IDEAS
FOR
CHANGE

Data generated by citizens in Catalonia: mapping and taxonomy

Giovanni Maccani

Javier Creus

Lucía Errandonea

December 2021

IDEAS
FOR
CHANGE

About Ideas for Change

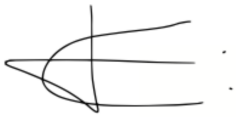
Ideas for Change is a pioneering research and consulting company in impact innovation. Design and build #FuturosQueMolan to contribute to the improvement of cities, organizations, public institutions and social entities.

We specialise in the design of open and contributory business models while enabling the exponential growth of organisations in the digital environment. In addition, we have extensive experience in the development of citizen participation strategies, in the promotion of collaborative economy initiatives and of data economy projects.

www.ideasforchange.com

Project Leader

Javier Creus, *Ideas for Change founder*



Research Team

Giovanni Maccani

Lucía Errandonea

Elsa Boloix

How to cite this report: Maccani, G., Creus, J., Errandonea, L., "Support to the Mapping of CGD Cases and Actors for the Project Catalunya un País de Dades", Ideas for Change, 2021.

Table of Content

1.Introduction

Over the last two decades, the active engagement of citizens in data-driven innovation has grown significantly across different levels. Citizens are increasingly empowered to contribute to innovative policy making and participate in socio-technical innovation (Hecker et al., 2018)[1] through so-called Citizen Generated Data (CGD) ecosystems. This is partly due to the rapid adoption of open innovation paradigms and the advancement and pervasiveness of today's digital technologies (Balestrini et al., 2017)[2]. This involvement can take many shapes and forms, and generally comes together under the umbrella of CGD.

In this study, we adopt (Wilson and Rahman, 2016)[3]'s definition of Citizen-generated data (CGD) as *data that people or their organisations produce to directly monitor, demand or drive change on issues that affect them. This can be produced through crowdsourcing mechanisms or citizen reporting initiatives. This is distinct from "big data" or social media data, which is indirectly created by citizens through interaction with media platforms*. We therefore focus on projects and initiatives where citizens play a conscious role in generating data sources for multiple, public, uses and purposes.

The potential of CGD is well known and well acknowledged, but the journey of fully achieving it is still at its infancy. The paradigms and technological advancements in terms of big data and decentralisation of processing capabilities (i.e. cloud computing), combined with open innovation trends and decreasing hardware and software costs, are some of the main drivers behind this evolution. In general terms, CGD opens up the potential of either generating new data, which was impossible to coherently gather, store and reuse before, or exploiting the pervasiveness of data points as complementary data sources harnessed to expand coverage and depth of existing data. In the most typical situations, the starting point is often citizens being not satisfied about the management of certain aspects affecting their everyday life. This could be, to provide some examples, in the form of lack of accurate and open data about a phenomenon from the public sector, or because of a lack of trust between citizens and government agencies, or to raise awareness about phenomena that receive little to no attention from the public (and policy) sphere, or even as a planned infrastructure to complement data granularity to institutional sources.

The field of CGD is arguably complex and currently scattered and, in most cases, fragmented. This short study aims at generating a coherent taxonomy to understand and map existing CGD initiatives and projects, with a specific focus on the Catalonia region in support of the definition of the *Catalunya un País de Dades* upcoming project. This document is therefore structured upon three sections. After this introduction, section two represents the core of this report. First, it presents the detailed objectives for this study and the methodology designed and followed. Then the actual CGD taxonomy generated from the mapping and analysis of 50 projects is presented across

---

[1] Hecker, S., Haklay, M., Bowser, A., Makuch, Z. and Vogel, J. eds., 2018. Citizen science: innovation in open science, society and policy. UCL Press.

[2] Balestrini, M., Rogers, Y., Hassan, C., Creus, J., King, M., and Marshall, P. 2017. A City in Common: a Framework to Orchestrate Large-Scale Citizen Engagement around Urban Issues. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp.2282-2294), ACM.

[3] Wilson, C., Rahman, Z. 2016. Citizen Generated Data and Governments. DataShift.

its three main elements: (1) the project; (2) the CGD itself; and (3) its destination. Each is presented in a dedicated sub-section which is further enriched with examples from the mapping exercise as well as its preliminary analysis. Finally, section three proposes some reflections from the study undertaken together with some concluding remarks.

2. An emerging Taxonomy of CGD projects

The purpose of this short study is therefore twofold: (1) mapping CGD-based projects and initiatives focusing mostly on the Catalan ecosystem; and (2) design and outline a taxonomy of these projects to be able to understand how these are structured and how these differ from one another. These two objectives are somewhat intertwined as the mapping exercise is the foundation upon which the taxonomy is generated and, on the other hand, the taxonomy is made available to inform further mapping and positioning of other new or emerging CGD projects and initiatives.

The outputs of this study are therefore two, i.e. consistent with the two objectives set at the beginning: (1) an emerging taxonomy of CGD initiatives; and (2) an integrated worksheet file where the 50 initiatives and projects are listed and mapped within the taxonomy itself. In terms of the mapping exercise, a living document has been created, shared, and will be continuously updated beyond the end of this study. At the time of writing this report (i.e. December 2021) it showcases the 50 CGD-based initiatives and projects initially considered:

Link to the CGD Map Living Document

As a mechanism to enable everyone to contribute to this mapping effort, we have created a survey-type of process to allow inputting new projects into the living document. New submissions are visualised in real time (i.e. upon completion of the survey) and highlighted in the main mapping document (link above) with different colour labels. These will be cleaned, validated and finally accepted on a bi-weekly basis.

Link to the CGD projects submission process

Regarding the taxonomy, this is presented and discussed extensively in this chapter, i.e. the core of this document. A full list of the projects considered is available in Appendix 1.

2.1 Methodology

To guide the aim of this study, i.e. to investigate and outline an emerging taxonomy of CGD projects and initiatives, with a specific focus on the Catalan regional context, we relied on the Information Systems (IS) literature. In this discipline, Nickerson et al (2013)[4] propose a method for taxonomy development in which taxonomies are seen as IS artifacts for bringing order to complex areas and potentially lead to new research directions, i.e. consistent with the scope of this analysis. In their paper, Nickerson et al. (2013) distinguish between *Empirical-to-Conceptual* and *Conceptual-to-Empirical* approaches, in which the former takes an inductive reasoning from the bottom-up analysis of existing practices and examples (in this case CGD-based projects and initiatives) and the latter a deductive one from the top down study of given theoretical constructs. Given the exploratory nature of this study, we adopted the former, i.e. the *Empirical-to-Conceptual* approach to taxonomy development.

[4] Nickerson, R.C., Varshney, U. and Muntermann, J., 2013. A method for taxonomy development and its application in information systems. European Journal of Information Systems, 22(3), pp.336-359.

The first step in this research process was about searching, ordering, and considering an appropriate amount of CGD-based projects and initiatives upon which the taxonomy has been built. The search and selection of projects was undertaken adhering to significant degrees of diversity among themselves (in terms of context, discipline, sector, type of funding, organizations involved, leaders, size etc.). For the purpose of developing the taxonomy, a total of 50 projects were established as an appropriate number to be considered, given the short timeframe of this study. The 50 projects were identified as a result of a consistent desk research effort across a number of global databases (e.g. EU Cordis), local ones (e.g. Citizen Science Office in Barcelona), our own experience, as well as from the webpages of organizations that are well acknowledged actors in this ecosystem (e.g. ISGlobal, Ibercivis).

Once identified and listed, the analysis was set to start. This has been conducted in a bottom-up fashion whereby projects were considered both individually (to understand their details and specificites) and as groups (to understand how projects differ among themselves and thus shape the taxonomy's elements). This iterative exercise has allowed us to gradually generate dimensions and categories of projects, i.e. the structure of the taxonomy itself, which is presented next.

2.2 Three overarching dimensions

To classify CGD, given the focus of this analysis, three overarching and consistent dimensions have been defined and taken into account: (1) the project or initiative; (2) the data generated as part of this; and (3) the final destination of this data and/or the project's findings and results.

First, the actual project or initiative itself is found as a useful dimension to report on the variety of natures underpinning CGD-based endeavours. Within this category, defining features and variables for each initiative can be further divided into two areas: (1) a general overview of the project, including its name and link, the sector(s) in which it is positioned, the geographical scope, the source of funding, the leading organizations (highlighting those from the local regional context of Catalonia), the duration and its status, i.e. whether it is active or terminated; (2) the main role of citizens in the project defined across growing levels of responsibility and ownership.

Second, the actual data generated and produced is taken into account across two sub-dimensions: (1) the nature of the citizens' contribution (i.e. giving access to their own data versus producing data themselves) and its type; and (2) the actual approach to data collection in terms of both the tool(s) leveraged and the timeliness of the data at stake.

Finally, the third overarching theme presents CGD projects and initiatives based on the final destination of its outputs as well as on its (sometimes intended) outcomes. In addition, projects are classified based on the level of openness of its outputs, and specifically of the CGD produced or the result of its analysis. This is taken into account by considering the presence (or absence) of open datasets (or content) and their associated open license.

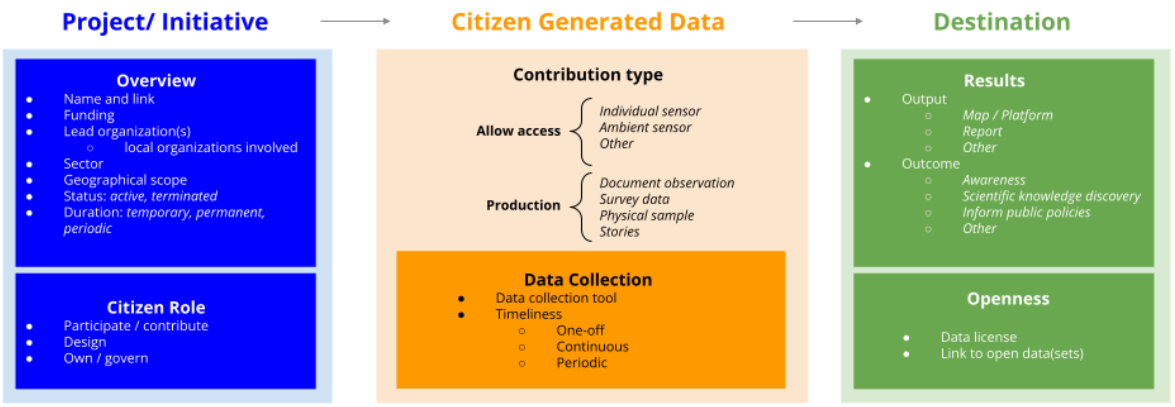


Figure 1: The CGD Emerging Taxonomy

2.2.1 Project

The first, and most obvious, differences among the projects considered are identified at the project level across a number of defining variables. Some of these are somewhat general in nature and include the project's name, sector, geographical scope, duration and status as well as its governance, i.e. the funding mechanism(s) and body(ies) and the leading organizations with a specific focus on the Catalan context. Other more articulated elements emerged from the mapping exercise and refer to the projects' governance arrangements with respect to citizens, i.e. their role in the initiative. These variables are tackled, defined, and enriched with examples and a preliminary analysis from the mapping exercise, in the following dedicated subsections.

