Improving open data quality through citizen engagement and data engineering

César García Sáez
cesar@lahoramaker.com
La Hora Maker
Madrid, Spain

Figure 1: Madrid Sample Open Data Error Message, 2022.

ABSTRACT

In this paper, we will focus on improving the quality of open data offered through open data portals by engaging with citizens using open source tools. To do so, we will evaluate current open source solutions from the data engineering field, selecting those better suited towards collaborative workflows. We will propose a methodology to evaluate errors in open datasets and notify public administrations, resulting in better overall quality and more trustworthy and transparent processes.

OpenSym 2022, September 7–9, 2022, Madrid, Spain

ACM ISBN 978-1-4503-9845-9/22/09.
https://doi.org/10.1145/3555051.3555075

CCS CONCEPTS

• **Human-centered computing** → *Open source software.*

KEYWORDS

open government, open data, data engineering, citizen engagement

ACM Reference Format:
César García Sáez. 2022. Improving open data quality through citizen engagement and data engineering. In *The 18th International Symposium on Open Collaboration (OpenSym 2022), September 7–9, 2022, Madrid, Spain.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3555051.3555075

1 MOTIVATION

Open Government Data (OGD) is regularly produced by public administration bodies. These datasets cover all kinds of topics, from health to taxes. They are distributed via Open Government Data Portals (OGDP) [4].

Over the years, several initiatives have focused on making the data on OGDP more accessible, useful and interoperable. Initiatives like 5 Stars data [2] created by Tim Berners Lee, propose a hierarchy of several levels of usefulness in open data. More specific frameworks have emerged to check the quality of data, putting the focus mostly on metadata [9]. However, most of these frameworks don't address a fundamental problem: errors on data values themselves.

Spain is one of the leaders in OGD, with a maturity rating of 95 over 100 [3]. Yet, according to recent research [1], more than 56% of Spanish OGDP don't offer enough means to enable professional reuse of the data they provide. For example, most OGDP offer a way to notify of dataset errors, but they might operate under a one-way paradigm, without any compromise to notify users back about any measures taken to address citizens' requests. Up to 75% of OGDP offer no mechanism to notify of recent changes on the datasets, thus Open Data publishers are acting as black boxes, changing the published information without any further details. How could others build upon these unstable foundations?

Blatant errors in published open datasets, can have cascading effects on the surrounding data ecosystem. This opaqueness and randomness can introduce mistrust and disaffection into citizens[6], limiting the potential impact of open data initiatives and eroding confidence in the public sector.

In Spain, several citizen initiatives have launched to tackle problems in published data, focusing mostly in obfuscated or hard-to-process formats. One of the most famous was "Adopta un senador"[5], a crowd-sourced process to extract information about Spanish senators' patrimony, released as thousands of images under PDF format. More recently, Jaime Gómez Obregon has devoted the last two years to parse and release information about public contracts in Cantabria. [7] He is working now on a national version, highlighting the challenges and blockages to release all this information in an accessible way.

Most of these actions are either short lived or quite individualistic. Development speed is key, therefore most tools used are not shared with the broader community (or properly documented). Outraged citizens get triggered by some of the discoveries, but can't participate in a constructive way besides retweeting the original threads. In most of the cases, there is not a formal complaint to the governing bodies, so there is no legal mandate to act upon.

In recent years, a new discipline called data engineering is emerging. It focuses on extracting, validating, and preparing data for analysis[8]. In this modern data stack, data engineering processes are often visualized as pipelines, where data is ingested on one end, checked, cleaned, and transformed, before being released on the other end. Given the need for transparency and observability, most of the available tools in this area are offered under open licenses. Furthermore, pipeline elements can be managed as code, created and maintained collaboratively.

2 AIM: RESEARCH QUESTIONS

In this paper, we aim to propose a collaborative approach to achieve better open data quality through citizen engagement. We will explore the principles and tools of data engineering focusing on the following research questions:

- Can data engineering tools be used to evaluate the quality of existing open data?
- Can these tools be used to provide collaborative, transparent and scalable workflows?
- How could citizens start using these tools and engage with the process?

