June 2017

The State of Open Government Data in 2017

Creating meaningful open data through multi-stakeholder dialogue

Danny Lämmerhirt, Mor Rubinstein and Oscar Montiel

OPEN KNOWLEDGE INTERNATIONAL

The State of Open Government Data in 2017

Creating meaningful open data through multi-stakeholder dialogue

Danny Lämmerhirt, Oscar Montiel, Mor Rubinstein[1]

[1] **Danny Lämmerhirt** is the research coordinator and researcher at Open Knowledge International. He also leads the methodology around the Global Open Data Index. **Oscar Montiel** is International Community Coordinator at Open Knowledge International. **Mor Rubinstein** is the Data Labs and Learning manager at 360Giving. She coordinated the Global Open Data Index research efforts between 2014-2015 and was the project manager for GODI 2016/17.

This is Open Knowledge International's first State of Open Government Data report. Based on key findings from our work on the Global Open Data Index (GODI) 2016/17, it outlines the obstacles to open government data publication, and suggests steps that will allow progress in the field of open data.

In our view, public institutions should align the data they produce with the needs of civil society groups, citizens and other users. As mentioned in Open Knowledge International's recent Data And The City report, data infrastructures - the frameworks on which data is produced and published - are not mere "raw" resources that can be exploited. They are best conceived as spaces for public participation, a lively ecosystem in which audiences creatively use data to engage with public institutions. This can lead to new kinds of relationships which strengthen calls for a range of emerging goals focused on transparency, accountability, public participation, public service delivery, technological innovation, and economic growth.

Yet, institutions are producing more information which is encoded in forms that are preventing data publishers and public users from communicating with one another. Dialogue is critical to produce relevant data that can be used by civil society, and GODI - more than only a benchmark - provides the platform for this dialogue.

For GODI 2016/17, we ran a public dialogue between governments and data users for the first time to foster the production of meaningful data. Below we share learnings and outcomes from this process and explain important variables to further elaborate this dialogue model. We also discuss GODI's future role in steering these discussions. This document is open to debate, to continue learning from your experiences. We would love to hear your feedback in our discussion forum.

1. The Global Open Data Index - improving the state of open data through dialogue

In 2013, the open data community was advocating for the publication of more open government data. But without having a clear picture of how much data had been published so far, the community struggled to make strategic advances. So GODI was created to help the community shed light on how much open data is published in any country and to allow national communities to drive advocacy in their own contexts.

Since its creation, the open data community has been at the heart of GODI and has helped to reshape it constantly. GODI is more than just a benchmark - it is also an interface between data publishers and users. In the past, government and users primarily engaged with GODI staff members but did not engage with one another. In a few exceptional cases, the open data users directly used the GODI results to consult with their governments but communication flows were asymmetrical, invisible or did not target the responsible actors. In short: the communication between auditor (civil society) and auditee (public institutions) was not conducted in an effective, streamlined process.

This year, we wanted to use the launch of GODI to spark dialogue and provide a venue for the ensuing discussions. Evidence shows that governance indicators drive change if they **embrace dialogue and mutual ownership of those who are assessed, and those who assess.** The public dialogue phase was initiated on the Open Knowledge International forum, which resulted in more than 187 different questions being raised by data users and government.

Through this dialogue, governments learned about key datasets and data quality issues, while also receiving targeted feedback to help them improve. At the same time it allowed the community to understand the **mechanisms** for how and why open data is released or not. Our public dialogue model shows that there is a need to discuss and learn from one another, and to close the loop between government and civic actors.

2. The state of open data in 2017

GODI 2016/17 identifies three critical obstacles preventing open data use: data is hard to find; not user-friendly; and not openly licensed. This section examines each obstacle, presents recommendations for open data decision-makers and demonstrates how public dialogue can contribute tackling these obstacles.

2.1 Data is hard (or even impossible) to find

Imagine searching for a book among poorly labelled library shelves. To find the right book, a great deal of searching and persistence would be needed. Nowadays we continue to search in public information experiencing similar findability problems, but across many more archives available online. Improving findability of data is crucial for everyone. If even open data experts struggle to find the relevant data, who else would be able to?

