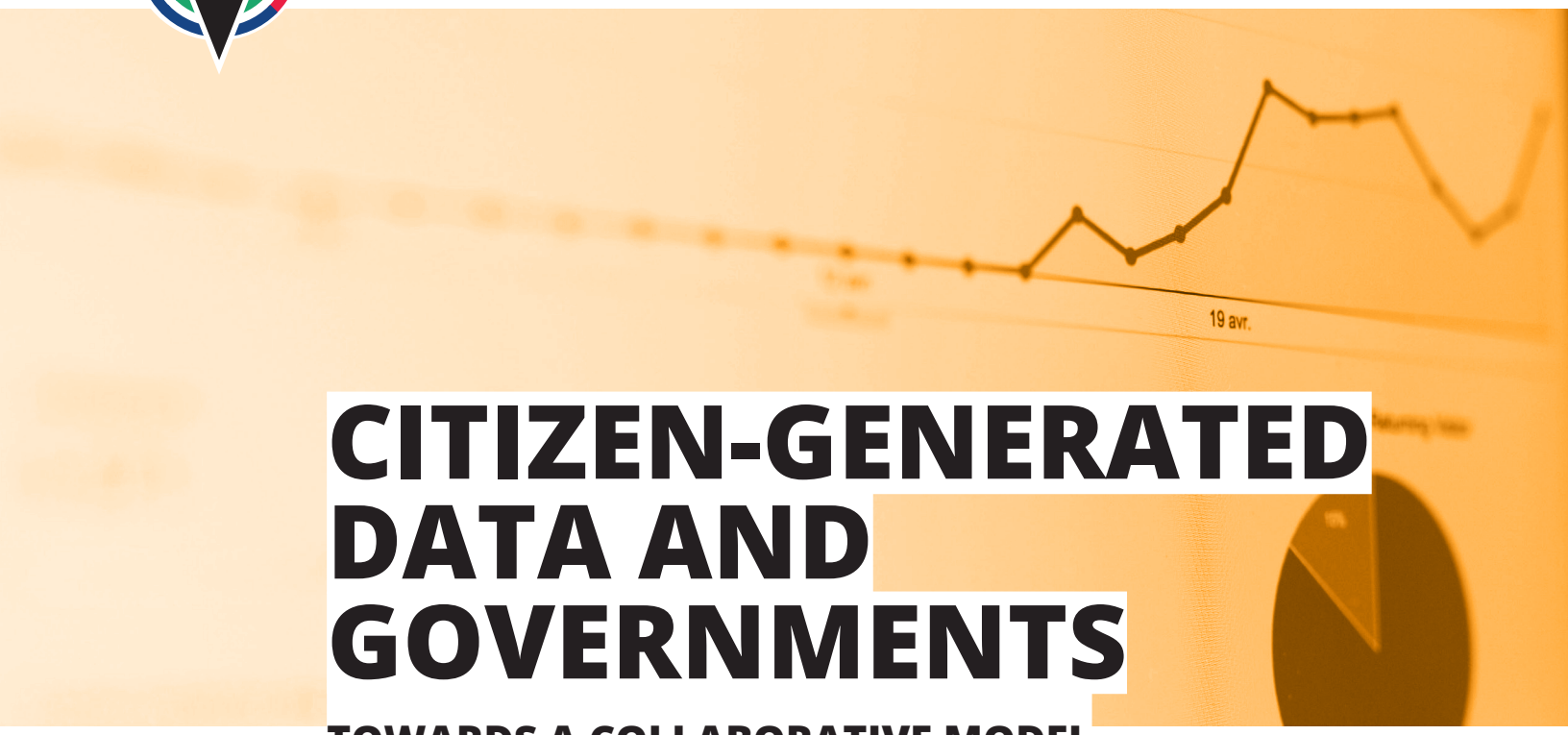
DATASHIFT



CITIZEN-GENERATED DATA AND GOVERNMENTS

TOWARDS A COLLABORATIVE MODEL

BY

CHRISTOPHER WILSON
ZARA RAHMAN

This piece explores how government hosting of citizen-generated data sets (CGD) can meet the needs of both governments and civil society, and open up opportunities for increased collaboration between government and civil society on collecting and sharing data, and using data to monitor and enhance progress on sustainable development. The piece begins by presenting the idea of government hosting for citizen-generated data sets, inspired by a recent conversation with the Innovation Lab in Buenos Aires. The following section discusses incentives, obstacles and benefits of this approach within the context of open data initiatives and development monitoring. The third section considers a model for this type of collaboration and suggests how it might be considered in countries with active monitoring and accountability efforts. The final section section proposes how additional research or practical efforts might help to develop this idea further, within the context of the "**data revolution for sustainable development**".

