appeal to participants. Carefully examine assumptions about the people you engage with.

CGD projects may include tribal communities, students, self-selected volunteers, professionals and topical experts, and many other groups. Working with many groups in a single project would likely require a distinct engagement strategy for each participant community. The aims of the project need to be clearly explained to manage the expectations of the participants.

1. Is your audience interested in what you think is important?
2. Have you clarified the duration of your project?

Mind your language when engaging people

A report by the UK Environmental Observation Framework studied motivations to participate in citizen-generated data projects. Among others, the report highlights that a misalignment of language, used by governments and citizens, may demotivate or divert people who have particular interests in a CGD project. It is important to listen to how people frame problems.

A recent study shows that there is no single concept applying for all CGD initiatives, that could express how people are involved, what their motivations are, or what they are doing in a CGD project. The study highlights that people engaging with CGD must consider how they frame CGD. For example **speaking of 'volunteers', 'amateurs', 'citizen sensors' or 'citizen scientists' may motivate or alienate different kinds of people**.

How do country culture and socio-economic factors influence your engagement?

Your country may have a different culture of volunteering. Volunteering may have positive or negative connotations, and may be more or less well established as a feature of civic life. In some contexts, and for some projects, volunteering cannot be expected from people if these have no resources. Stipends, contracts, and other forms of remuneration can compensate for people's time.

1. Is the context you operate in conducive to volunteerism? Would people depend on CGD projects as an income source?
2. What dynamics could payments create between the people you engage with?

Can CGD build on existing participatory channels in your government?

Governments have started to integrate CGD in their usual engagement channels with citizens. For instance, South Africa's Department of Monitoring and Evaluation has developed guidelines to include community-based monitoring in the work of administrations and public services. Provisions in policies may underline the role CGD can play. In the US, the Clean Water Act contains several provisions strongly arguing for the involvement of citizens in producing data for the EPA's monitoring programs.

1. Do frameworks exist for these administrations to guide engagement with citizens?

2. Are these frameworks conducive to include participatory monitoring, data collection, and deliberation around sustainability data?

How open or closed should your project be?

Membership can be key to govern roles, responsibilities, but also the access to data collection tools, and data collected. How open your project is for participation can depend on different questions:

1. Do you aim to collect confidential or personal data?
2. Do you want to ensure to collaborate with accredited data collectors? (who will be allowed to visit sites? Who have trust in communities? Who are trained?)
3. What access restrictions should people have to data and why?

How many contributors do you need?

CGD projects can come at all scales, from local interventions, to large-scale data collection efforts. What group size you want to achieve depends partly on the amount of data you want to collect and to what extent tools enable your community to scale data production. Revisit how much data needs to be gathered, where and when (revisit question 'How much data is needed to generate sufficient insights?').

How much data can be gathered with your tools (revisit question 'How conducive are CGD approaches to scaling?')?

Step 4: What resources are available to support CGD?

Citizen generated data may be free at the point of data collection, but it is (often) not cheap. Resources, coordination, infrastructure, and organisational changes are usually required to effectively support CGD. The following questions can help you think through the resources required to run a CGD project:

How much work will the coordination and facilitation of partnerships require?

Some CGD initiatives may require more partnership building depending on the data collected, and the role of partners in collecting, disseminating and using data. Collecting data on common property (e.g. data collected on the street, in observations, etc.) may not require permission to collect data, but collections on private property, households, or public facilities may require arranging meetings, appointments, permissions, and conduct data collection. In other cases, you will need to involve people in the definition of the scope for your initiative. Targeted outreach and iterative development of data models may be necessary:

1. Do you need permission to collect data? How much time will the relationship building with the relevant organisations take?
2. Do you need to reach out to citizen groups prior to data definition? What participation formats are planned to engage people, and how long are these formats planned to run?
3. Is it helpful to engage with CGD initiatives around methodologies, and to formalise partnerships and responsibilities?

Example: Black Sash has formalised several partnerships with government in form of Memorandums of Understanding. This helped ensure buy-in from the government, define responsibilities and allow Black Sash and community-based organisations to get formal access to public facilities to collect data. In the case of Black Sash, some of their partner offices have taken a full year to prepare for the community gathering.

How much support is needed during data collection?