Sector

Beyond the name of the project and the relevant links to access its websites and other resources, one of the most immediate variables useful to classify CGD-based efforts refer to the main sector in which these are intended to provide a positive contribution. While some could be considered at the intersection between two or more sectors (e.g. *CitieS-Health* operates in the multidisciplinary field of environmental epidemiology - i.e. at the intersection between environment and health), the main discipline has been highlighted, starting from the operating field of the leading organizations or communities involved. In only two cases (i.e. *Fumuts Ros de Olano* and *Olot Community*) this distinction was not possible as both communities have established their focus on both environmental sustainability and sustainable mobility.

In terms of findings, of the 50 initiatives considered, more than half (28, i.e. 56% of the total) focus on environmental matters of different kinds and from different angles. In particular, air and noise pollution and microplastic monitoring - classified as *environment* in the taxonomy - and biodiversity are the most represented with 13 and 7 related projects respectively, followed by: water quality and management (2 projects), weather related initiatives (2 projects) and one further initiative focused on climate change more generally. Three additional clusters have been identified in relation to: health (8 projects), socially relevant matters (7 projects), mobility (4 projects) and data and connectivity (3 projects). Less popular, but still represented in the sample with one project each, are: circular economy, food, aerospatial, urban planning and design, and energy. A summary of the distribution across sectors is provided in the following figure.

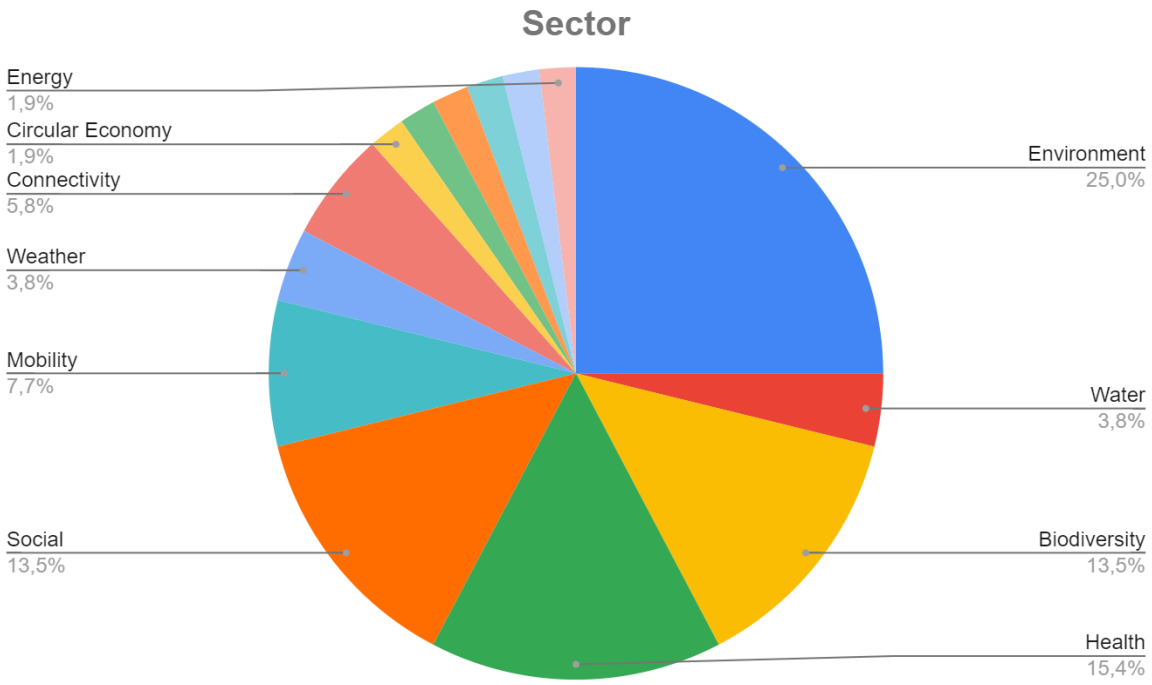


Figure 2: Distribution by sector

Project governance

The second distinguishing category in the taxonomy reflects the actual governance of the project or initiative. In this report, we define governance as the accountability framework for the project or initiative at stake. This includes, for the scope of this report, two main aspects. First, the actual funding structure and the leading organization (or organizations in case of consortia), and second, the actual accountability framework applied to citizens' participation in the project. The latter is tackled separately in the section below dedicated to the *Citizen Role.* With respect to the former, 68% of projects were originally financed by either the European Commission (mostly through the Horizon 2020 program) or by government agencies or actual departments (at different levels). As shown in the figure below, 18% of projects (i.e. 9) are classified as *mixed funding sources.* In these cases typically the project has started with one cycle of funding (e.g. EU) and its application has been extended through government funding.

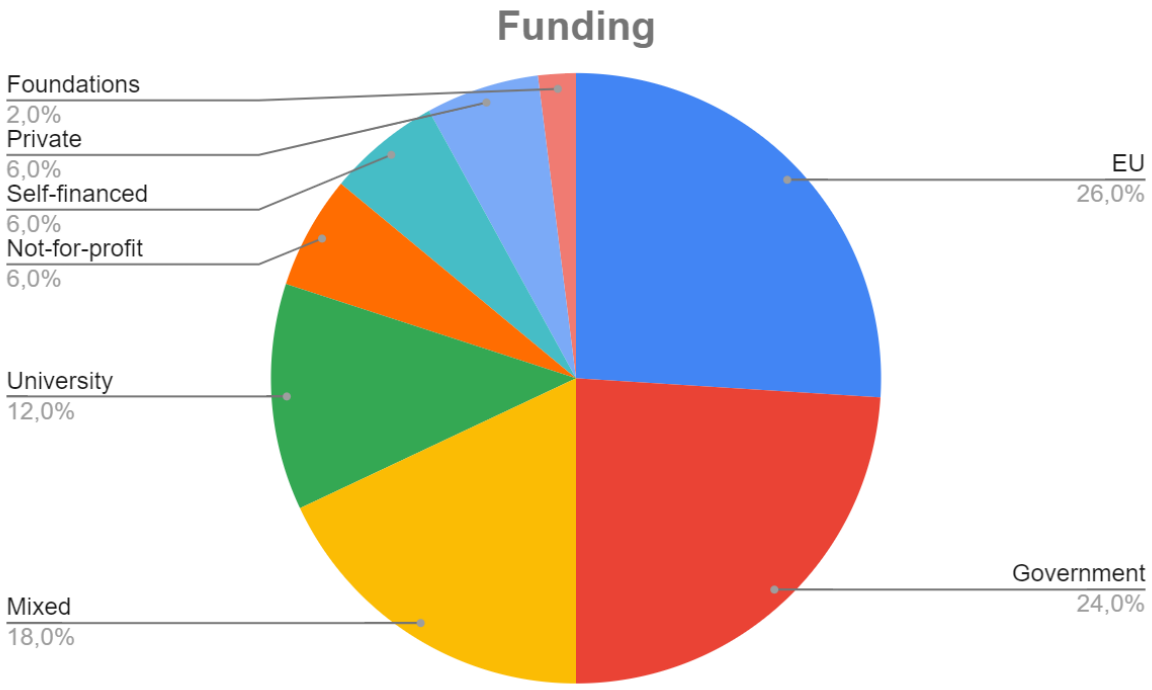


Figure 3: Funding type distribution

In relation to government-related funding, a further classification can be made across: national government agencies, local and regional Catalan government (i.e. Ajuntament de Barcelona and Generalitat de Catalunya), and local (publicly-led) research centres and agencies. At the national level, an important role in the ecosystem is played, among others, by the *Fundación Española de Ciencia y Tecnología* (FECYT) which financed (or co-funded) 10 of the 50 projects encountered. Combined, the *Ajuntament de Barcelona* and the *Generalitat,* are behind 6 projects. Finally, dedicated centers and agencies were also found to play a crucial role in this space. These typically operate in the sector and discipline where the project is situated and, importantly, represent stable entities, a key requirement for sustaining CGD-related actions over time. Examples include the *Servei Meteorològic de Catalunya* (leading and funding the *Red de Observadores Meteorológicos)*, the Centro Regulación Genómica (funding *Saca la Lengua),* the *Institut de Recerca de la Biodiversitat* (funding, among others, *BioBlitz Barris* and *Líquenes en Barcelona*).

Besides 3 projects each being funded by private organizations and not-for-profit ones, those financed by universities and those that appear to be self-financed deserve more attention. Regarding the former, while universities are keystones for many projects regardless of the source of funding, in a typical research project fashion, 6 of the mapped initiatives result to be funded by universities. Finally, 3 additional projects appeared to be self-financed. Interestingly, all these are led by communities of citizens themselves. Typically, these established communities purchase (low cost) sensing technologies and avail of their openness to generate data useful for their different causes and objectives. As elaborated more in detail below, these are typically the results of specific problems (e.g. traffic and pollution in Olot) affecting the community who generate data to both increase the granularity of understanding of their problems themselves and as a form of protest.

In terms of leading organizations, universities and consortia (often led by universities or characterized by their strong participation) combined lead more than 50% of the

projects considered. In terms of local universities involved, a variety of Catalan organizations have been found. These include, at the Catalan level, *Universidad de Barcelona,* the most represented in the sample leading 7 projects, and one project led by each of the following: *Universidad Autónoma de Barcelona, Universidad de Girona, Universidad Pompeu Fabra,* and *Universidad de Vic.*

Worth noting that *ISGlobal* in Barcelona appears to be an important pillar of the local CGD ecosystem, mostly, as their affiliation would suggest, in the health domain. Also, and even more at the national Spanish level, *Ibercivis* appears to be another not-for-profit organization that is assuming a significant leading role in this context, across domains and disciplines.

To conclude, as commented above, three projects are led by the same communities that are also responsible for the funding of the project/initiative.