3 METHOD

We will start doing a literature review to isolate those metrics already used by (open) data quality evaluation frameworks that can be computed. Taking a proven open source data engineering architecture as the basis, we will create a sample pipeline to ingest a few datasets from OGDP and return specific metrics.

Based on these initial tests, we will explore how to manage this architecture under a collaborative paradigm, employing the mechanisms provided by the tools used. We will also address how to raise alerts when we discover errors in the datasets. We will propose a workflow to send and track formal notifications to the responsible public bodies. Finally, we will provide a set of metrics to verify data quality improvement.

4 RESULTS

This study will provide the following results:

- Several applications of data engineering principles applied to open data quality evaluation.
- An open source architecture to accomplish this task in a collaborative way, including the roles for different actors.
- A workflow to evaluate and notify of errors in open datasets.
- One or more metrics, to keep track of improvements.

These outcomes will provide a reliable method to reduce the number of errors in open government datasets. They will also foster transparency and trust, providing a standardized way to notify public bodies and keep track of their responses.

5 CONCLUSION

In this paper, we will evaluate several popular open source tools in the data engineering field, assessing their usefulness to evaluate open data quality. We will also create an initial architecture to support collaborative workflows.

Based on the this reference architecture, we expect that new researchers and public bodies will start incorporating these data engineering tools into their work spaces, sharing their expertise back with the broader open source community. We also hope that

some modern data stack practitioners will also become interested, and eventually engaged, into Open Data ecosystems.

Further field work will be necessary to test these outcomes with active data user groups, open data advocates and public administrations. This will allow us to address the following open questions:

- How could we make this process as simple as possible for new users?
- What kind of improvements can we observe after applying these tools over long time periods?
- Does this approach contribute to citizen engagement, reducing disaffection and mistrust?

All these topics will be addressed in future research.

REFERENCES

[1] Alberto Abella, Marta Ortiz-de Urbina-Criado, and Carmen De-Pablos-Heredero. 2022. Criteria for the identification of ineffective open data portals: pretender open data portals. *Profesional de la Información* 31, 1 (2022).

[2] Pieter Colpaert, Sarah Joye, Peter Mechant, Erik Mannens, and Rik Van de Walle. 2013. The 5 stars of open data portals. In *Proceedings of the 7th International Conference on Methodologies, Technologies and Tools Enabling E-Government (MeTTeG13), University of Vigo, Spain.* 61–67.

[3] Ministerio de Asuntos Económicos y Transformación Digital. 2021. España continúa entre los líderes del open data en Europa un año más. https://datos.gob.es/es/noticia/espana-continua-entre-los-lideres-del-open-data-en-europa-un-ano-mas

[4] Susana de Juana-Espinosa and Sergio Luján-Mora. 2019. Open government data portals in the European Union: Considerations, development, and expectations. *Technological Forecasting and Social Change* 149 (2019), 119769.

[5] Javier de la Cueva González-Cotera. 2012. Praeter Orwell: Sujetos, acción y open data ciudadana. *Argumentos de razón técnica: Revista española de ciencia, tecnología y sociedad, y filosofía de la tecnología* 15 (2012), 13–37.

[6] Manuel Gértrudix, María-Carmen Gertrudis-Casado, and Sergio Álvarez-García. 2016. Consumption of public institutions' open data by Spanish citizens. *El profesional de la información (EPI)* 25, 4 (2016), 535–544.

[7] Jaime Gómez-Obregón. 2020. Contratosdecantabria.es. https://contratosdecantabria.es/

[8] Meike Klettke and Uta Störl. 2022. Four Generations in Data Engineering for Data Science. *Datenbank-Spektrum* 22, 1 (2022), 59–66.

[9] Sylvain Kubler, Jerermy Robert, Sebastian Neumaier, Jürgen Umbrich, and Yves Le Traon. 2018. Comparison of metadata quality in open data portals using the Analytic Hierarchy Process. *Government Information Quarterly* 35, 1 (2018), 13–29.