INTRODUCTION

This piece reflects on government hosting of citizen-generated data (CGD) and how this can meet the needs of government and civil society. It was inspired by a recent meeting with Argentine civil society organizations and government representatives, including the Buenos Aires Government's Innovation and Open Government Lab.[1] The meeting was organized to explore government perspectives on CGD and to look for opportunities to promote its use, and discussions covered issues of comparability and sustainability of CGD, as well as its potential to complement, validate and enhance official statistics. One of the most novel and exciting ideas that surfaced in these discussions was the potential for government open data portals, such as that managed by the Buenos Aires Innovation Lab, to host and publish CGD, and what benefits that might hold for the application and sustainability of such data.

Hosting CGD on government portals is a novel concept for many, with exciting potential to improve both the quality of national statistics and public trust in government data (see box on the Argentine context). Lack of sustainability models and limited methodological expertise are common challenges for CGD initiatives, which might be at least partly addressed by government hosting of CGD on their own data portals, or stronger collaboration between government and civil society. For many governments, lack of trust in their own data by citizens and patchy data coverage can be powerful incentives to explore such arrangements.[2]

This piece explores government hosting of CGD and how this can meet the needs of government and civil society. Following this introduction, we explore the potential benefits and obstacles both civil society organizations and government representatives may face when collaborating on government hosting of CGD. The third section suggests components for a model that civil society and government representatives could adopt to support the successful implementation of such an initiative. The closing section notes the dramatic lack of international experience with such solutions, and suggests some additional steps that could be taken to further our collective understanding of government and civil society data collaboration in general, and hosting schemes in particular.

---

1 Laboratorio de innovación y Gobierno Abierto de la Ciudad de Buenos Aires, see http://www.buenosaires.gob.ar/modernizacion/gobiernoabierto

2 For example, huge data discrepancies are reported on the numbers of disappeared people in Mexico, or air quality in Beijing. In these cases, CGD can provide crucial alternative datasets to important topics, and in some cases, put pressure on governments to collect better data themselves. A related situation can be seen in cases where data on important topics is simply not collected by governments or institutions, such as in the US, where collecting data on individuals killed by police forces is optional.

WHAT IS CITIZEN-GENERATED DATA?

Citizen-generated data (CGD) is 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, often organized and managed by civil society groups. This is distinct from "big data" or social media data, which is indirectly created by citizens through interaction with media platforms.

There is much enthusiasm about the potential of CGD to raise citizens' voices and to contribute to the "data revolution", but can also be criticized for its lack of representivity or statistical rigor. For more on CGD, see **this briefing note**.

THE ARGENTINE CONTEXT

The politics of institutional data in Argentina are marked by distrust of "official" statistics. The accuracy of data on inflation, generated by the National Institute of Statistics and Censuses (INDEC),[3] for example, is widely questioned. [4] Some investigations suggest that manipulation of this data goes as far back as 2007,[5] and continues today. As of September 2015, the official inflation figure is around 15%, but independent analysts estimate it could be almost double that in reality.[6]

This perception isn't confined to data on inflation. In 2014, INDEC also stopped publishing poverty statistics, citing "the difficulty of bridging the new and old consumer-price indexes," though the Economist attributes this to political reluctance to show increases in poverty levels.[7] This led some international ranking initiatives, such as the International Monetary Fund's World Economic Outlook Database, to exclude Argentine data in indices.[8]

But the lack of confidence in official data at home is perhaps even more troubling; the limited availability of reliable data can make developing effective economic policy impossible. As described by a former employee of INDEC, without the data, it's impossible to know what effect government policies are actually having.