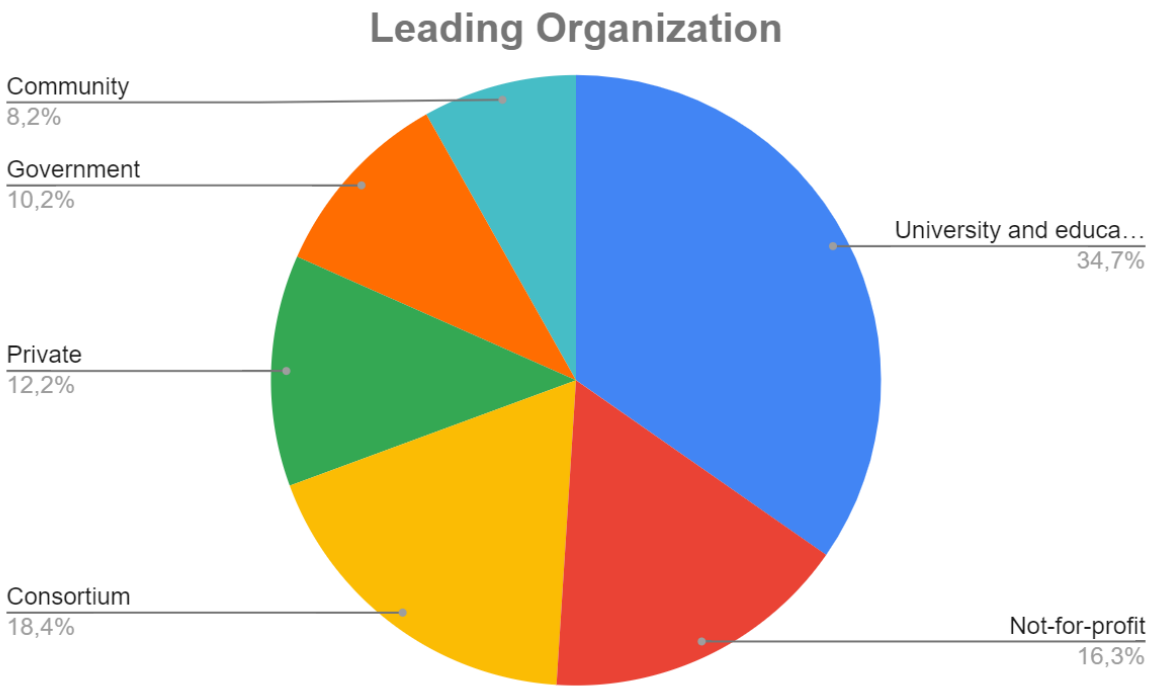


Figure 4: Leading Organizations distribution

Geographical Scope

As described above, the scope of this study is to provide some preliminary mapping of CGD-based initiatives and projects with a specific focus on Barcelona and Catalonia. However, it is important to understand whether these efforts concentrate in this area only, or if Catalonia/Barcelona are one leg of a wider project or initiative. In particular, we identified six clusters of project when taking into account their geographical scope:

1. Barcelona (18 projects): these endeavors focus either on a specific neighbourhood in the city (e.g. *BioBlitz Barris, Fotoveu Gotic)* or on the city as a whole (e.g. *Projecte Endémic, FoodMapping, Juegos Para el Cambio Social)*.
2. Barcelona and EU (10 projects): these are typically European funded projects which usually include distributed interventions across geographically spread locations and contexts. The purpose is usually to foster generalizability of the results by demonstrating their value across different socio-political-economic contexts (e.g. *WeCount, CitieS-Health).*

3. Catalonia (5 projects): of these five further projects, two are led by governmental agencies, two by universities in the region (i.e. *Universidad de Girona* and *Universidad de Vic*) and one by a local community (in Olot).
4. Spain and Spain and EU (9 and 3 projects respectively): even though the scope is posed at the national level (or like in the case of Barcelona and EU above - as part of a multi-pilot EU project across different countries) these projects have been taken into account either because their are led (e.g. *Saca la Lengua*) by or included in the consortium local partners (e.g. *Generation Solar* led by *Universitat Pompeu Fabra*).
5. Global (5 projects): this final category includes projects that have a global focus, including contributions from Catalonia and/or Barcelona. These projects are represented by a proven and tested IT infrastructures - typically a visualization GIS platform fed by individual distributed inputs, either from sensors or from dedicated mobile applications - and are now used widely across countries. Examples include the *Smart Citizen Kit, Odour Collect, Wheelmap,* among others.

Status and Duration

The final category with respect to the project's main characteristics emerges to be related with the timing and status of the interventions.

First, considering the status of the projects selected, 33 appear to be still active (some as described below with a temporary schedule dictated by the funding structure) and 14 terminated. Three additional projects were labelled as *inactive.* The main reason for this is because, while their websites suggest that the actions are still undergoing, there is no evidence of activities conducted in the last few years. It is noted that the pandemic situation might have affected these projects' ability to continue working on schedule on their tasks and visions.

With respect to the duration of the project, This is important to consider to distinguish between those efforts that are **temporary** in nature, and therefore focus more on specific tasks, hypothesis or research questions, and those that act more like a **permanent** source of data, information and knowledge. In this way, we identified three categories: permanent (42%), temporary (52%) and periodic (4%). With respect to the latter, those considered periodic are somewhat permanent but their actions are based on seasonal efforts (e.g. *Planttes,* concentrating its efforts during the time of high risks of pollen allergy) or annual ones (e.g. *Vigilantes del Aire*).

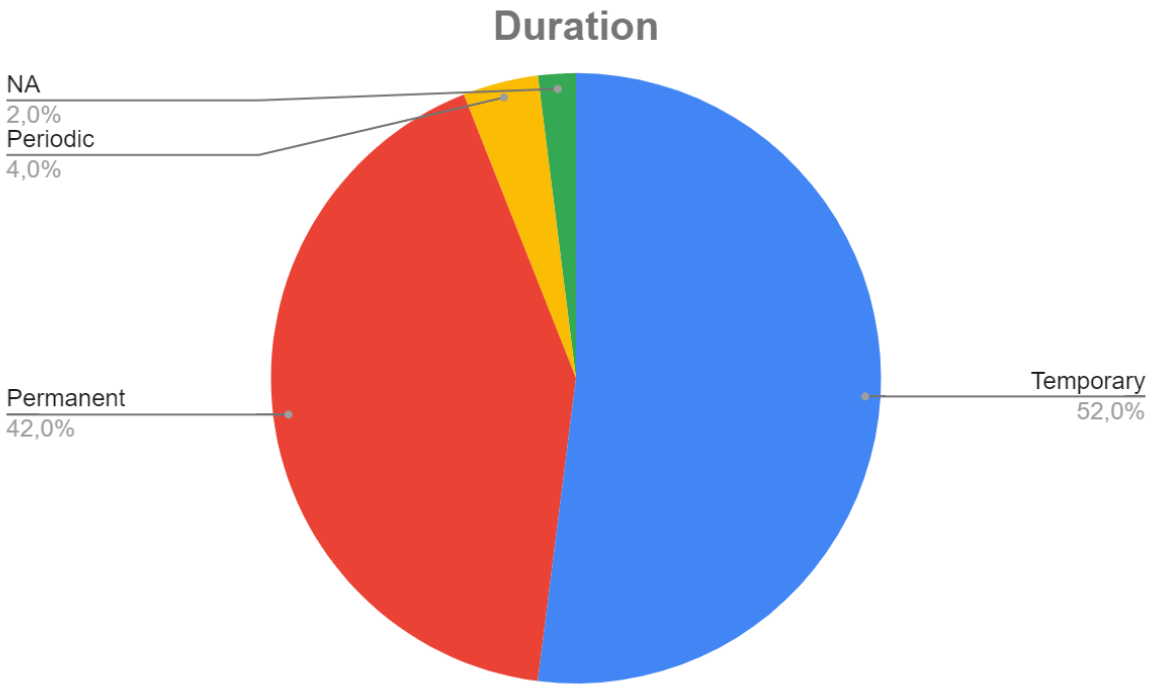


Figure 5: Projects' Duration distribution

Besides the focus - i.e. temporary projects oriented towards experimentation and research purposes, and permanent projects oriented towards sustainable information and data sources - a few considerations can be made based on this distinction.

First, it is interesting to reflect on the fact that achieving a permanent status is often the openly declared objective of certain experimentation-based projects - typically those funded by the EU. However, beyond the funded period, these typically lack the resources and business models to continue their efforts. From another perspective, those that appear as permanent projects are sustained by stable entities (e.g. public sector agencies like *El Servei Meteorològic de Catalunya*) or commercial business models (e.g. *Smart Citizen Kit*), or established communities (e.g. *Fumuts Ros de Olano, Olot*).

Second, 71% and 100% of projects in the contexts of health and connectivity respectively are temporary. This, in first approximation, suggests that experiment-based projects tend to be temporary. It is also noted that most of these do not include objectives around future sustainability (as argued in the previous paragraph) but complete a full cycle in a given timeframe.

Third, all projects considered in the biodiversity discipline and all those community-led self-financed initiatives are permanent.

Fourth, connected to the previous point, it appears that all projects that avail of external platforms (e.g. biodiversity projects connected to the global platform *iNaturalist,* or all community projects leveraging systems like *Telraam, the Smart Citizen Kit, Twitter*) tend to be permanent. This is probably related to the fact that update and IT maintenance costs are not covered by the project itself. This has the potentiality of becoming sustainable following an initial development of a socio-cultural infrastructure to enable the consistent collection and processing of relevant data over time.

Citizen Roles

Partially inspired by the citizen science discipline, and particularly by the literature focusing on different types of citizen science projects based on the phases where citizens assume active roles, we defined three collectively exhaustive categories. These reflect three different levels of citizens' accountability with respect to the project or initiative.

1. **Participate / Contribute**: this first layer, by far the most common among the CGD-based initiatives considered, includes various levels of active involvement which, however, are limited to data contributions to tasks or research processes that are fixed and pre-defined by those organizations leading or funding the project itself. In other words, these projects typically leverage data collected by citizens for given purposes. It is noted that these participations and contributions may be underpinned by completely different levels of granularity of the tasks. As an example of low task granularity, the project *Cobertura Mòbil,* leverages GPS data provided by citizens who simply agree once to give access to this information. On the contrary, some projects within biodiversity demand higher efforts for contributing, such as taking pictures, classifying them, adding text-based characteristics when uploading them etc.

2. **Design**: at this second level, in addition to participating and contributing, citizens are typically responsible for the design (or more often the co-design) of certain aspects of the project or initiative, which are however governed by other bodies or consortia. This is the case of all citizen science projects included in our sample, whereby citizens typically actively participate in the design and selection of research questions and the co-design of technologies and/or other aspects of the defined intervention. Other projects included in this cluster include the citizens' responsibility of taking some critical decisions during the planning and implementation phases. For example, in *Mapa Sonoro Barcelona* citizens decide where measurements should be taken from.