We identified four problems and suggest possible solutions for open data findability, whether in portals or to make it more accessible through common search engine queries:

Problem	Solution
Citizens still need to check many different places on the web to stitch data together	• **Understand what data can be related to one another** (either based on data standards, citizen consultations, or similar processes). • **Highlight next to the data** where to find **any additional data that is related** to a specific dataset**.** • **Create a data portal to publish open data nationwide.** The data can be hosted on a platform like CKAN or elsewhere. What is important is that the data is accessible from there.
Data may be hidden deep in websites and names of links are neither meaningful nor self-explaining	• **Use well-labelled links to make browsing for data more intuitive**. If open data cannot be placed on a homepage, make sure that users can intuitively click through a website.
Bad naming or website indexing forces users to experiment with queries	• **Give data files comprehensible names.** Names matter for search engines and humans. • **Tag data files** so that users do not need to know the exact name of a file when using a search engine**.**
URLs are not permanent and lead to empty or broken websites	• **Make data permanently accessible.** Data should be made available at a stable Internet location indefinitely and in a stable data format for as long as possible.

How our public dialogue phase facilitated progress towards findability:

- **Governments showing where data is actually published:** See the discussions about land in India or maps in Brazil.
- **Raising GODI's value as a link registry:** Government feedback enabled to add more relevant URLs to GODI's results, making it easier for our users to find it.
- **Improving user experience on government websites:** Some governments that engaged actively in the dialogue responded after GODI publicly flagged poor findability. For example, Colombia directly implemented changes to their site structure to increase findability, as well as the Mexican government, which implemented tags and built packages of the data for bulk download.

2.2. Government data cannot readily be used

Different audiences have very different needs for data. To align government information systems to these needs, governments should develop a holistic, user-centric understanding of data quality. However, as we explained in our blogpost on this topic as well as during a discussion on the Open Knowledge International forum, there is no single definition of what good data quality is. In regards to GODI, we identified two larger quality issues:

- **Quality of the data itself:** Datasets do not always contain key elements that would make them useful or processable. Data may also not be published timely or aggregated on a high level. Sometimes datasets can contain unusual coding, or other elements that hamper to interpret them in the first place.
- **Technical quality:** Data should be machine readable and easy to access. Governments increasingly publish data online. Yet, too often we notice trade-offs between data quality and quantity. For example, **governments do not publish raw data** which is unsatisfying for developers, researchers or topical experts. **HTML is the most popular format for information publishing**, which is a bad practice to open data publication since it is not easily accessible to technical purposes. Furthermore, **only a fraction of all datasets is published in machine-readable formats.**

More specifically, we identified seven problems and suggest possible solutions for open data usability, regardless of the publication platform:

Problem	Solution
Data is not available or hasn't been considered by publishers to be crucial data because there are different routines and priorities to produce data inside the government.	• **Understand and rethink data production chains.** Check if there are institutional routines or sectoral guidelines that inform how government data is produced and communicated. Do these conform to the principles of open data?
Data may be published according to relevance criteria of government, but these might not align with the needs of primary user groups.	• **Understand user needs.** This may be achieved by running focus group sessions or developing user personas (see this link for examples of how US cities did this). User personas enable to understand data needs as well as the risks of opening data.
Governments publish data in many forms, from cadastral maps, to interactive bubble charts of budgets. Some of these help non-experts to make sense of data more easily, **but they do not show raw data**.	• **Publish raw data that is accurate and precise. As principle 2 of the Open Data Charter states:** "To the extent possible, release data in its original, unmodified form, and link data to any relevant guidance, documentation, visualisations, or analyses. To the extent possible, release data that is disaggregated to the lowest levels of administration, including disaggregation by gender, age, income, and other categories."

