

Open Data for Governance and Research: Unlocking Urban Insights

Giorgio Piccitto (Università degli Studi di Milano-Bicocca); Domingo Scisci (Università degli Studi di Milano-Bicocca); Sonia Stefanizzi (Università degli Studi di Milano-Bicocca)

Open data is a tool with enormous potential (Gurstein 2011; Krotoski 2012), both to implement open government and participation practices, and to improve the effectiveness of academic research, by increasing data transparency, accessibility, cost-saving and reproducibility (Boulton et al., 2011).

Such data becomes particularly crucial in understanding new trends in cities' governance, grounded in the current general trend towards data-driven decision-making (Neves et al., 2020). However, in this scenario of hectic development of digital technologies and information-intensive productive processes, several impediments have been recognised that prevent the open data process from fully realising its positive effects (Zuiderwijk et al. 2021).

In this context, we will present the "Open Government Data" project, which aims to extract value from administrative data and provide the scientific community with new tools to understand the governance processes of local institutions. Starting from the open data provided by some Italian cities and regions, the project aims to create new ready-to-use research data to be made available to all researchers.

The working method is based on two specific aspects.

Firstly, typical data curation practices will be applied, such as using metadata, documentation and distribution standards that will guarantee the FAIRness (Briney, 2015; Wilkinson et al., 2016) of the data.

Secondly, the "code-first" approach will be adopted, which consists in focusing attention and work on the code that generates the data, as opposed to the data themselves, facilitating the activities of documenting the operations performed, in a transparent and collaborative manner (Gentzkow and Shapiro, 2014). This approach allows great improvements in terms of reproducibility and accountability, since it keeps track of what has changed, when, by whom and why.

Consistently with this strategy, the entire data life-cycle from raw to public data has been managed following the typical software development workflow. First, a repository managed by a version control software (Git) was created, which ensured the tracking of all changes and made collaboration easier. The code was hosted in a repository manager (GitLab) that allowed the research team to track issues, bugs and requests collaboratively, linking issues to the code used to

re-solve them.

This procedure enabled the rapid and frequent release of data to the research team, still in the field phase, following an agile-like methodology. Such DevOps-oriented approach also allowed the definition of standardized procedures for setting up the survey instruments (questionnaires), technical documentation and metadata needed for long-term preservation.

This second aspect - which consists of treating the entire data lifecycle according to the typical software development workflow - represents one of the most original aspects of the project, which will guarantee its sustainability even after its completion. The strengths and weaknesses of this approach will be discussed, particularly highlighting the need for more IT skills in the social sciences and for new professional figures who can integrate technical skills and domain-specific knowledge.

The project aims to define and propose a new rationale to produce and interpret the Public Administration Open Data (PAOD), to overcome all those accessibility issues that prevent the full realization of the Open Science (OS) paradigm and that reproduce structures of social inequalities (on the basis of individual material and competencies resources).

Specifically, three broad goals are defined: 1) to establish the Public Administration Open Data Observatory; 2) to develop new research lines, inspired by and coherent with the new proposed 'open' and 'accessible' way of conceiving data; 3) to create the Innovative Teaching Lab, through which will be developed some good practices in terms of multimedia production and MOOC (Massive, Open, Online Courses) provision.

The project is explicitly targeted to increase social and economic efficiency and sustainability in local social contexts (Vega-Tinoco et al., 2022). Indeed, the availability of open data contributes to the citizens' empowerment and mobilization and can be used as a benchmark to ask for public policies and interventions and to improve the effectiveness of evidence-based intervention.

Several promising concrete applications are expected, allowing to overcome some of the most known problem related to the open data quality and accessibility, especially in terms of completeness and accurateness of the information. One objective will be the systematization of the information at a sub-city level. Concretely, it will be created a database of individual socio-demographic characteristics at the 'town halls' (Municipi) level in Milan, that would be freely consultable.

This database will trigger new avenues also in Academia, permitting research at a very fine-grained territorial level. These datasets, indeed, could be used to pass from the aggregated to the micro level, recreating a dataset at the individual level by weighting each combination of the combined

variables. Such a method, recently applied (with Eurostat data) in a cross-country study (Ortensi et al., 2024), could be adopted in the study of urban phenomena allowing the use of multivariate analyses techniques.