Government statistics have now become a hot topic on the Argentine campaign trail,[9] matched by the launch of civil society fact-checking initiatives and interest in open data among government institutions. The Innovation and Open Government Unit was created in 2011 and two years later reformed as an Innovation Lab, with a four-part mandate, entitled: Open Data; Data Laboratory; Open Innovation; and Urban Sensorization.

Responsible for publishing open data sets and connecting data users in and out of the national and local Argentine governments with the data they need, the Innovation Lab is a unique institution in the Argentine context. It takes a progressive approach to hosting alternative types of data and supports collaboration between government and civil society. The Lab might be uniquely positioned to address the shortcomings of public data and public trust in Argentina, and their keen interest was a primary inspiration for the piece.

---

3 http://www.indec.gov.ar/

4 http://www.indec.gov.ar/diferencias.asp

5 Bauer, Michael "Argentinien: Wechselgeld vom Bäumchen," Der Standard, September 7, 2015. Accessed October 19, 2015. http://derstandard.at/2000021590346/Argentinien-Wechselgeld-vom-Baeumchen

6 Reuters, "Argentina says July inflation accelerates to 1.3 pct m/m," August 14, 2015. Accessed October 19, 2015. http://www.reuters.com/article/2015/08/14/argentina-inflation-idUSE6N0V209V20150814

7 H.C, "Fishy Figures," The Economist, September 17, 2014. Accessed October 19, 2015. http://www.economist.com/blogs/americasview/2014/09/statistics-argentina

8 "World Economic Outlook Database", International Monetary Fund, Oct 6, 2015. Accessed October 20, 2015. https://www.imf.org/external/pubs/ft/weo/data/changes.htm

9 Baker, Vicky, "Fact-checking Argentina's elections,", September 14, 2015. Accessed October 19, 2015. https://www.opendemocracy.net/vicky-baker/fact-checking-argentina-elections

INCENTIVES, BENEFITS AND OBSTACLES TO GOVERNMENT HOSTING OF CITIZEN-GENERATED DATA

INCENTIVES AND BENEFITS

Hosting CGD on government portals potentially offers a number of advantages for both government and civil society organizations.

For government actors mandated to maintain and promote data portals, incorporating CGD can significantly broaden the scope and coverage of those portals, adding depth and context in sectors where government data exists, and filling gaps in sectors where it doesn't. Inclusion of CGD in these portals implies that it meets certain thresholds for methodological rigour and sustainability, either prior to inclusion, or through data cleaning and institutional arrangements implemented during the inclusion process. The addition of complementary, high-quality CGD sets can be especially useful for statistical and monitoring professionals in government, offering novel opportunities for National Statistical Offices (NSOs), line ministries and national development and planning agencies to validate their own data and have access to credible and complementary data in areas where data collection has been traditionally weak.

Doing so also can also improve public perceptions of government data initiatives. Many governments invest significant resources in promoting their "data performance" in international fora such as the Open Government Partnership[10] or international rankings such as the Open Data Barometer. Including CGD increases data portal coverage and gives the impression of collaboration with civil society, which can also be important at home for deflecting criticism that governments are cherry-picking or manipulating official data to support policy aims. For governments that are already engaging with civil society around open data, publication and hosting of CGD on government-managed data portals provides an excellent opportunity for improving such engagement, as well as developing the capacity of civil society organisations and citizens to generate data that meets high methodology and data structure standards.

For civil society, including CGD on government data portals can improve the profile, accessibility, use and quality of that data. Government data portals will in most country contexts have wider audiences than most civil society groups, and can disseminate CGD to a wider spectrum of potential users than civil society would be

10 http://www.opengovpartnership.org/

able to do alone. This includes government data users, and the stamp of legitimacy conveyed by having data hosted on a government data portal can implicitly validate datasets in the eyes of government workers and statisticians. This makes the data more likely to be accessed and used by those developing policy and making policy decisions. Similarly, official data portals are much more likely to offer application programming interfaces (APIs), providing a low-cost opportunity to get CGD into the hands of app developers and service providers.