3. **Govern**: finally, at the highest level of accountability, projects or initiatives may be owned and fully governed by citizens themselves. In this study, this was the case of those actions led and financed by the communities themselves.

Concluding, to provide an overview across these three roles that citizens can assume, or three growing levels of accountability, the figure below shows how the vast majority of projects considered can be positioned within the participate / contribute cluster (76%), followed by the design one (18%) and by those fully governed by citizens.

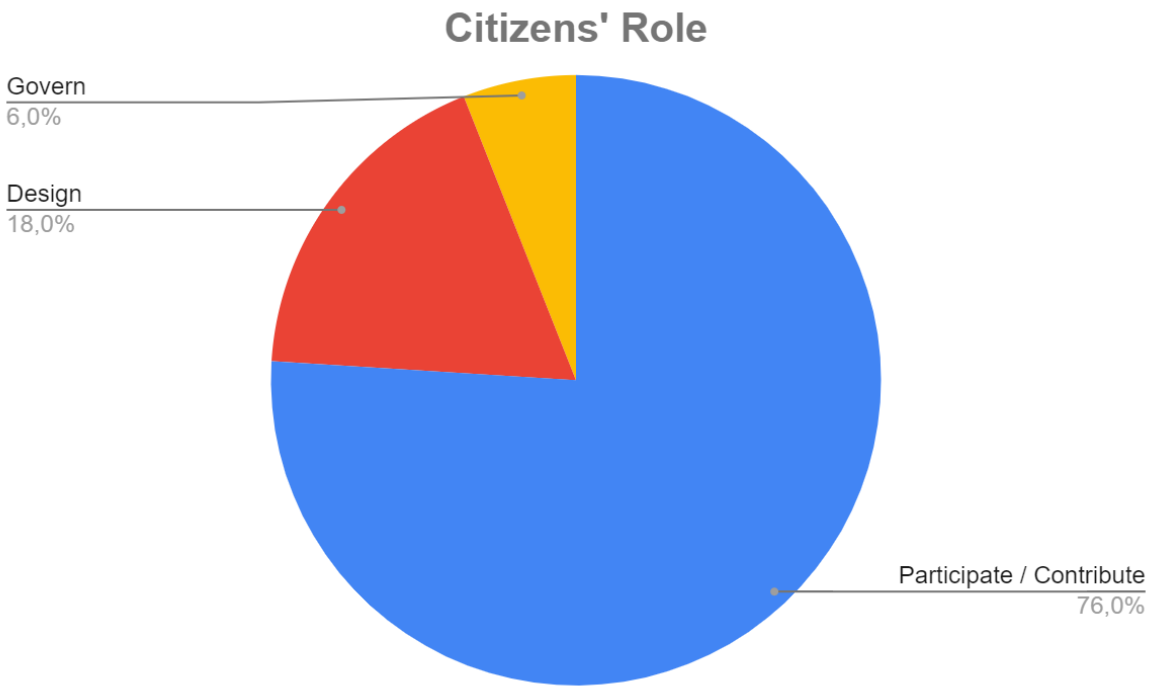


Figure 6: Citizen Role distribution

2.2.2 Data

As the second overarching category, in this space we explored how distinct projects differ in terms of what data is collected and how. These two elements were broken down into two categories including two variables each. First, we reflect on the contribution type in terms of both the task performed by citizens to enable the gathering of the relevant data and the type of data collected. Second, we reflect on the data collection in terms of both what instrument or device is leveraged as well as the timeliness of the data.

Contribution type

As mentioned above, this dimension considers the task performed by citizens and the type of data collected. These two elements are connected in the sense that certain tasks allow for certain data to be collected. The two main tasks (and the related data) are defined as Access and Production, and are tackled separately below.

Access: in this first category, the task is defined generally as giving/enabling/allowing access to data generated by other instruments or devices. These can be divided into two types of sensing technologies that differ in nature as well as in the resulting task from the participating citizen. First, we identified **ambient sensors**, defined in this taxonomy as sensing technologies that gather data about an environment and its conditions. The condition is that the technology is separate from the personal devices of an individual (e.g. her or his smartphones), i.e. has a separate hardware product. The second, instead, refers to access given to **individual sensors**. In these cases data collection usually follows citizens' permissions to provide access to individual data provided by, for example, sensing technologies ingrained into our everyday digital devices (e.g. location data from smartphones). The table below provides an outline of the data that is typically

collected with each approach within this Access category as well as some relevant examples from the sample.

Access		
Contribution Type	**Typical data collected**	**Examples of projects from the sample**
Ambient sensors	Noise, air quality, traffic and mobility	*Making Sense, Decode, Smart Citizen Kit, Olot Community, Fumuts Ros de Olano, Mapa sonoro Barcelona.*
Individual sensors	GPS, health data, web traffic data	*Bee Path, Cobertura Móvil, CitieS-Health, Salus.coop.*

Table 1: Access Contribution Type

As a general observation, projects based on individual sensors tend to be temporary (as they are typically based on short time access to individual sensitive information - e.g. GPS location 24/7 in the case of *Beepath,* or health data in the case of *Salus.Coop*). On the contrary, ambient sensors-based actions tend to be more enduring over time and often establish objectives around becoming the reference approach, standard, and platform, for example in the context of monitoring certain environmental variables like air quality or noise.

Production: differently from the previous, projects labeled within the Production category entail a higher commitment for generating the relevant data as it is produced in some form by the participating citizen. The typology emerging from the bottom up analysis includes at least four different types of data produced by citizens.

First, we identified data contributions in terms of **Documental Observations**. These are typically in the format of photos, videos, text, or audio content reporting the situation of a specific phenomenon of interest or a problem at stake. The most common example refers to biodiversity projects, often based on citizens submitting evidence of existence and/or status of plants or animal species. Other examples focus more on the power of photovoice methodology to report and subsequently understand other important elements of urban environments such as their food environments and how they are perceived and experienced (*Food Mapping*) or the mapping of the diverse degrees of accessibility (*Wheelmap*).

The second type refers to **survey data**, i.e. structured data provided by citizens in the form of a questionnaire. While only one project has been found to solely rely on this source of evidence, several other initiatives leverage this method of data collection as a complementary source to the main one obtained. As an example from the latter, projects relying on ambient sensors leverage complementary qualitative and quantitative information to enable a better interpretation of the hard data produced by the sensor. For example, data about noise may be influenced by a public protest organized in a given day passing by the sensor's location. Gathering this complementary data allows capturing these exceptional circumstances and thus a more reliable and meaningful interpretation of the sensor's data.

As the third type, we identified **physical samples**. In this category, the provision of data is indirect in the sense that the actual data is produced usually by a scientific and/or

technical group that generates data from the analysis of physical samples provided by citizens. Across the 50 projects considered in this study, 7 rely on physical samples of various kinds. Examples include strawberry plants as air quality biosensors (*Vigilantes del Aire*), diffusion tubes (*Citie-Health*), tap water samples (*Aigua BCN*), samples of microplastic (*Paddle Surfing for Science - PlastiPlancton BCN*), or even saliva's samples (*Saca la Lengua*).

Finally, all remaining data types were classified, for completeness, under the category Other. Examples here are varied, ranging from providing structured analysis results through gaming like in the case of *Genigma* (to investigate alterations of cancerogen genomas), training an urban planning related algorithm (*Arturo*), co-creating data licenses (*Decode*), stories (*#Cuentalo*), or new scenarios for public policies (e.g. *HOOP, SEEDS, Juegos para el Cambio Social*).

In the same fashion as for the previous element, a summary table is provided below.

Production		
Contribution Type	**Typical data collected**	**Examples of projects from the sample**
Documental observations	photos, videos, text, audio.	*Planttes, Wheelmap, BioBlitz, Food Mapping*
Survey data	responses to pre-defined and structured multiple choice questions	*Observatorio Ciudadano de la Sequía*
Physical samples	concrete material (e.g. microplastic), analog and bio sensors, health-related samples	*Pescadors de Plastic, Vigilantes del Aire, xAire, Projecte Endemic, Saca la lengua, Aigua BCN*
Other	stories, scenarios, analysis outputs, reports	*Genigma, Arturo, CSI-COP, INSpire*

Table 2: Production Contribution Type

The actual distribution of these types of data collected across the 50 projects considered is provided in the following figure. As shown, Documental Observations are the most common in the sample (43.6%), followed by physical samples (17.9%), and survey data (10.3%), while 28.2% within the Production category leverage other types of data collection. As a general observation, all projects in the biodiversity sector provide Documental Observations, although in different formats depending on the initiative.

Finally, it is noted that, if compared to projects based on Access, those based on Production entail a higher commitment and time for participating citizens to contribute (e.g. installing a sensor once versus documenting a neighborhood food environment through photography and text submissions). This leads to a reflection on the need to dedicate significant effort to community maintenance and sustainability (in addition to community building also required in Access projects) in order to establish a valuable and meaningful data gathering process. In this way, it can be argued that the nature and type of data collection is linked with the actual governance of the project whereby stable

entities and permanent structures are more likely to achieve a long-term sustainability of the community of citizens undertaking data generation tasks.