Government publishes in PDF or HTML and not in machine readable formats.	• **Ensure that data is processable.** Raw data must be published in machine-readable formats, which need to have consistent values. This also to verify consistency of data by checking for missing data values and alike.
Datasets aren't easy to understand and lack more context to be useful. Files are named in implausible ways, or have an incomprehensible structure.	• **Add metadata to ensure that data can be understood by citizens.** Spreadsheets are not self-explaining especially if they contain expressions that are not used in daily language. Any special terminology needs explanation by using metadata (data about data). There are many ways to add metadata. In any case, metadata should be machine-readable and **make metadata easily findable.** Metadata must be published close to a datasource and clearly refer to a piece of data.

How our public dialogue phase facilitated progress towards usability:

- **Discussing appropriate levels of detail and completeness:** Governments explained the blockages of producing and publishing data at a certain level of detail (see as example Uruguay's comments on producing transactional spending data).
- **Explaining how data can be made more understandable:** The GODI team could flag good examples of data understandability in the forum (e.g. weather data documentation in Canada).
- **Discussing appropriate file formats** for data publication, as exemplified by Australia's national laws.

2.3. Open licensing is rare practice and jeopardised by lacking standards

Only a small fraction of government data is truly openly licensed. Widely acknowledged tools such as Creative Commons 4.0 licenses are rarely used. Most of the time, governments

use customised terms of use or modified licenses. This is a critical problem. The **proliferation of custom terms can cause incompatibilities between licenses** - posing major blocks for data use (see our blogpost on the topic). Risk aversion and the fear for unlawful data use can culminate in unnecessary or ambivalent clauses, which in turn can cause legal concerns, especially around commercial use. Standard open licenses are intended to reduce legal ambiguity and enable everyone to understand use rights. Yet, **many licenses and terms contain unclear clauses**, or are not obvious what data they exactly refer to. Below we enlist the most common problems with licensing and possible solutions to these issues:

Problem	Solution
It is unclear whether copyright protection applies to data or not.	• **Does the data and/or dataset fall under the scope of intellectual property (IP) protection?** Often government data does not fall under copyright protection and should not be presented as such. • When government data is in the public domain by default, make clear to end users what that means for them.
Governments choose license terms that do not fall under the Open Definition or are not officially acknowledged as being open.	• **Use standardised open licenses.** Open licenses are easily understandable and should be the first choice. The Open Definition provides conformant licenses that are interoperable with one another. To guarantee a license is compatible, best practice is to **submit the license for approval under the Open Definition**.
The license does not entirely clarify what data it applies to.	• Exactly **pinpoint** within the license what data it refers to and provide a timestamp when the data has been provided.

License terms may not be immediately findable, or are published on a different webpage that is not linked to the dataset.	• **Clearly publish open licensing details next to the data**. It should be both human and machine-readable. Maintain the links to licenses so that users can access license terms at all times. It also helps to have a license notice 'in' the data. **Highlight the license version** and provide context how data can be used.
There are mixed messages about copyright in the sites and platforms where data is stored.	• Re-evaluate the web design and **avoid confusing and contradictory copyright** notices in website footers, as well as disclaimers and terms of use.
The way some additional clauses are written can be confusing for the user.	• Whenever possible, **avoid restrictive clauses** that are not included in standard licenses.

How public dialogue facilitated progress towards open licensing:

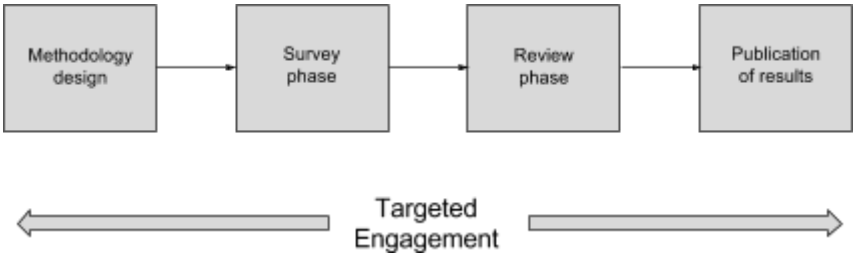
- **Flagging restrictive or ambiguous clauses -** see the case of Belgium budget data .
- **Clarifying legal rights attached to data** (see here).
- **Discussing best practices how to signal public domain status** (see the case of Mexico).