Different CGD projects may require different supervision prior or during the data collection phase. Usually, CGD is steered by a leading organisation managing training, engagement, communications, equipping people, and other tasks. Deliberative formats such as community scorecards may require strong facilitation skills (e.g. during focus groups and community meetings) in order to understand power dynamics in communities, and to gather unbiased data. Here is a short checklist of things to consider:

1. How much training is required per person to reach a sufficient level of expertise?
2. Is the task easy or clear enough that people can instruct themselves (e.g. via online courses or tools)?
3. Do you rely on experienced validators to cross-check data collected? Are these available in your team or do you need to recruit them?
4. How many hours of communication with your community do you expect throughout the project? Base your assessment on a rough count of questions that might arise, as well as your planned communications activities.

What tools do you plan to use for gathering data?

Your communities may have different habits of using media, or different literacy levels. Consider:

1. What media are people habituated with?
2. What blockages could arise when people are not familiar with the tools you want them to use?
3. Have you tested your tool with the communities to ensure user-friendliness? What have you learnt for the design of your tools?
4. How accessible and usable is the necessary infrastructure, and how does it enable the inclusion of citizens?
5. Does the data collection require an online connection to be usable? Is the project entirely computer- and web-based? Could this exclude certain groups of people from collecting data?

University College London runs the project MOABI to monitor natural resource use in the Democratic Republic of Congo. To make the monitoring app usable by indigenous people, the app interface is codified in symbols that local populations can easier understand. http://rdc.moabi.org/en/

Black Sash is a South African NGO using community scorecards to advance social justice. A key learning of their work is to design tools usable by communities. In collaboration with Code for South Africa and other organisations, they increased the user-friendliness of their tools: https://www.blacksash.org.za/index.php/introduction-to-community-based-monitoring

The Making All Voices Count project has summarised some lessons for users of civic technologies, which may be helpful to think through CGD as well: https://researchfindings.tech/

Tool acquisition and maintenance - what costs will be incurred?

CGD initiatives can work with different tools. Sometimes it suffices to use consumer devices such as smartphones and open source apps. In other cases accredited technology needs to be acquired by citizens to be able to collect data. Governments, NGOs and universities can provide accredited technology to collect data, or can help calibrate and maintain tools. To estimate incurred costs, you may ask yourself:

1. Does the data I monitor require special tools people usually don't own?
2. How many people need to be equipped with the tool, and are there opportunities to share the tool?
3. How much does it cost to maintain the tool?
4. Are there organisations already providing accredited tools? Is there an opportunity for me to help these organisations acquiring and distributing tools?
5. What support could these organisations need to scale the provision with more monitoring tools?

Organisations that develop open source tools or loan equipment: Community networks like Atlantic Water Network are hubs for equipment, training, and exchange, and good organisations to support and collaborate with.

Atlantic Water Network loan equipment to citizen groups, help calibrate tools, and offer training. The Equipment Bank *"functions as a library for water quality monitoring kits for citizens, community groups, volunteers, who have any form of interest in the health of their lakes, rivers, streams, even ocean coastlines, to go out and gather data."* https://atlwaternetwork.ca/

The Public Lab publishes a suite of open source hardware tools people can use: https://publiclab.org/

The Open Data Kit is a customisable tool for mobile data collection. It allows to design surveys and can be integrated with other tools like OSM: https://opendatakit.org/

The United States Environmental Protection Agency has a dedicated website with guidance material, quality assurance guidelines, but also explanations how data can be sent to the EPA to be further used.

Which funding sources are available to your audiences?

Citizen groups, community organisations, NGOs and other members of civil society can have different resources available to partake in CGD projects. For instance, CGD efforts can rely on project cycle funding, which may not be sustainable in environments with reduced donor support.

How much donor support does civil society in your country receive, from which donors?

Step 5: Making citizen-generated data public

Citizen-generated data may be of further use when made accessible to other groups. There are several considerations as to when and how data should be made accessible. Some projects deal with sensitive types of data, or rely on the trust of contributors that data is safely shared. Different approaches exist and you may need to inquire following questions when engaging with CGD initiatives:

What licence could you apply to your data?

If you provide infrastructure for CGD projects to host data on, you have several options to licence data. You may wish to make data freely accessible to anyone with restrictions at most applying to provenance and restrictions that retain the openness of the data. You may also want to limit the reuse for certain use cases, for example not allowing commercial reusability, or non-derivatives of data. For instance, some initiatives argue that opening up data and information about the data can increase trust in data by making data and collection methods verifiable. Some types of data, such as environmental data, or factual data on public facilities might be more desirable to open up than others. You may ask yourself:

1. Does the my legal context provide for ownership rights to data as part of intellectual property protection, or neighbouring rights?
2. Is the data collected to be considered in the public domain (such as for factual data in the United States)?
3. What licensing restrictions are most suitable to your case and the communities you are working with? For example, is it desirable to prevent commercial reuse? Is is desirable to provide data only for certain purposes?
4. Do you want others to freely combine and reuse the data you publish? What legal incompatibilities could your licence choice bring? How would this impede the ability of others to reuse data?