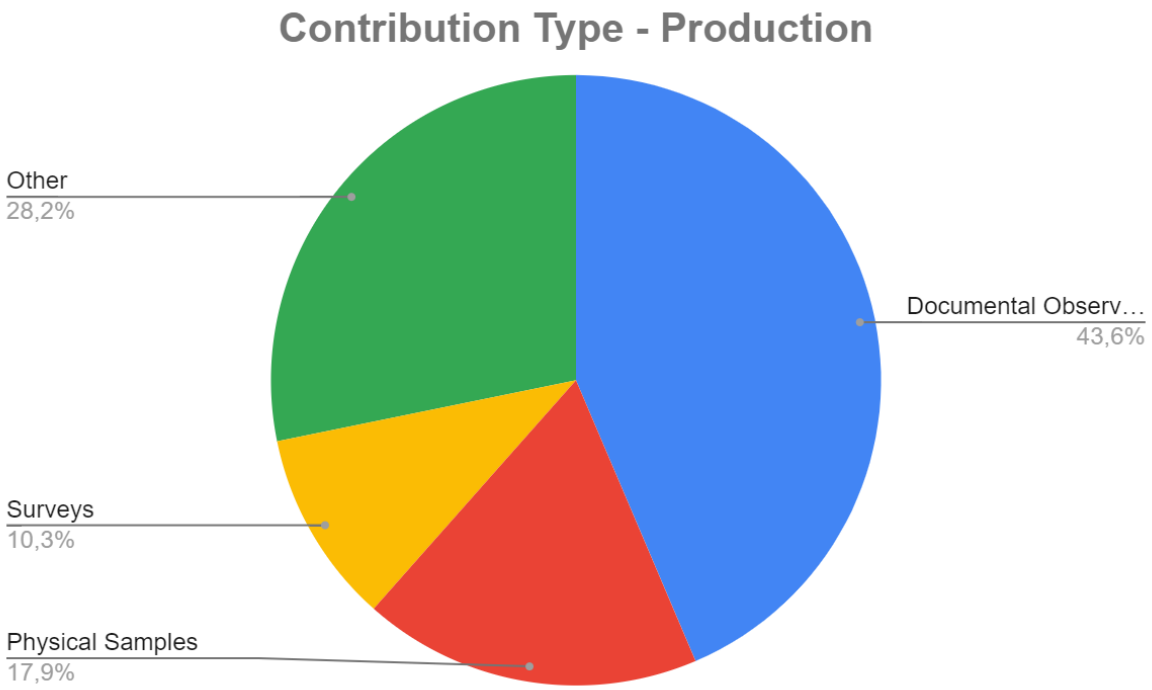


Figure 7: Contribution Type - Production distribution

Overall, the majority of projects were found to belong primarily to the Production category (39 project - i.e. 78%), while 12 projects (22%) belonged to the Access category[5]. In addition to the trends and tendencies highlighted above, it is worth noting that within the Access category of the total of 8 projects relying on ambient sensors, 5 focus on noise pollution, 4 on air quality, and 3 on traffic and mobility. Indeed, the tendency observed is that individual projects relying on access and specifically ambient sensors tend to incorporate different focuses (e.g. *Fumuts Ros de Olano* gathers data about the three phenomena). This suggests that once a community is formed around ambient sensors-related matters, adding additional sensors and focuses to the initiative results to be facilitated as each addition regards one further technology on top of an existing socio-cultural infrastructure, i.e. the community itself and its governing and communication mechanisms.

Data collection

This second category within the Data dimension, reflects how the different projects in the sample differ in terms of the actual data collection tools employed and the timeliness of the data gathered and analysed. These two elements are tackled separately below.

With respect to the actual tools used for data collection, we identified seven different tools leveraged across projects.

1. **Mobile application**: in total, 17 of the 50 projects considered leverage dedicated mobile applications for gathering and structuring CGD. Interestingly, all projects

[5] It is noted that one project has been labelled as both access and production as it leverages multiple and diverse types of data and data collection instruments.

relying on mobile applications belong to the category Participate / Contribute when looking at the Citizens' Role in the project. The vast majority of these are connected to platforms where the various data collected from the ensemble of citizens is integrated and visualised. Depending on the nature of the application designed, these can accommodate different types of data ranging from Documental Observations both in the form of episodes of interest (e.g. *Odour Collect* in the case of odour pollution episodes) as well as photography/video/text (e.g. biodiversity projects like *BioBltz, Observadores del Mar*, *Líquenes en BCN*), to Individual Sensors (e.g. *Salus.Coop, Cobertura Móbil*).

2. **Web application**: to a lesser extent (i.e. 6 projects), web applications are also used to gather CGD. These typically substitute mobile applications and are based on data input through a web browser. The webpage is typically designed to facilitate the submission of structured data through a survey-like form. For example, *Generation Solar* gathers information about solar panels installed at people's places through a web form where contributors need to input a wide range of characteristics of their plants, the brands, the performance, location etc.

3. **Automated sensors**: a total of 8 projects leverage automated sensors for data collection. While these can all be classified as ambient sensors, the opposite is not always true (i.e. some ambient sensors, e.g. diffusion tubes in *Cities-Health,* are not automated sensors). These cover typically three elements: noise, air quality and mobility. In most cases, especially those led by communities, low cost sensing technologies are used, e.g. those based on Raspberry Pi or Arduino computing devices.

4. **Analog sensors**: in this area, 3 projects leveraging two analog sensors were identified: *Vigilantes del Aire, CitieS-Health,* and *XAire.* The first uses strawberry plants as air quality biosensors. Air quality parameters are extracted from the bio-magnetic analysis of the particles that deposit on its leaves over time. The second and the third both use diffusion tubes to gather data about NO2 pollution. It is noted that in these cases the citizens generate the resulting data in an indirect manner as they deliver the analog sensor to a team of technical analysts that are in charge of their analysis and of the actual data generation.

5. **Workshops**: interestingly, in 8 projects data is collected through collective interactions and exercises, in most cases in co-creation and co-design formats. These are typically the result of organised community workshops, which can vary from those led by a scientific committee and those organised by the communities themselves. The role of citizens in these projects is elevated compared to the previous in that they typically see them also responsible for co-designing the interventions or some of their key aspects. Examples include, but are not limited to, *SEED, CoAct, Juegos para el Cambio Social, INSpire.* Results are typically in the form of structured content, like a proposal for a new public policy, scientific experiment, or an entire system like in the case of *HOOP* where citizens co-design new circular economy scenarios to harness the potential value of organic waste.

6. **Special equipment**: in somes cases (5 in total) data and evidence is collected through artefacts that are specifically designed for the purpose of the relevant project or standard equipment for collecting and storing particular physical samples. Examples of the former include *Paddle Surfing for Science and*

PlastiPlancton BCN whereby surfers are installed a little net underneath their boards to collect microplastics close to the shore areas. Examples of the latter include the equipment to collect saliva or urine samples in *Saca la Lengua* and *Aigua BCN* respectively.

7. **Other**: this last category was defined for completeness and includes all tools that did not fall within previous categories. For example, within the project *Red de Observadores Meteorológicos* granular data about weather events is gathered through dedicated communication channels (e.g. emails) established between the *Servei Metereologic de Catalunya* and the network of observers.

Overall, the distribution of these different data collection tools among the 50 projects considered is depicted in the figure below.

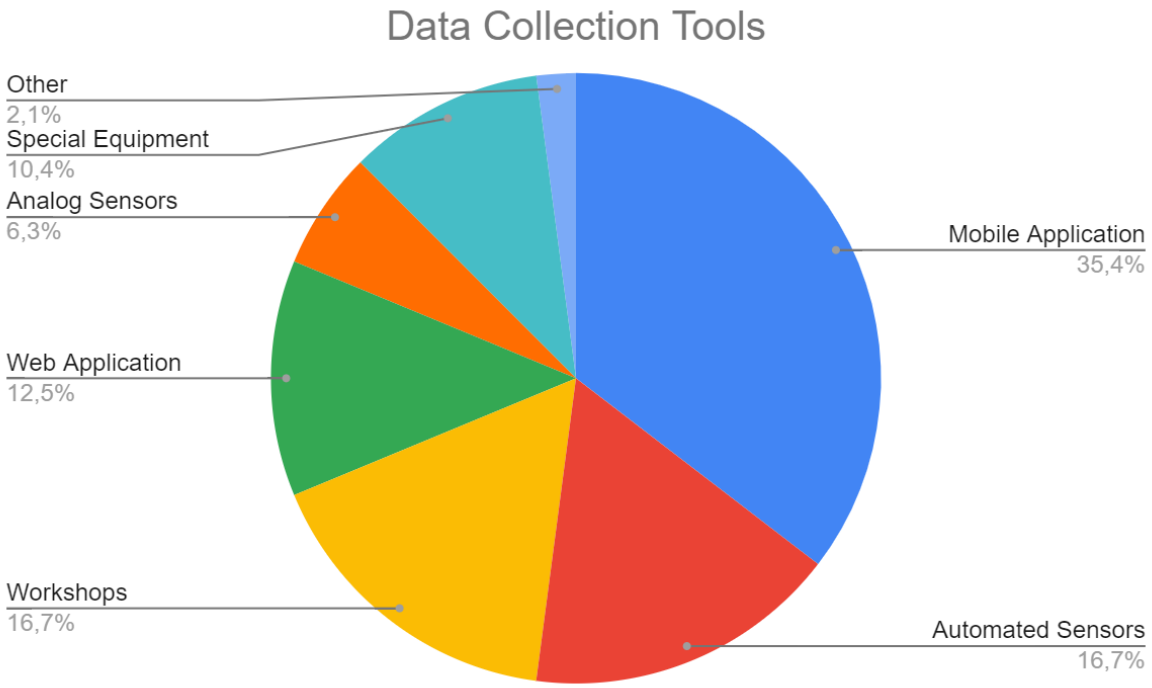


Figure 8: Data Collection Tools distribution

As the second main focus within Data Collection we reflect and distinguish CGD based on their timeliness. One of the main barriers in open (government) data more generally is to achieve a situation where data is provided in a timely-enough manner to be valid and valuable. In other words, while many advocate for data in real time, this is not always possible nor the recommended scenario. We therefore distinguished CGD based on whether these are: continuous, one-off, or periodic.

- **Continuous data**: in this first case, i.e. the case of automated sensors, data is provided in a continuous manner, independently from external events. For example, air quality and noise sensors provide data to the platform in near-real-time, as opposed to only when the situation is particularly problematic. This is also consistent with the use of this data, that is to understand certain conditions under different circumstances.

- **One-off data**: in certain cases, the continuity sought above is just not relevant and contributions are based on one-off data inputs, defined in this study as those

data provided in discreet occasions. Examples include two types. First, experiments based on single contributions such as a saliva sample in *Saca la Lengua* or tap water and urine samples in *Aigua BCN.* Second, these may be the cases of projects where data is generated and delivered only in some special occasions like in *Odour Collect* where citizens provide geo-localised information about odour episodes they experience.

- **Periodic**: finally, in some cases, data is sought under certain times of the year or exceptional conditions. In this case the data is defined as periodic as it is generated only for these timely bounded periods. As an example, the project *Planttes* seeks to monitor the incidence of allergies in a given area from gathering evidence about the status of the flowers of certain plants. This is obviously relevant mostly during Spring time.

In conclusion, taking into account the timeliness of the data can be crucial when planning integration of CGD into the existing open data initiatives and portals. The recommended exercise, however, is to start from the phenomenon of interest and to reflect on what type of data may be needed to address it, rather than from existing projects that do not always showcase best practices or optimal conditions.

2.2.3 Destination

As the third and last overarching category, the taxonomy developed in this study takes into account how different projects vary in terms of both the outputs produced and the outcomes sought and (when available) achieved. This distinction is important as outputs focus on where the data finally ends up and how it is reported, whereas outcomes reflect the actual objective and goal of what is wanted to be achieved through the CGD collected and analyzed during each project or initiative. Below, we reflect on these two separately. Finally, a reflection on what data licenses are used in the sample of projects considered is also provided.