3. Creating meaningful avenues for better data through dialogue

Running a public dialogue following GODI's publication of results for 2016 provided unique incentives: governments could contest the results, influence their ranking, or flag their willingness and investment to improve their data. Open data communities could also challenge our findings and use the results to evaluate government performance publicly. We understand that this is a unique situation and that it is unsure whether the model can be sustainable. This

section addresses the key learnings from 2016 and what they teach us about **redesigning engagement channels throughout the entire process of GODI.**

In order to **establish a meaningful relationship between auditor and auditee,** governments need to be part of the equation so we can ensure two things: that the assessment is not a mere policing exercise from the civil society side; and that governments have the opportunity to adopt better practices following this exercise. It is key to understand the incentives of different stakeholders and to design **mutual benefits for all stakeholders.** The 2016 public dialogue shows that this relationship can be established without losing GODI's civic focus.



The four stages of the GODI process

But this raises questions about when public dialogue is most needed and whether it is possible to **expand targeted engagement and co-ownership** throughout the entire GODI process. To provide answers, we summarised our learnings from GODI 2016/17 into three areas:

1. The need to **rethink who to engage with throughout GODI stages**
2. The need to **increase active outreach and sustainable engagement** throughout GODI stages
3. The need to **develop more targeted feedback for data publishers and users**

3.1. Rethink who to engage with throughout GODI stages

There are many different groups that are interested in the GODI process. What learnings do we draw from the public dialogue about engaging stakeholders? Who could contribute to the process and how? How does each stakeholder benefit? The following examples show how different stakeholders interact in the discussion:

Open Data Communities

This group is where our main users are and is the one we engage with during most of the GODI process. This work is divided into different regions where local groups or organisations help outreach by place or country. In general, these communities belong to the Open Knowledge Network but we have realised that through our extended outreach, we have received submissions from groups and individuals we hadn't reached before. The public dialogue showed that constant engagement is instrumental for others to discuss the survey process and findings, to spot datasets more easily, to reduce workload, and ultimately to avoid submission fatigue among our volunteers.

Government officials

Government officials played a largely passive role in past GODI versions, only engaging at the end of the review process. **Yet, it is crucial to ensure buy-in and ownership of those who are assessed.** Therefore it is important to engage actual *data producers*, not only open data champions within government. More targeted outreach could enable them to allocate resources -both human and financial - in a more effective way. Engaging governments in the survey and review stages also provides the most representative datasets at the desired level of detail. This could reduce some of the costs involved in our community-based survey phase. Cooperation with other global initiatives such the Open Data Charter or the Open Government Partnership is crucial to foster more participation by government actors.

International NGOs

Originally we partnered with international NGOs who employ or know topical experts and give advice around the design of key datasets. The partners operate at a global scale and have an interest in understanding availability of data in their respective fields. These partners proved to be crucial anchor points for GODI to connect with topical experts on the ground, such as parliamentary monitoring organisations or environmental activist groups.

Topical reviewers

Promoting sustained dialogue throughout the whole process will allow our expert data reviewers more context and the ability for governments to understand the decisions made regarding reviews, as well as proactive reactions to the data publication in terms of quality and findability.

Topical experts with clear data needs

Another important, yet currently indirect, GODI user group are topical experts. In order to reach these experts, we normally rely on international NGOs (see above) and researchers to map data ecosystems (more detail below). Currently, most of the topical experts are based in the global north. We acknowledge it is necessary to reach out to topical experts who work on social change, even if their focus is not open data. Measuring valuable datasets will also mean understanding what meaningful is for these other types of users pushing for social change, focusing in the global south, where we have found more data gaps.