Inclusion of CGD in government portals will often be predicated on a number of conditions. These might relate to data structure and consistency, licensing, methodological transparency and rigor, and data sustainability. Improving data sets' quality and usability until they meet such thresholds will be a valuable undertaking (though it may sometimes prove impossible). Data portals that implement specific standards for data formats will help to make CGD more interoperable and comparable with official data. This, in turn, will enable civil society actors with limited statistical resources or expertise to use CGD to check facts or identify gaps in official data with confidence.

There are, however, obstacles to government hosting of CGD in many countries. Most obvious is the profound lack of trust that dominates discussions about data between civil society and governments the world over, a trend that is often firmly rooted in a more general political acrimony.

Despite this, hosting CGD on data portals can be an effective way to bypass these dynamics and lay the groundwork for more productive government-civil society collaboration on data and monitoring issues. Data portals are often maintained by non-traditional actors within governments, such as special units attached to executive branches, or ministries not otherwise associated with the production of statistics. Where these bodies are institutionally removed from ongoing data disputes between civil society and national statistical offices or line ministries, hosting collaboration can provide fruitful grounds for identifying unlikely advocates, even in countries where there is deep-seated scepticism about government manipulation of data.

DEFINITION: APPLICATION PROGRAMMING INTERFACE (API)

In computer programming, an application programming interface (API) is a set of routines, protocols, and tools for building software applications. In other words, APIs are sets of requirements that govern how one application can talk to another, making it possible for applications to share data, without having to share all of their software's code. A good API makes it easier for third-party programmers to build an application using the data, and to remix it with other applications or data.

POTENTIAL OBSTACLES

For governments, the most significant substantive obstacle to hosting CGD may be the quality of CGD sets. Most data portals will have clear requirements for data structure and machine readability. Statisticians (whether directly involved in data portals, or representing NSOs), may also insist on methodological criteria, or object to the inclusion of data that does not meet statistical standards. Some of these objections will be political, and require political solutions. Others will demand methodological rigour in data collection that cannot be retroactively addressed. Many, however, can be addressed by adding caveats and contextual data to explain what the data does and does not represent, or by standardizing or restructuring data sets. Engaging a data scientist not implicated in institutional politics can be a useful way of navigating concerns about the quality of CGD sets.

For civil society organizations, ensuring methodological rigour can prove a significant obstacle to hosting data on government portals. Not all civil society organizations producing data have in-house capacity to conduct statistical analysis, or apply and enforce rigorous methods in data collection. Hiring data scientists or statisticians to try and address such issues retrospectively can be expensive and might run counter to the incentives of advocacy organizations, who may prioritize forward-looking and policy-oriented actions over methodological housekeeping. Concerns about losing control or ownership over data or about how data will be presented and used can also obstruct progress.

THE STATISTICAL CRITIQUE OF CGD

Civil society's contribution of data to the "data revolution" is a cause of concern for some statisticians, largely due to questions about the data's representivity and quality. A core tenet of statistics is that a small group (of survey respondents, of export products or of water sample readings, for example) accurately represents the values of the larger group. This principle is what allows basic statistical measurement, and implies rigorous methods for selecting the groups of things that are polled and measured for official data. Crowdsourcing, which has become both a popular method for civil society to create CGD, as well as a buzzword in policy circles, cannot be representative in a statistical sense because it is individual members of the "crowd" who determine themselves whether or not they want to provide data. This leads many statisticians and other measurement professionals to dismiss crowdsourced data, and the other novel forms of civil society data they associate with it.

In addition, CGD sets are often vulnerable to critiques regarding the rigor of their data collection methodology, especially sampling methods. These concerns are at times valid, and can reflect a general lack of statistical expertise among civil society organizations.

Such obstacles are not insignificant, but if they can be overcome, there is clearly a great deal to be won by both government and civil society actors. The potential advantages of improved public perceptions of both government and civil society, more strategic engagement between government and civil society and better data accessibility can provide powerful incentives to work through such obstacles.

In addition, successfully hosting CGD on government portals can improve the quality and utility of data in a number of ways. Most critically, it can contribute to improving the monitoring and impact of important policy initiatives.