Outputs

With respect to outputs, three general primary outputs[6] were identified in this study: (1) platforms or databases; (2) reports; (3) social media[7]. The latter was the case of one project only, i.e. *Prou Transit,* a community-led initiative dedicated to generating data about traffic, mobility, and pollution from integrating evidence from existing sources, and its reporting on a dedicated social media page (Twitter). Reports, on the contrary, were found to be quite common across the projects considered, representing the main output of 19 of these (i.e. 38%). This is the case of most projects based on citizens' indirect contribution to data generation, i.e. when the process between the citizens actions and the final results is mediated, through synthesis and/or analysis, by technical and/or scientific actors. These include for example health projects based on the collection of physical samples (e.g. of saliva or urine), the analysis of the strawberry plants' leaves, among others. A potential solution to move from reports to more reusable platforms refers to the practice of digitizing these results and making them available through more interactive and navigable datasets or visualizations. However,

[6] It is noted that several projects deliver different types of outputs. Those considered here are the most prominent ones.
[7] It is noted that outputs were found for a total of 48 of the 50 projects as for Red de Observadores Meteorológicos and Salus.Coop no specific, defined, outputs have been identified.

these practices do not seem to be widespread in the sample of projects analysed in this study. Finally, platforms and database - related outputs deserve more attention and reflection.

The main reason why we deepen on these is in the value carried by these outputs for the scope of this study. In particular, CGD (openly) available through platforms and databases substantially increases its value if compared to data presented in lengthy scientific reports. This value is centred in the potential (immediate) re-use of this data and the much easier and less time consuming process to do so in platforms, rather than from reports. Indeed, data is typically produced and made available in machine-readable manner and contains important attributes (e.g. time stamps, location accuracy etc.) as well as metadata and information to aid its navigability and understandability.

In particular, when considering platform and database outcomes, two general types were identified in the sample.

First, 3 projects' outputs are made available through platforms containing structured content. These somewhat represent an evolution from solely delivering reports as the main outputs of the initiative shift to digitizing and structuring these reports' content into structured organised platforms. Examples include *INSpire*, where a wide range of experiments are co-created and undertaken by participating citizens, or *#Cuentalo* where women's domestic violent experiences are organised, structured and reported.

Second, the remaining 25 projects (i.e. 50% of the sample) rely on outputs in the form of GIS Platforms, i.e. showcasing geo-localised CGD onto maps. Within these, the majority relies on dedicated platforms, i.e. designed and built for the specific project and thus accommodating inputs from that initiative only. Of these, for some this is their natural solution as there would not be enough justification for integrating with other solutions. For example, *Wheelmap* is unique in its focus and thus relies on its GIS platform. Others instead could potentially be integrated with other, global, solutions that are similar in scope. Examples in this space include biodiversity projects such as *RiuNet* and *Observadores del Mar.*

On the contrary, the 8 remaining projects are integrated with existing global platforms, i.e. the data generated in the specific project feeds a global, established database and platform. Across the projects considered, three global platforms emerged as being leveraged. These are:

- iNaturalist, i.e. a global biodiversity and CGD platform. In this study we have encountered several biodiversity projects that leverage these systems (i.e. also including the dedicated app for data collection), such as *Ritme Natura, Liquenes BCN* and *Bioblitz.*

- SmartCitizen.me, i.e. the platform where all the data collected globally by the Smart Citizen Kit sensors (i.e. measuring air and noise pollution) is visualised and made openly available. *Making Sense* and *Decode* are examples of projects in this space.

- Telraam, i.e. the platform where traffic information (i.e. number of vehicles, bicycles, trucks and pedestrians in transit and their speed) generated by the

Telraam sensors is visualised. Examples of projects and initiatives include *Fumuts Ros de Olano, WeCount,* and *Olot.*

As a more general reflection, some conclusions can be preliminarily drawn based on the connection between data collection tools and outputs (given the relatively small sample, these are treated as tendencies rather than established patterns). These links are shown in the figure below.

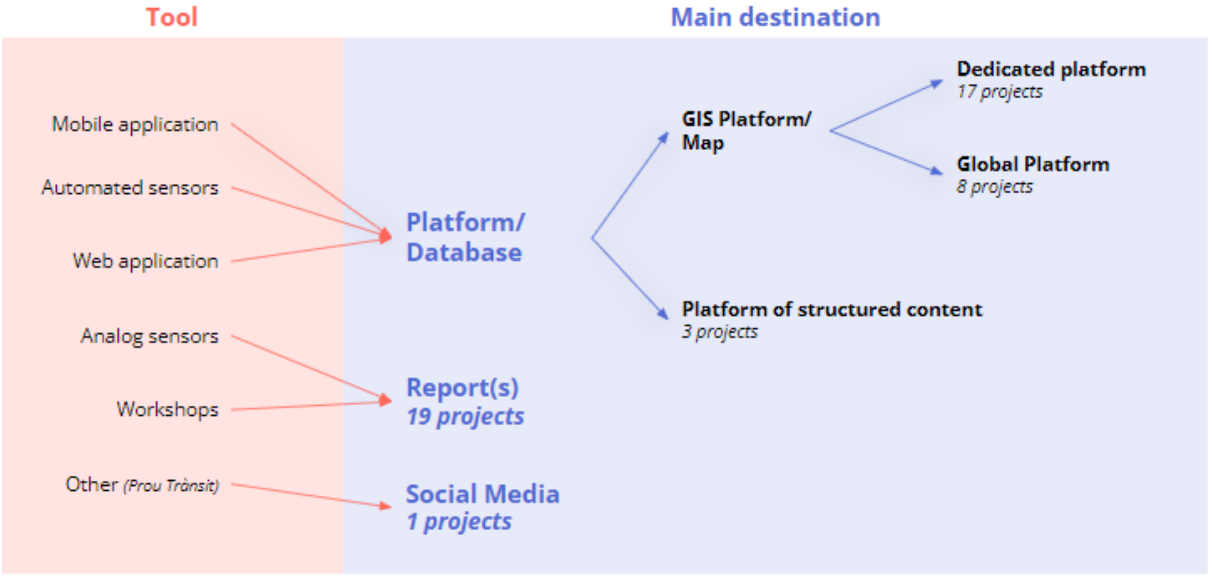


Figure 9: Data Collection Tools and Destinations

In general, we observed that patterns exist between the data collection tool employed and the final destination of the results and findings of the project. When data is gathered through dedicated mobile/web applications or through automated sensors, these are usually designed and implemented in parallel with a dedicated destination platform. What is less common is to have a dedicated app for a given project feeding into a global platform. For example, those projects whose results are provided in iNaturalist (i.e. a global platform) also rely on the iNaturalist mobile application for data collection (e.g. *Bioblitz*). It can therefore be argued that, in those cases where CGD is provided through platforms or databases, data collection instruments and destinations are part of the same information system, based on the same standard.

On the contrary, when data is collected through analog sensors (e.g. strawberry plants or diffusion tubes for air quality) or generated at workshops, the relationship between data collection tools and destinations is mediated by an intermediary step. In other words, in these cases the output is not the CGD itself, but the result of its analysis, interpretation (and sometimes manipulation and cleansing processes) typically conducted by domain expert organizations.

While it may seem trivial, this distinction has an implication on the (potential) outcomes of these CGD-based projects, especially in terms of the ability of reusing the results for different purposes. This is tackled next.

Outcomes

Regardless of the outputs, it is important to also reflect on the outcomes (or intended outcomes) of these projects. Initially, from the bottom up analysis we identified three clusters of outcomes across projects.

1. A relatively low number of projects considered, has at its core the objective of contributing to developing new or improving existing public policies, or in general assisting the work of the public sector in delivering public services. Across these projects whose aim is to **inform public policies**, there is typically a direct channel (different extents were observed from a communication channel established to the relevant public sector agency itself financing and/or leading the project) between the CGD initiative and the relevant public sector organization. In most cases, however, it appears that the evidence provided by the CGD initiative is complementary to other sources of evidence taken into account in a given phenomenon. For example, the *Red de Observadores Meteorológicos* provides the *Servei Meteorològic de Catalunya* with additional data to complement the more traditional data already collected by the agency.

2. Certain projects aim at achieving **scientific discoveries**, i.e. the main aim of the project is to better understand a phenomenon of interest through systematic and rigorous research endeavours. This cluster includes all projects in the context of health, the majority of projects in biodiversity, and in general those led by research agencies and universities.

3. As the third class, five projects in the sample aim at **raising awareness** about certain topics of interest. Typically, these projects do not have the rigor or the systematic / academic nature to be either leading to scientific discoveries or be considered as is for informing new policies. Rather, these projects aim at raising consciousness about problems affecting our environment and societies. For example, *#Cuentalo* leverages CGD for raising awareness about domestic violence to women. However, the data is somewhat unstructured and no evidence is found of either a clear purpose to target new or existing policies, or transparent rigorous research processes guiding data collection and analysis.

However, a reflection about both scientific discoveries and raising awareness led us to expand this classification to include also two elements at the intersection between these two classes of outcomes and informing public policies. In other words, some projects explicitly argue that both scientific discoveries and raised awareness about certain topics can also be seen as an intermediary step toward achieving policy impact, i.e. the basis upon which new policies can be implemented or existing ones can be improved. We therefore added two additional clusters in this taxonomy:

4. **Scientific discovery / informing public policies**: unlike projects solely dedicated to scientific endeavours, this cluster encompasses initiatives that include in their structures the transfer of scientific knowledge to policy makers to be considered as a new source of evidence for their work. This happens in different ways. For example, in EU-funded projects sometimes these links are explicitly mentioned within its core objectives (e.g. *WeCount*). In other cases, policy makers are formally part of the consortium (usually not leading it otherwise projects would tend towards a direct contribution to policy).

5. **Raising awareness / informing public policies**: like the previous cluster, projects in this group ingrain a link with policy makers (or explicitly attempt connecting with them). In the sample considered in this study, these can be further divided in two categories. First, we identified community-led initiatives where the typical approach is to first create a community to gather and co-create actions around a problem that affects them, i.e. a matter of their concern (e.g. air quality and noise pollution in *Fumuts Ros de Olano* and *Olot*). Action in these cases is about generating CGD as evidence for the severity, magnitude and diffusion of the problem experienced. This evidence is subsequently leveraged to enable informed confrontation with policy makers (often in the form of protests or activism). Second, certain projects start small and aim at raising awareness about phenomena that are of interest to the civil society or cohorts within it. Over time, driven by proven usefulness of the CGD-based solution, these evolve into valuable systems that can potentially aim at informing policy makers. An example that falls in this second category is *Wheelmap.* While it started as a contributory tool to assist people with disabilities in finding accessible places, its trajectory is now towards informing relevant government departments about where investments in accessibility are needed.