Example 1: HOT as well as OSM use the Open Data Commons Open Database Licence (ODbL 1.0). This licence is a so-called share-alike or 'copyleft' licence. This means that all works using OSM data must be licenced under the same terms as ODbL when made public. This has the goal to retain the openness of data, and to prevent that data gets published under a closed licence.

Example 2: Atlantic Water Network's Atlantic DataStream uses terms of use which limit the use of data only to research and educational purposes.

Further readings:

A study on intellectual property in citizen science projects outlines different licencing models for communities, as well as possible issues for licensing.

The Open Data Handbook, developed by Open Knowledge International, provides guidance how an open licence can be applied to your datasets.

The Open Definition is the gold standard for open works and open data. It defines elements of an open licence, and lists several recommended standard open data licences.

How much flexibility might citizens need when licensing their data?

CGD initiatives handle data licensing in different ways. Some initiatives which host data apply terms of use that apply by default to all datasets uploaded. In other cases, people have more discretion as to what licences they want to apply when transferring data onto data infrastructure.

1. What concerns could people have when uploading data to your infrastructure?
2. How stable and predictable should the provision with data on your infrastructure be? Could you avert users if people revoked the right to use data stored on your infrastructure?

Example: LandMark allows data uploaders to to select the accessibility to data. A data sharing agreement regulates how data will be shared, who can use and access data, or how data should be described and quoted. In addition, their data sharing agreement provides for the possibility to revoke licensing terms. As the website states: “Individuals or institutions contributing data to LandMark retain full ownership over their data. Contributors can choose, at any time, to remove or update their data displayed on LandMark.”

What is legal interoperability and why does it matter?

Open licences are legal arrangements that grant the general public rights to reuse, distribute, combine or modify works that would otherwise be restricted under intellectual property laws. Yet, not all open licences are equal. There are major differences in between open licences. People and organisations might want to create their custom licensing terms. This however can lead to incompatibilities between licence terms, and ultimately create legal uncertainty for people who want to use and build on CGD. This phenomenon of creating new, possibly incompatible licences is call ‘licence proliferation’. You may ask yourself:

1. What is my intended purpose to licence data? Are there reusable standard licences that provide for my purpose?
2. How much time would it take to clear rights and develop new licensing terms instead of adopting existing standard licences?

Incompatibilities across open licences.

This report outlines the problem of open licence proliferation and provides recommendations as to what reusable standard licences could be used to enhance legal compatibility across licences.

Step 6: Consider risks, responsible data use, and protection

CGD may deal with sensitive or personal data, and may highlight people's experienced problems as well as render them vulnerable. In these cases, principles of responsible data production and use, data protection, but also ethical and legal considerations are crucial. In the following we outline some of the challenges and questions CGD initiatives may be facing. For further information, we refer to the suggested readings added in the end of this section.

To what degree are citizens informed, and can influence how data is being used?

CGD initiatives are often initiated and organised by a leading group, such as an international NGO, a group of researchers, an NGO, a university, a government agency or others. Goals of citizens and leading groups might not align, and citizens might not be aware, or in control, of how data is being used. This is particularly important because CGD can deal with more or less sensitive types of data, but also because use cases of data might not align with what the data was collected for:

1. Did I ensure that all citizens are informed about how data produced by them, or about them is being used?
2. Can citizens consent (and revoke consent) as to when and how data is produced, accessed and used?
3. Are citizens able to manage with whom they share the data they have collected?
4. Are citizens (or groups thereof) able to define terms of use, access modalities, and other governance tools to control data use?

Example 1: Landmark gathers land rights data from indigenous peoples, using a data sharing agreement to enable people to consent and to revoke the right to share data.

How would citizens expose themselves to risks if they collect data?

CGD is collected under different circumstances. People may need to venture into difficult territory, be exposed to wildlife, or operate in politically difficult situations. If the security of people cannot be assured, CGD may not be a suitable approach to gather data. Take into consideration following questions:

1. Would the data collection require entering difficult terrain?

2. Are the communities you engage with under threat?
3. Does the general political climate, or dynamics across interest groups put data collectors at risk?
4. Could other groups be interested in the data you collect, and what unintended effects could this bring?