CONSIDERING A MODEL FOR GOVERNMENT HOSTING OF CITIZEN-GENERATED DATA

As the popularity and implementation of government-led open data initiatives grows, the opportunities for civil society to explore hosting opportunities with government also increases. At the time of writing, collaborative efforts led by Open Knowledge estimate that more than 400 open data portals are being operated the world over.[11] According to the Open Data Barometer, as of 2014 most countries had some sort of government-led portal or publication initiative for open data.[12]

Each country context will be different. Incentives for hosting CGD are likely to be strongest among government actors in countries that are actively promoting their open data performance, evidence-based policy making, or support for the "data revolution" internationally. Civil society may find it easier to promote the utility of its data in countries where there is a strong precedence for the purchase of private sector data, or where data publishing initiatives are under-resourced. Specific political, financial and socio-economic considerations will inevitably determine the potential scope and success of civil society and government collaboration on monitoring data. When feasible, however, such arrangements could be powerful.

11 See http://dataportals.org/

12 The Open Data Barometer index for 2014 notes that 67% of the 86 countries surveyed have "evidence of a national data catalogue providing access to datasets available for re-use. Access to the data could be provided directly on the catalogue or indirectly through pointers to the place where the data is located" (score of 3 or more). Data is available at: http://opendatabarometer.org./assets/data/ODB-2014-Survey-Ordered.csv.

HOW GOVERNMENT HOSTING OF CDG MAKES FOR BETTER MONITORING DATA

The incentives and obstacles for civil society and government organizations to collaborate on data collection and use will vary from country to country. When feasible, however, government hosting of CGD could dramatically increase countries' capacity to use data for monitoring and improving development initiatives.

Greater coverage, prominence and accessibility

Hosting multiple types of data on government data portals increases a portal's relevance and prominence in a diverse data landscape. This increases the chances that people seeking data will go to government portals as a first port of call, increasing use in line with portal mandates. Data sets hosted on such portals can in turn be expected to reach wider audiences, while adherence to common and accepted data standards and formats can facilitate the download and use of relevant data by third parties.

Conservation of resources

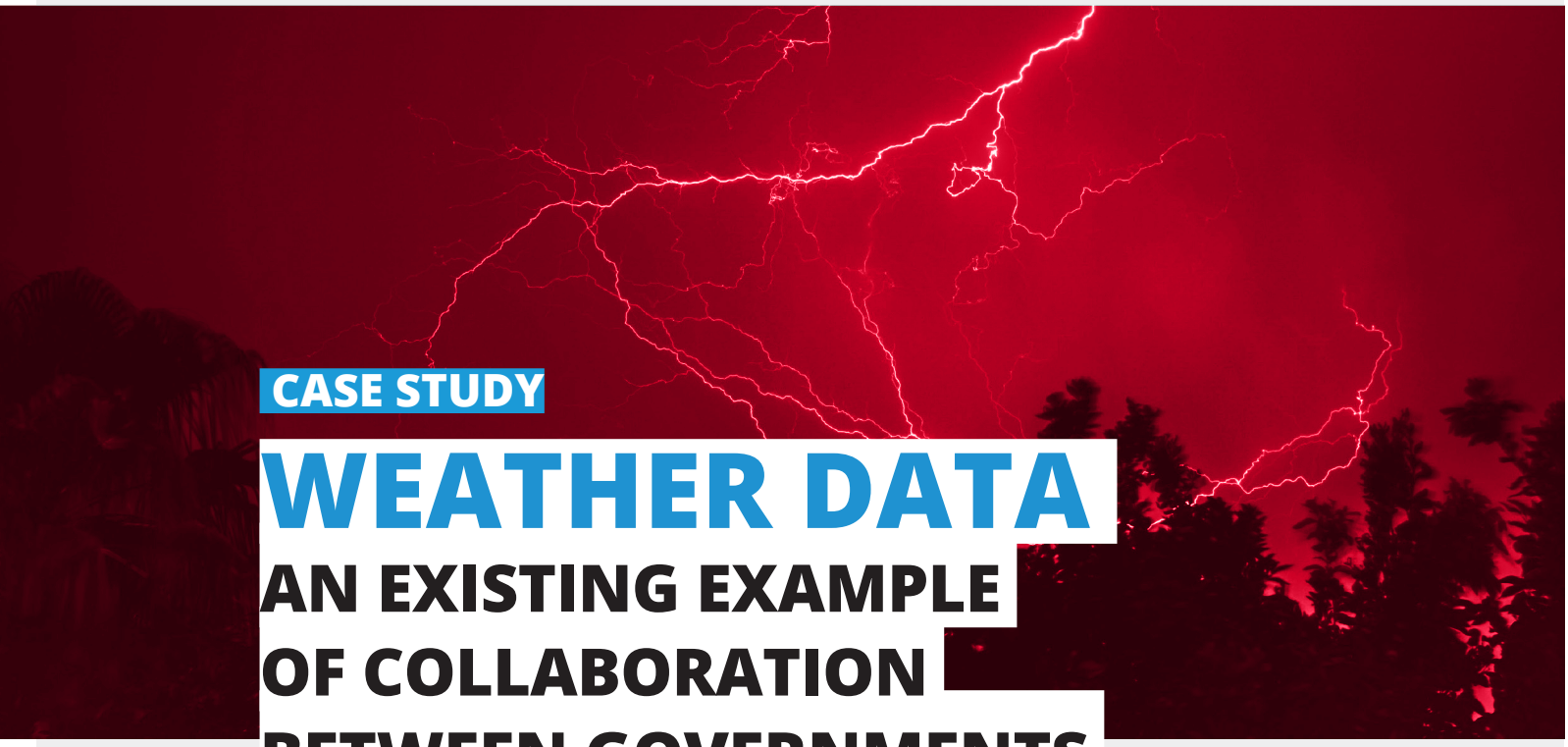
CGD typically gets produced outside of government data budget lines, and can often cover issues and sectors that would otherwise either be purchased from the private sector, incurring a financial cost, or where there would be no data at all. Civil society organizations that host data independently have to pay more to do so; if a government entity hosts it for them, they may be able to spend more time on increasing the data's quality, or using their limited resources to support outreach or participation in the data.