Overall, these five different layers of outcomes are represented in the figure below. As shown, a significant proportion of projects from the sample (i.e. 40%) focus on scientific discoveries and 24% at the intersection between scientific discovery and policy making, while they are equally distributed across the three remaining categories.

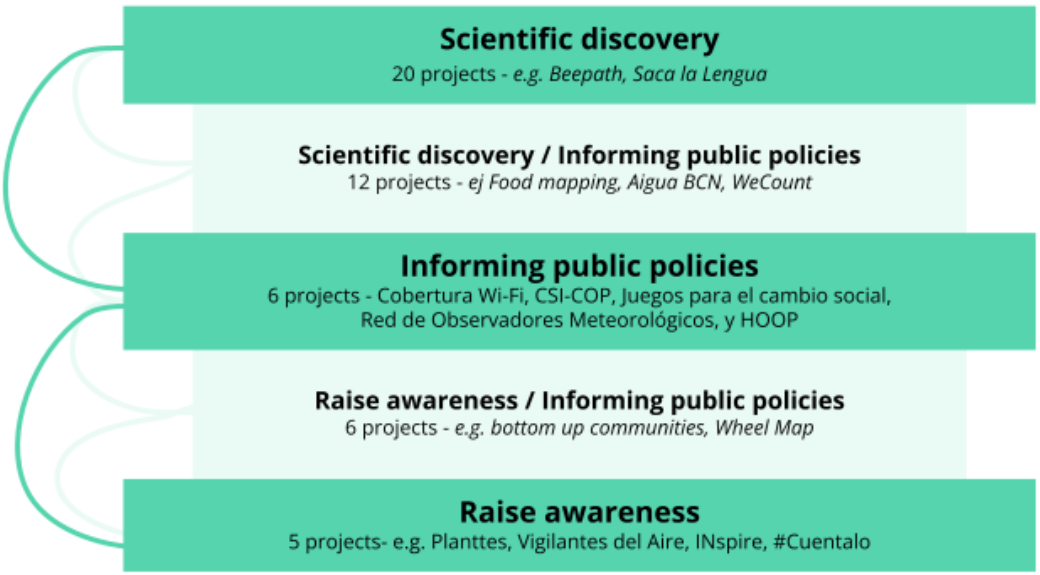


Figure 10: Outcomes' Classification

Data license

When tackling re-usability of the data and its degree of openness, it is important to consider the license associated with the datasets being produced within each project. However, it must be taken into account that among the 50 projects considered, only a

subset provides datasets as an output (see above). From another angle, while also reports can be associated with a license (open or not), we consider only datasets in these reflections.

In total, we found 22 projects whose output are structured datasets to which an (open) license has been assigned. Collectively, these datasets adhere to a total of ten different licenses. An overview of these is provided in the table below.

License	**Brief description**	**Projects in the sample**
Creative Commons By Attribution (CC-BY)	Full reuse and manipulation only if credits are given to the author.	4
CC-BY-NC (Non-Commercial)	CC-BY + Full reuse not for commercial purposes	5
CC-BY-SA (Share-alike)	CC-BY + Full reuse and manipulation keeping the original license	2
CC-BY-NC-SA	The previous three combined	4
CC-Copyleft	Same conditions as CC-BY-SA for open software	1
EUPL (European Union Public License)	Same conditions as CC-BY-SA for open software	1
GNU Lesser General Public License (LGPL)	Same conditions as CC-BY-SA for open software without the need to release the source code in case of integration of LGPL in new, proprietary, software.	1
Open Database License (ODbL)	Same conditions as CC-BY-SA	1
Salus Common Good	Co-created user-driven license for personal data.	2

Table 3: Open Data Licenses overview

As shown in the table above, Creative Commons in their (NC and SA) variations are the most commonly used. The choice of which license to adopt seems to be dependent on individual choices (or contractual restriction with the funding body/bodies) and less on

the type of project. In other words, we did not identify any preliminary pattern that would suggest clear links between the type of data collected, the tools leveraged, or the sectors in which the projects operate, and the license adopted.

3. Reflections and Conclusions

This study presented an emerging taxonomy of CGD-based initiatives and projects to further understand this complex and multifaceted ecosystem. This is enriched by the generation of a living document dedicated to the mapping of existing and past projects within the taxonomy itself. This mapping exercise will be continuously updated as a living repository of initiatives, mainly focusing on the Catalan context.

This study showed in particular the variety of possible approaches to CGD from three, interrelated dimensions: the project, the data itself and its destination. Each of these three dimensions has been unbundled and reflected upon in this report. In particular, we reflected and outlined how different projects differ in terms of their: (1) governance structure; (2) citizens' roles; (3) data structures and attributes; (4) data contribution type; (5) data collection tool(s); (6) outputs; (7) outcomes; and (8) their degree and type of openness (when in place).

When reflecting overall it is useful to choose a starting point related to how, ideally, an optimal or successful project would be positioned in the taxonomy. However, this analysis showed that three distinct classes of outcomes can be identified, i.e. scientific discovery, raising awareness, and informing public policies. Both the first and the second appear to be less complex if compared to a scenario whereby the outputs are adopted by complex public structure to clearly contribute to new policy development and/or to the improvement of existing ones.

This simple reflection leads to an important argument that is that CGD alone appears mainly to be, at the current technological-social-cultural-political conditions, a powerful instrument to enable a more granular or accurate understanding of a given issue. This process is usually underpinned by cycles of understanding, investigating, and coherent reporting, consistent with the focus of systematic research endeavours (typically those leading to scientific discoveries) and communication campaigns (typically those leading to raising awareness). From a more contextual perspective, this analysis shows how CGD can be leveraged to further understand a given issue, at a greater level of granularity and/or to add attributions to existing data thus improving its quality. In both cases, CGD is leveraged to inform the need for a more solid and rigorous data collection effort in certain areas. Making the last step, i.e. bridging to the policy level, is less common in the sample considered.

One way to further unpack this problem is to consider two extreme ways of CGD provision to the public sector: supply-pushed and demand-pulled. In particular, this analysis shows the greater effectiveness of CGD initiative in influencing policy making processes, when data is demanded by its final users (i.e. policy makers), if compared to the vast majority of situations where data is supplied to them. The problem usually does not lie in the fact that the data is of low value or quality. Rather, the process of positioning this data within the existing public sector infrastructure is problematic from a variety of standpoints. For example, data is generated about a phenomenon without taking into account what data already exists in this respect, and without planning how CGD can be integrated into these existing datasets (or practices). Without a specific demand, or a dedicated planning exercise from the beginning, public sector agencies often find themselves swamped with new data, of different standards and formats, with

different associated licenses, and, in some cases, where sustainable provision over time can not be ensured (see more below). As another example, often considering new sources of data (and thus of information and knowledge about a phenomenon), implies a change of working to reap the full potential of CGD this data by public sector people and systems. In this way, enabling and managing change in the public sector is a well acknowledged issue within and beyond the public management literature.

The recommendation is therefore to **strengthen the relationship between supply-pushed and demand-pulled** approaches. This means achieving alignment between citizens' motivations to allocate effort and commitment to (produce or give access to) CGD and the specific need in terms of policy making and/or improvement. Policy dialogues, hackathons, datathons and other contests, and, most of all, co-created participatory approaches, are the main instruments found in this study to address these challenges.

The nature of the phenomenon of interest then dictates the appropriate model to be followed. Also considering aspects such as duration of the project (i.e. temporary, permanent, or periodic), two extremes can be identified: (1) **holistic**; and (2) **ad-hoc**. The former identifies a situation whereby the CGD resulting from a project or initiative feeds into policy making mechanisms or databases in a continuous manner. An example could be the integration of CGD about a continuous phenomenon (see also timeliness of the data above) like air or noise pollution into the existing Open Government Data (OGD) portals. In this case, CGD is typically integrated in OGD once thresholds for methodological rigor and sustainability are met (see below). This reflection suggests the suitability of Permanent projects for addressing scenarios that require a holistic approach. On the contrary, the latter typically focus on supporting or enabling a more informed decision making process with respect to isolated decisions (or demands). In this way, Temporary projects appear to be the best fit.

This idea can be extended through reflecting further on the **relationship between government agencies and citizens** (communities), especially in terms of governance of the project. While an optimal configuration across the board does not exist (it may depend on several other factors such as the sector, the level of maturity of existing policies in the specific context, holistic or ad-hoc approaches, among many others), three main models can be outlined.

First, governments themselves may lead a project or intervention when seeking to incorporate CGD for their consideration and use (i.e. the typical demand-pulled approach to CGD). These cases are not common in our sample, but it can be argued that a gap exists between data users and producers, whose engagement and empowerment can be essential to generate data of enough accuracy, accessibility, interpretability, and validity more generally. To establish these conditions, the evidence collected in this study suggests that formalising the relationship with the community of citizens is a potential way to address some of these challenges (e.g. *Red de Observadores Meteorológicos*). However, to make the CGD useful (i.e. the project's Output) its inclusion may be the result of certain cleansing and manipulation processes. These, while improving the datasets for their intended use, may lead to situations where questions may be raised about the integrity of the original CGD.

Second, we identified at least three projects where the efforts are led by the communities themselves (i.e. one typical supply-pushed approach to CGD). These often

tend to allocate effort and resources to raise awareness and inform policy makers about an issue that they experience and that affects them in terms of their quality of life. However, as shown through the community-led examples in this study (e.g. *Fumuts Ros de Olano, Olot*), their ability to reach outcomes beyond raising awareness appears to be limited. Besides the lack of alignment explained above, the process of bridging to the policy level may be inhibited by limited skills (both technical and domain specific), lack of human, IT, and financial resources, and the usual lack of sustainable (business and funding) models to fulfill the need of producing timely-relevant, reliable and valid data about a phenomenon.

Third, collaborative models seem to address several of these challenges, although establishing and maintaining these collaborations may be more time and resource consuming. These configurations are expected to address problems of community-led initiatives in terms of provision of various levels of skills and sharing of time and resource burdens across the actors involved. Potential questions about legitimacy and integrity emerging in government-led models could be solved by undertaking participatory and common decision making processes for both sides to appreciate (1) how to preserve the integrity of the CGD and (2) how this should be interpreted and/or manipulated for it to be usable and valuable.