Does the CGD initiative you engage with collect personal or sensitive data?

CGD does not only render populations and their problems visible. In some cases, it may also make these groups more vulnerable.[4] Some CGD initiatives may deal with sensitive data. Harassmap and Utunzi, for example, enable to collect cases of harassment and violence against women and LGBT communities. Other sensitive data may include political opinions, ethnic origin, or people's beliefs. Some CGD initiatives collect data in such a way that it is attributable to individuals and groups.

The scope of personally identifiable information/personal data varies, but new advances in technology suggest that any piece of information able to identify a person should be considered PII.[5] This includes beyond people's names, home addresses or other common identifiers also IP addresses of digital devices, location information, or personal names attached to the datasets people collect (e.g. satisfaction surveys). Depending on the context you operate in, data may put people at risk if not shared responsibly.

1. Does the data collected include personally identifiable information in its widest sense?
2. Is the data collected sensitive or politically charged?

Have you considered relevant steps to ensure data protection?

Citizen-generated data should follow existing recommended data protection principles, in order to ensure that sensitive and personal data is only collected and used for a formerly specified and legitimate purpose stated at the beginning of data collection. Data minimisation, as well as limitations to storage and purpose are important measures to keep in mind. In addition, organisations handling personal data should consider protection against unauthorised access and processing as well as accidental loss or damage of data. Some of the considerations:

1. Which of this data is absolutely necessary to collect, and do you have options to minimise the amount of data collected?
2. Have you defined limitations as to how long data is stored? Does this time span correspond to the original use purpose of the data?
3. Have you clearly articulated the purpose of data collection, and how data will be used?
4. Does the technological process used to collect and process data enable privacy from its inception (privacy by design?)

[4] Group privacy: Linnet Taylor 2016, McDonald: Ebola. A big data disaster.

[5] https://privacyinternational.org/sites/default/files/2018-09/Data%20Protection%20COMPLETE.pdf

5. Have you taken organisational and technological measures to ensure that data collections are protected against unauthorised access or processing?

Further readings:

Privacy International published a guide on how to ensure data protection: https://privacyinternational.org/sites/default/files/2018-09/Data%20Protection%20COMPLETE.pdf

The responsible data group helps humanitarian and other organisations develop responsible data principles across different sectors: https://responsibledata.io/

The global MyData network develops human-centered principles for the management and use of personal data: https://mydata.org/declaration/

What to do next?

Try specific toolkits for governments and CGD designers

A toolkit to mobilise and organize a community of citizens around open source technology

https://waag.org/sites/waag/files/2018-03/Citizen-Sensing-A-Toolkit.pdf

A toolkit for governments to engage with remote sensing:

http://making-sense.eu/publication_categories/toolkit/

The toolkit shows "how open- source software, open-source hardware, digital maker practices and open-source design could be used effectively by local communities to appropriate their own sensing tools to make sense of their environments and address pressing environmental problems."

A toolkit to run satisfaction surveys of public service delivery (South Africa): https://www.dpme.gov.za/keyfocusareas/cbmSite/CBM%20Documents/CBM%20Toolkit%20V1.pdf

U.S. Federal Crowdsourcing and Citizen Science Toolkit: https://www.citizenscience.gov/toolkit/

The UK Environmental Observatory Framework has created a list of toolkits, reports and guidance material to help agencies select and engage with citizen science.

http://www.ukeof.org.uk/resources/citizen-science-resources

US Environmental Protection Agency provides methods manuals, official protocols, and protocol certifications, and it endorses groups that follow its guidance. https://www.epa.gov/air-sensor-toolbox

Familiarise yourself with specific approaches and questions around concepts related to citizen-generated data

The European Citizen Science Association has published a collection of Citizen Science guidelines and publications, including guidance on how to evaluate outcomes from those initiatives, discussions around intellectual property, or how to set up different participatory formats such as 'bioblitzes'. https://ecsa.citizen-science.net/blog/collection-citizen-science-guidelines-and-publications

Visit and add on to our extended list of CGD initiatives

In our research, we could only scratch the surface and present some ways of doing CGD. As part of this research we have compiled a list of more than 200 CGD initiatives. The list includes the name of the initiative and links to their web presence. We suggest to visit the sites to explore further.

This report was funded in part by a grant from the United States Department of State. The opinions, findings and conclusions stated herein are those of the author[s] and do not necessarily reflect those of the United States Department of State.

Global Partnership for Sustainable Development Data

OPEN KNOWLEDGE INTERNATIONAL

Public Data Lab