Improving data quality

Shared hosting will often involve standardized data formats and structures. This means that there are opportunities to harmonize and increase interoperability between government and CGD sets. This can also provide unique opportunities to assess the quality of a given data set, or find areas where additional data collection is needed. For example, while data on national access to primary education may not be disaggregated by gender or ethnicity, CGD (though not representative and with limited coverage) may cover these issues, providing important insights or suggesting where subsequent data collection should invest resources. This can feed directly into strategies for data collection or policy action.

Provides a frame for collaboration around data, monitoring & development programming

Collaboration on data hosting provides government and civil society with a "safe space" where they can collaborate on data. Although it is ostensibly removed institutionally and thematically from controversial issues such as the monitoring and priorities of development programs, data hosting can provide a foundation for collaboration in those areas and help strengthen relations across civil society and government.



CASE STUDY

WEATHER DATA

AN EXISTING EXAMPLE OF COLLABORATION BETWEEN GOVERNMENTS AND CITIZEN-GENERATED DATA

Within the realm of citizen science, governments and avid citizens have been collaborating for a long time. Specifically within weather and climate data, data gathered by individuals has long been an essential part of many national weather forecasting institutions.

In the UK, for example, citizen scientists contribute to the data gathered and used by the Met Office, which is the UK's national weather service and one of the world's leading providers of climate services. Citizens have contributed to historical data via the Old Weather initiative,[13] allowing the effects of global warming over time to be observed and analysed more clearly. The project uses the power of the crowd to transcribe weather observations written in historical shipping logs recovered from archives around the world.

13 http://lifeofdata.org.uk/node/old-weather/Culture/

Amateur weather observers also contribute regularly to weather data used by the Met Office and by climate scientists, using sensors to measure variables such as rainfall and air pressure. The Climatological Observers Link (COL)[14] provides guidance on data formats and standards to follow, and observers are encouraged to follow these protocols so that the data they upload can be easily integrated with other data sets. The monthly bulletin produced by the COL, collating the data that they received in the preceding month from serious weather observers, is considered to be of sufficiently high quality that it is archived by the UK Met Office.