As mentioned above, **sustainability of the data provision** can also act as a barrier inhibiting the achievement of the full potential of CGD for innovation (within and beyond the policy making domain). Certain phenomena, especially those dynamic in nature such as air and noise pollution and mobility, require continuous (or periodic) data provision for it to be of any use. This is particularly relevant for Temporary projects dealing with Continuous data (e.g. *WeCount*) which typically lack the resources to sustain the CGD initiative beyond the funding period. The extant literature promotes approaches to transfer the socio-technical infrastructure nurtured and created during the project to entities that are more stable and enduring over time, ideally the actual final users of the projects' outputs. In Barcelona, the *Oficina de Ciencia Ciudadana*, schools and the universities are currently the primary actors in this way. Full adoption by actual government departments appears not to be common in the sample considered. The main reasons identified revolve once again around lack of resources, the strict rules currently regulating the procurement of services (and data) to public agencies, and some still open questions on how to ensure quality and accuracy of the data provided.

Collectively, these trends and challenges suggest focusing future research efforts across three interrelated elements: **technological, legal, and social**.

First, from a technological standpoint, we highlight a need to rethink the current architectures and to incorporate emerging trends keeping in mind the objective of increasing usability of the CGD, i.e. increasing its quality and trust from all parties. In this direction, we distinguish two connected elements: (1) considering emerging technologies like Artificial Intelligence (AI) and Distributed Ledger Technologies (DLTs); and (2) foster the implementation of decentralised data management mechanisms and practices. With respect to the former, on the one hand, DLTs hold the promise of addressing existing challenges related to both risks to privacy and security and to potential situations where integrity of the data may be compromised. The idea behind these developments is that citizens may want to contribute CGD to a given entity without revealing their identity while demonstrating that they are entitled to contribute such data (e.g. they are residents of a city or trusted data providers, etc.). On the other hand, AI was found to aid

the analytical capabilities of those projects in the sample relying on the *iNaturalist* IT architecture and infrastructure. Although not clearly present in the project considered, other applications of AI are outlined in terms of (Ceccaroni et al., 2019)[8]: enabling adaptive management and orchestration of citizens and their communities in their effort in generating CGD; provide an increased level of personalisation of the citizens' experiences in CGD ecosystems, through for example leveraging customised incentive mechanisms to trigger everyone's motivation to participate; and providing customised training capabilities (in different languages) thus finally improving data quality, which remains a key challenge in this space. With respect to decentralised data management, new practices are currently being established to enable individuals to integrate their own data and store it securely in decentralised databases, which can be understood as personal web servers for one's data. These would increase a more informed and rightful participation of citizens where they are empowered to decide what data to share, under what conditions and with whom in a seamless manner. Solid[9] and MyData[10] are some examples of current efforts devoted to these purposes. However, to further establish such decentralised mechanisms, we advocate for more research in the domain of decentralised identities, i.e. the underpinning decentralised and trustworthy foundations for enabling decentralised data management consistently. While advancements have been made in the field of electronic transactions within public services at the EU level, e.g. the eIDAS (electronic IDentification Authentication and trust Services)[11], also sometimes including blockchain technologies like in the case of the more recent EBSI initiative[12], the discourse with respect to an integrated decentralised personal identification system is still in its infancy. In this case, we refer to decentralization of identity as the technological and legal (see below) ability of a given system to shift data governance and management control from a central body (or system) to a distributed network whereby the individual has the ultimate authority to control her or his own data and its access rights.

These reflections on the future of IT infrastructures upon which CGD efforts are undertaken, lead to the need to reflect on how the legal and rights-related landscape should evolve in parallel. The key focus is on establishing the right for people to donate data (that they produce or give access to) that is used at the time of taking decisions around issues affecting them. Also, consistent with the decentralization shift argued in the previous paragraph, a citizen may want to apply specific privacy enhancing licenses over this data for it to be shared under conditions that she or he establishes. *Salus.Coop* and *Decode* represent key examples in this way from the sample considered in this study. However, we recommend further exploring the future of licensing in order to trial co-created ones for common good (e.g. *Salus.Coop*) across different disciplines to enable an in-depth understanding of the current challenges and barriers for achieving this vision across domains.

Furthermore, we argue that to fully establish and routinize these new decentralised IT architectures and licensing systems, appropriate social infrastructures should also be designed and implemented. An example of forefront application in this way is

---

[8] Ceccaroni, L., Bibby, J., Roger, E., Flemons, P., Michael, K., Fagan, L. and Oliver, J.L., 2019. Opportunities and Risks for Citizen Science in the Age of Artificial Intelligence. Citizen Science: Theory and Practice, 4(1).
[9] https://solidproject.org/
[10] https://mydata.org/
[11] https://administracionelectronica.gob.es/ctt/eidas#.YcHWomjMLIU
[12] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/EBSI

represented by the emerging concept of data cooperatives, originally created to overcome the often common asymmetric relationship between data producers and users. At their core, data cooperative structures are owned by their membership and thus improve accountability, while advocating on behalf of the CGD producers and subjects. Early applications of these concepts demonstrate their ability to build higher levels of trust between the parties (at least those related to the supply and the use of the CGD), and hold the potential of integrating two worlds that have been distinct to-date, i.e. open (government) data and personal data organizations.

Ultimately, we reflect on the concept and the boundaries of what is acknowledged to belong to the CGD ecosystem. In this study we adopted a comprehensive definition, binding our unit of analysis within the scope of projects and initiatives based on the informed and active production/collection and analysis of data for given purposes. Data that is indirectly created by citizens (e.g. big data from digital footprints, social media activities) was left beyond the scope of this research. However, recent developments (enabled by the recent data protection regulations) are demonstrating how some instances of this data "indirectly" created by citizens can effectively become citizen data as these are now entitled and given the legal rights for data subject access and data portability requests. An example is the initiative Workerinfoexchange[13] which assists workers, especially those employed in gig economy platforms (e.g. Uber, Glovo), in requesting, creating, managing (and sometimes analysing) the data about themselves that is collected, stored and used within the organizations they work for. On their website, they describe this process as follows: *when you invoke these rights* (i.e. subject access and data portability), *any organization that has data about you, must provide you with a copy of this data, including information about why they collect this data, who they share it with, and if they make any automated decisions about you, as well as the logic of those decisions. We collect this data to support you and other workers in exercising your rights.* All in all, the argument relates to the fact that by exercising their own rights (i.e. rights of access, object, rectification, erasure, portability, and restriction), citizens can enable a shift from indirectly generated data to actual (decentralised) CGD. Through these reflections, we argue for the need to reconsider the foundational definitions of CGD towards also including data from the so-called Big Data ecosystem to which control is shifted to the individual, consistent with the decentralization trend described above, and enabled by the existing (and emerging) legal (digital) rights.

[13] https://www.workerinfoexchange.org/

Appendix 1

List of projects analyzed

Name	Link	Sector
RiuNet	http://www.ub.edu/fem/index.php/ca/inici-riunet	Water
Foodmapping	https://www.foodmapping.cat/	Food
BioBlitz Barris	https://bioblitzbarris.net/	Biodiversity
Líquenes de Barcelona	https://liquensdebarcelona.net/	Biodiversity
BioBlitz	http://bioblitzbcn.museuciencies.cat/	Biodiversity
Observadores del Mar	https://www.observadoresdelmar.es/	Biodiversity
Cobertura Mòbil	https://www.elperiodico.com/es/tecnologia/20160223/catalunya-detecta-las-zonas-con-poca-cobertura-gracias-a-una-app-4922116	Connectivity & data
Ritme Natura	https://ritmenatura.cat/	Environment
Odour Collect - D-NOSES	https://odourcollect.eu/	Environment
Beepath	http://beepath.org/	Mobility
Salus.coop	https://www.salus.coop/	Health
Genigma	https://genigma.app/ca/	Health
Planttes	http://www.planttes.com/?page_id=46&lang=en	Health
Wheelmap	https://wheelmap.org/?locale=es	Social
Floodup	http://www.floodup.ub.edu/	Water & meteorology
Cicada.cat	https://cicadacat.wixsite.com/index	Biodiversity
OpenTEK	https://opentek.eu/licci	Climate change
Arturo 300mil	https://300000kms.net/case_study/merce/	Urbanism
Observatorio Ciudadano	https://observasequia.es/	Water

de la Sequía		
Red de Observadores Meteorológicos	NA	Meteorology
MammalNet	https://www.mammalweb.org/es/?view=projecthome&option=com_biodiv&project_id=115	Biodiversity
Paddle Surfing for Science and PlastiPlancton BCN	https://www.asensiocom.com/surfingforscience/en/	Environment
Censo personas sin hogar / Fundacion Arrels	https://www.arrelsfundacio.org/es/censo/censo-barcelona/	Social
Saca La Lengua	https://www.sacalalengua.org/	Health
Aigua BCN	https://www.isglobal.org/-/aiguabcn	Health
FILMAR	https://cetaceos.webs.ull.es/bioecomac/filmar/	Biodiversity
Decode	https://decodeproject.eu/	Connectivity & data
Prou Transit	http://proudetransit.emiweb.es/	Environment
#Cuéntalo	http://proyectocuentalo.org/	Social
Citi Sense	NA	Environment
Mapa Sonoro Barcelona	http://www.bitlab.cat/projectes/mapa-sonor-de-barcelona/	Environment
WeCount	https://www.wecountmovilidad.eu/	Mobility
#Servet	https://servet.ibercivis.es/	Aerospacial
Smart Citizen Kit	https://smartcitizen.me/	Environment
Making Sense	http://making-sense.eu/	Environment
Fumuts Ros De Olano	https://twitter.com/fumutsrosolano	Environment & mobility
OLOT	https://eu-citizen.science/project/110	Environment & mobility
xAire	https://www.ub.edu/web/ub/es/menu_eines/noticies/2021/07/001.html?	Environment
Vigilantes del Aire	https://vigilantesdelaire.ibercivis.es/	Environment
CitiesHealth	https://www.citieshealthbcn.eu/en/resultats	Health
CSI-COP	https://csi-cop.eu/	Connectivity & data
HOOP	https://hoopproject.eu/	Circular economy

SEEDS - Science Engagement to Empower Disadvantaged adoleScents	https://seedsmakeathons.com/	Health
Juegos para el Cambio Social	http://www.ub.edu/opensystems/projectes/games-for-social-change/	Social
inSPIRE	https://inspiresproject.com/	Social
GenerationSolar	https://generationsolar.ies.upm.es/	Energy
CoAct	https://coactproject.eu/	Social
Pescadors de Plastic	https://mon.uvic.cat/pescadors-de-plastic/	Environment
Fotoveu Gotic	https://ajuntament.barcelona.cat/ciutatvella/ca/noticia/fotoveu-gotic-una-reflexio-veinal-sobre-el-turisme-massiu_765420	Social
Projecte Endèmic	https://projectendemic.com/	Health