3.2. Increase active outreach to audiences throughout GODI stages

Alongside discovering new audiences to engage with, it is necessary to evaluate GODI's engagement channels. Currently GODI has the following channels to identify and engage stakeholders:

- **Fellowships and in-house research to understand data ecosystems:** This work was aimed at developing user stories about data elements in different categories by conducting targeted research. This includes to identify and interview key stakeholders, from multilateral organisations and international NGOs, to national government officials, local communities and individual data users. *These insights are instrumental to identify and collaborate with data users more closely*. An example is our open land data fellowship with Cadasta Foundation. The fellowship helped Cadasta Foundation to advance their knowledge around key datasets and user personas. Similar efforts have been made through in-house research on five data categories, and are now followed by a research fellowship to understand the users of water quality data in four Asian countries.
- **Community work:** Since GODI is a community-based assessment, we rely heavily on the submissions done by volunteers from around the world. Since last year, we have relied on "community wranglers", who are leaders in their region, in order to gather as many submissions as possible from individuals and groups in their own networks. This has allowed us to get a sense of dynamics in some regions, like the Balkans, Southern Africa and the Caribbean. On the other hand, engagement in regions like MENA and some parts of Western and Central Africa proved difficult for this edition of GODI, pointing us to a clear need to improve engagement in these regions. To improve this process, we need to have better planning to engage with old and new communities in a

more meaningful and proactive way, throughout the entire process, also to tackle a possible fatigue that we have noticed in the community.

- **Social media:** The most common channel for us to communicate with the GODI network are our social media platforms. Stakeholders communicated centrally through the forum but these communications have been mostly to large extends uni-directional, until the start of the dialogue phase. We have identified that communication throughout the whole process needs to be two-sided and constant in the relevant matters for users and governments.

3.3. Develop more targeted feedback for data publishers and users

The public dialogue has shown that GODI's process and its results are not self-explanatory and that more targeted feedback should be provided. Some data users and governments indicated that they struggle to interpret the results of our assessment. However, presenting the results in a way that makes them most usable can prove challenging. This year our research team enriched each dataset results with review comments and a review diary. Representing the reality of open data is complex, and customising feedback for each country and situation would cause high documentation costs - something our standardised, easily readable survey intends to buffer. Some feedback we already know of includes:

- To document any methodology changes and proactively communicate these in a central place (examples how we dealt with this concern include this blogpost, on how we define key datasets and how we evaluate them).
- To exactly pinpoint the issues underlying our results (why is data not findable, why exactly do we subtract points for accessibility, etc).
- To communicate more clearly that GODI measures different degrees of openness.
- To present and reward smaller progress towards publishing open data. Both parties regard this as especially constructive for their work (an extensive forum discussion can be found here).

User surveys can help to understand *how people are using* different parts of the GODI interface and what could make GODI's results more actionable. The public dialogue showed us that both the open data community and government care deeply about the feedback GODI provides to them. The results should provide a realistic estimation of open data.

4. Conclusion

To date, there is a mismatch between the data governments provide to the public and the needs of data users. Government information systems are a space for public participation where different audiences creatively use data to engage with public institutions. A closer dialogue and mutual ownership are needed in order to bridge the mismatch between supply and demand of data. This report presents how GODI creates and fosters this dialogue.

Drawing on our experience with GODI 2016/17, we diagnosed three problem areas: data is hard to find; not user-friendly; and rarely openly licensed. In a public dialogue phase, GODI allowed data users and governments to engage with the results, correct entries and thereby learn from one another. As has been shown, public dialogue can facilitate progress towards all three of these areas and has been perceived by participants as highly relevant and important for their work.

Yet, creating venues for public dialogue and sustaining forward is not without challenges. It is key to rethink engagement channels throughout the entire process of GODI, and also to understand the incentives of different stakeholders to participate. Emerging questions include:

- Who could contribute to a dialogue process and how?
- What would the benefits for each stakeholder be?
- Which aspects of GODI do stakeholders value most?
- How should we convey GODI's results in the most meaningful way?

Because these questions cannot be answered without our users, we would love to invite you to discuss our report, its findings and the future of the dialogue model in our forum!