The Met Office also has its own site where it gathers observations from weather observers across the UK, called the Weather Observations Website, or WOW[15]. The site allows for both manual inputs – ie. without any special equipment necessary – and automatic data inputs, from observers who have access to compatible Automatic Weather Stations. WOW is a collaboration between the Met Office, the Royal Meteorological Society, and the UK Department of Education.

The project offers a number of concrete examples of some of the potential benefits drawn from government and civil society collaboration. Notably, the data gathered by amateur observers is, once uploaded to the site, "kept indefinitely"–so, the data will be kept online even if the observer who contributed the data stops contributing. Although this policy is being reviewed every twelve months, for now at least, it provides sustainability to the data created through this initiative. For the contributors, this brings a number of benefits: knowing that the data they as individuals are collecting will be brought together with other relevant data sets, and having their data kept online in an easy to find repository. Extra motivation to the contributors is given by the Met Office's official and public recognition of the observers' data as a legitimate and useful data source.

For the Met Office, being able to easily compare their "official data" with data sets coming in from serious amateur observers allows them to verify that what they are receiving and reading is accurate, and gives them alternative measurements for forecasts (perhaps from areas where less coverage is available). Recognition of the initiative's importance in the UK has led the Australian government's Bureau of Meteorology to partner with the UK Met Office to build a WOW site for Australia[16]. This demonstrates that, when the circumstances are right, initiatives of this sort can be replicated across different geographic regions.

14 https://www.colweather.org.uk/index.php

15 http://wow.metoffice.gov.uk/weather/view?siteID=878216001

16 http://www.bom.gov.au/wow-support/

CONSIDERATIONS FOR A MODEL

Government hosting of CGD could be **holistic**, in the sense that government portals host and incorporate all relevant data that meets thresholds for methodological rigor and sustainability, or **ad hoc**, if specific data sets are incorporated into data portals on the basis of isolated decisions or demand.

The degree of collaboration may also vary. Inclusion of data sets may be predominantly **government-led**, when government actors seek out CGD that is already available and feature that data. Doing this unilaterally, without contacting the civil society organizations that produced the data is likely to be less constructive, partly because engagement with data producers can be essential for securing useful contextual information about data provenance, accessibility and sustainability. Scenarios that are **collaborative**, or **civil society-led**, are more advantageous. Here, civil society groups respond to gaps in existing data that can be filled, either in response to government requests, or according to their own assessment of available data in a given area or sector.

Investment of time and resources is also required to include CGD in government portals, including the interrogation of data sets to determine that they meet methodological and structural thresholds for inclusion, as well as the cleaning and standardization of data sets that do. In some cases it may be feasible for portal managers to conduct the majority of this work, though doing so may raise civil society concerns in some context about the integrity of data that is cleaned and hosted solely by government.

Given that civil society organizations will often lack the human, financial or technical resources to manage such processes alone, **collaborative models may be the best solution**. Workshops and other events in which portal managers and civil society organizations together evaluate and prepare data for inclusion in official portals can provide legitimacy to process, while also allowing for financial and time burdens to be flexibly shared according to specific contexts and demands. Collaborative investment of time and resources can also help build civil society statistical and data capacity, lowering government transaction costs in future collaborations and CGD incorporation.

Generally, a holistic and collaborative approach to hosting, in which investments of time and money are shared by government and civil society, is likely to best meet the needs and incentives of all parties. This will inevitably begin with conversations between CGD producers and the managers of online data portals, in which each party's incentives and opportunities will dictate the scope for collaboration. However, a few specific activities may help to further such a conversation, and facilitate a more collaborative and mutually beneficial process.

