Harvard Data Science Review • Issue 5.3, Summer 2023

# Open Data and Code at the Urban Institute

**Graham MacDonald**[1]

[1]Office of Technology and Data Science, Urban Institute, Washington D.C., United States of America

**Published on:** Jul 27, 2023

**DOI:** https://doi.org/10.1162/99608f92.a631dfc5

# Background

The Urban Institute is an organization whose mission is to provide evidence, analysis, and tools to people who make change to ultimately empower communities and improve people's lives. We define 'people who make change' broadly as policymakers, government agency employees, advocates, community leaders foundations, corporate leaders, and other similar actors.

Though Urban as an organization has a number of goals, I would categorize our primary drivers as 1) to make impact toward our mission and 2) to fundraise effectively to support that impact and the organizational supports that make it possible. This is important to consider later when I discuss organizational priorities around open data and code.

Similarly, Urban conducts work broadly across many policy areas, however, I would summarize them succinctly as 1) conducting policy research and evaluations, 2) providing technical assistance on implementation, 3) producing data and data tools, 4) providing advisory services, 5) convening experts across sectors, and 6) translating research and communicating it to targeted audiences. In support of this work, Urban sometimes posts both the data and code powering the data on its website, Urban.org.

# Main Thoughts

## Existing Initiatives

Urban is home to a number of existing initiatives intended to make progress toward more open data and code. The first is Urban's Data Catalog, to which all researchers who wish to publish code on Urban's website must submit their data and document their submissions to a minimum extent. The second is Urban's central library of quality assurance materials and trainings, which promote open science standards, reproducibility checks, automation in programming, clear organization, and quality documentation throughout. The third is Urban's automated R (and soon Stata) themes, templates, and guides, which allow researchers to more easily automate the full research pipelines from data collection to publication in R. Urban has processes in place to comply with the requirements of third parties, such as the American Economic Association (AEA) Data and Code Availability Policy for submission to AEA journals and Inter-university Consortium for Political and Social Research requirements for submitting data to their third party digital data repository on behalf of grantors such as the National Institute of Justice, among others, to whom Urban is required to submit or may submit voluntarily as part of funded project requirements. These organizations have specific requirements for submitting data, codebooks, and code that facilitate replication and further use.

And finally, Urban has a central team of approximately 20 data scientists and research programmers who are available to conduct code reviews and reproducibility checks.

Urban continues to make improvements in all these areas, including adding supporting resources for quality assurance, such as an internal code library, providing additional documentation and examples on GitHub for certain projects, improving our automation of publishing systems to extend to additional content on our website, and revamping and improving our data catalog experience.

## Successes

As a result of these efforts, Urban has seen a number of successes that have led to substantial benefits to the organization. For our external users and partners, our publicly available data are now better documented, with a clear license for use; citations are clear and available; and our impact through open data is easier to see and track. For example, we are able to more easily view exactly which resources have been accessed and downloaded, interact with users through our centralized datacatalog@urban.org email address, and conduct follow up with these users and via citation searches online periodically to determine the use and impact of specific data sets.

We have a clearly established Chief Data Office within our Office of Technology of Data Science, and data initiatives and day-to-day operations, including our centralized email and administrative functions, are supported by our 20-plus strong team of data science and research technology experts. All these efforts are advised by a standing council of nine, which includes members of our Technology and Data Science, Communications, and Research teams, and is led by our chief data officer. This centralized effort has led to increased growth and usage. Our users range from researchers to local community groups to policymakers across the spectrum, and have grown from an average of 500 unique monthly users at our launch in late 2019 to an average of more than 1,500 unique monthly users in early 2023. Similarly, we have tripled our number of unique page views from around 2,000 to 6,000 in that same timeframe, on average. From an initial start of 20 external data sets, we have grown to 86, a number of them regularly updated and maintained, as of this writing.

Better quality assurance materials and process automation have led to a more streamlined review process that saves time and allows for rapid iteration and even innovation under tight timelines. The processes and systems in place have also allowed for more redundancy and reduced stress on team members when they work well, especially when these efforts resulted in improved documentation, onboarding, communication, and collaboration across teams with diverse skillsets and backgrounds.

## Challenges

Despite our efforts and successes, significant challenges remain. Our organization is decentralized and funded by many different parties across government, philanthropy, and the private sector. Many of these funders in recent years, especially in the philanthropic sector, have shifted their focus away from core research and more toward work that generates impact. I worry that this shift will lead open data and code efforts to be seen increasingly as ‘optional’ when so many other ‘more impactful’ activities are vying for the next marginal

funding dollar. In any case, as a result of this landscape, these open practices are only adopted on a voluntary basis or in certain cases where required by the funder or journal at Urban.

Researchers and organizational leadership are not always directly incentivized, outside of the few funding requirements we do observe (such as from the Sloan Foundation, the Arnold Foundation, the National Science Foundation, and others), by existing priorities to tackle these challenges. While centralized quality assurance and open code are seen as a priority at Urban among all staff, in terms of centralized organizational initiatives, they have been at times overwhelmed by even higher priorities, especially in light of my thoughts on funders' changing priorities in this space.

In the meantime, researchers continue to prioritize quality control in their own decentralized individual projects and efforts. However, in my experience the majority are not motivated by quality-control arguments to adopt newer open data and code practices, and the strongest motivation remains funder and journal requirements. Most researchers I work with see their work and current processes as high quality, just defined differently across the organization and their respective fields. More importantly, in my view, is that open data and code efforts are often seen as an additional layer of bureaucracy and busy work on top of existing requirements, and thus they are perceived as reducing the agency and academic freedom that many researchers highly value.

## Conclusion

From my view here at Urban, which operates on a unique mix of government, foundation, and corporate funding resources, more openness and reproducibility is unlikely without increasing the requirements on researchers and institutions from those funding them and disseminating their work. Ultimately, despite the short-term perceived costs of increased bureaucracy, I believe these requirements will bring larger benefits to the field and are worth considering.

It would be wise for advocates of open data and reproducible research to call for funders and journals to require reproducibility checks at a minimum, and open data where possible as a next step. I would also be in favor of third party reproducibility checks and/or marks that certify that a third party check has been passed and certain materials are available for reproduction.

These requirements would improve our clients' confidence in the quality of the complying institutions, and clearly help us to differentiate important policy signals from the noise. It would also pave the way toward a future where more replication studies are feasible. Urban currently has systems, processes, and materials in place to comply with these requirements, and the field now has sufficient examples from peer organizations and journals to enable the rapid spread of best practices once requirements are in place.

## Disclosure Statement

This article is an expanded version of remarks given in the Conference on Reproducibility and Replicability in Economics and the Social Sciences (CRRESS) webinar series, which is funded by National Science Foundation Grant #2217493.

---