COMPONENTS FOR A COLLABORATIVE APPROACH

Data sets mapping

An obvious first step towards understanding how CGD can complement what is already included in a data portal is by mapping what data is available both in and outside of that portal. Basing such work on conversations between both civil society and government actors about the nature and purpose of the mapping (what to include, who the potential users are, where similarities and differences lie) can pave the road for further collaboration down the line, establishing trust and providing guidelines for later decision-making. Mappings can also be used to identify the data that is not yet available, but which is in demand among government, civil society, or the users of data portals. In some cases, government data portals provide a specific page for users to 'request' data, which provides useful insights into the type of data that is in demand.[17]

Workshops on inclusion criteria, structures, methods and uses

Workshops are a strong practical tool for establishing the criteria for data set inclusion in government portals. This can be done on the basis of, or in parallel with a mapping exercise, and should be structured to ensure that civil society perspectives on data standards are well accounted for. Organizing events around training or discussion on specific technical issues, such as data structures, collection methods and data usage for development monitoring, can be an effective way to ensure broad participation and strengthen collaboration. Such an approach has the added value of building civil society statistical and methodological capacities, while also helping governments to identify novel ways in which non-representative CGD can complement and otherwise strengthen official statistics.

Secondments and fellowships

Managers of government data portals will likely need to specifically allocate resources in order to incorporate a significant number of CGD sets, and this work may often fall to small teams that are already overworked and underfunded. Where feasible and resources can be secured (also through third parties), options should be explored for establishing fellowships or secondments for staff from civil society organizations producing data, to work within open data portal institutions, specifically on the process of incorporating CGD sets. This arrangement could be a powerful mechanism for strengthening dialogue and collaboration around monitoring data. It could also increase technical capacities among civil society organisations, even when implemented on a very short-term basis.

17 https://data.gov.uk/data-request

Collaborative monitoring

The entire premise of this piece on government hosting assumes that civil society and government data can be useful for monitoring and enhancing development programs. Following, or in parallel with, efforts to include CGD sets, government and civil society actors should collaborate on efforts to monitor and enhance government and civil society development initiatives through the application of multiple data sets. This may involve the development and identification of novel monitoring methods and data analyses.

This section has suggested some of the characteristics that define approaches to government hosting of CDG, and some activities that can strengthen such efforts. These components will manifest themselves differently in different country contexts, as resources, incentives, data availability and political dynamics also vary.

MOVING FORWARD

This piece has offered some preliminary thoughts on the potential for a specific kind of collaboration that could contribute to the discourse on the "data revolution for sustainable development", as inspired by a specific conversation in a single country. At the time of writing, examples of this kind of collaboration seem to exist largely within the citizen science space, so speculations about this type of collaboration in other areas may well be fanciful. It nevertheless seems that there are a number of mutual advantages for both government and civil society to be gained from such collaboration, and that further exploration is not only a good idea, but inevitable.

As global discussions about evidence, monitoring and the "data revolution"–particularly in the context of the Sustainable Development Goals (SDGs)–continue to gain speed, as the global norms surrounding open data further entrench themselves in international and national debates, and as our communal understanding of how to use novel types of data in complementary ways increases, collaboration around government hosting of CGD can be a small but powerful way to strengthen the potential impact of CDG at the country level. To enable this requires further thinking, as well as practical efforts.

In practical terms, even preliminary conversations between civil society and government actors about the potential use and utility of CGD can provide tremendously important insights, not only for the feasibility of this kind of hosting arrangement, but regarding the perceived and actual utility of CGD for monitoring development progress. We need more conversations between governments and civil society, between statisticians and those leading crowdsourcing initiatives, and between international policy and national accountability movements. We also need better documentation of these discussions, with a special focus on where and why CGD is (or is not) adding value to monitoring and development programming.

DataShift will continue to work on these issues, by supporting local partners in four pilot locations to build their capacity to produce and use CGD. Additional insights and efforts by peers in other countries are also essential, and we encourage organisations working on CGD initiatives to continue documenting their experiences.

Picture by Leandro Kibisz, https://www.flickr.com/photos/loco085/8661487925

DataShift is a multi-stakeholder, demand-driven initiative that builds the capacity and confidence of civil society to produce and use citizen-generated data to monitor sustainable development progress, demand accountability and campaign for transformative change. Ultimately, our vision is a world where people-powered accountability drives progress on sustainable development.

DataShift is an initiative of **CIVICUS**, in partnership with **the engine room** and **Wingu**. For more information, visit www.thedatashift.org or contact datashift@civicus.org.

CIVICUS | the engine room | Wingu Tecnología sin fines de lucro