Dear Chair and Committee Members,

Thank you for requesting that I testify today about these proposed amendments to the city’s open data law. My name is John Krauss, and I am the technology fellow at the GovLab at NYU, whose mission is to “improve people’s lives by changing the way we govern,” including through the “increased availability and use of data.”

New York City is one of the leading open data cities in the United States. By sheer number of datasets, New York City publishes more than twice as many on its portal as Chicago, which is second in number of datasets, and three times as many as San Francisco, which is third.¹

The numbers are impressive, and New York, spurred by 2012’s open data law, has published a large number of datasets.

What we’re discussing today are a handful of improvements to that law, many of which were inspired by prior testimony from the open data community that brought flaws in the published data to light.

Within the open data community, there has been agreement for years that unanswered questions and unresolved problems in the data are impeding the use of the open data portal. To understand this issue, we downloaded all 293 comments from the portal 2011 to the present and analyzed them.² We broadly divided the comments into four categories:

1. Problems needing a fix, such as, “70 poll sites on this list have blank entries for the LOCATION field.” (70 comments)
2. Questions needing an answer, such as, “Will this be updated regularly?” (106 comments)
3. Comments that need no response, such as: “Hard to trust any of this data. Very few respondents.” (32 comments)
4. Responses, such as “Thank you for your feedback.” (82 comments)

¹ Count of datasets on municipal Socrata open data portals, from http://www.opendatacache.com: as of 10/1/2015, New York City has 1,363 datasets, Chicago has 573, and San Francisco has 438.
² All data can be viewed at https://docs.google.com/spreadsheets/d/1ZqIEfl7h-CXnbHcOX-i4l3fp3VcPkwqKMB4B9bQrg/edit?usp=sharing
Slightly less than half of the questions (51 of 106) were answered. On average, half a year (180 days) elapsed between question and answer. The average unanswered question on the portal is over a year old.

Actual problems fared worse. About two thirds (46 of 70) haven't been fixed. On average, unfixed problems were reported over a year ago. Over half (28 of 46) never even received a confirmation that the problem was being looked into. Of unfixed problems that were confirmed (18 of 46), on average over a year has elapsed since confirmation without a fix.

Int. 916 (data compliance audit), Int. 915 (portal must be up to date with website), Int. 900 (geospatial format), Int. 898 (data dictionary), and Int. 890 (data retention) all address concerns problems brought up by open data portal users. As others testifying today will tell you, simple compliance with the Technical Standards Manual would fix most of these issues. New legislation requiring data conform to standards it was already supposed to conform to is redundant.

What is needed are people who can respond to comments by actually fixing the issues brought up in them. We need people to fix this data, not more requirements it be fixed.

The current arrangement, where DoITT manually ferries reports of problems to agencies that may or may not have resources to resolve them, does not work. We need open data expertise on the front lines, responding to these problems and empowered to fix them. This means funding for staff and resources within an organization dedicated to open data. It also means a private right of action by open data users to put them in the same class as FOIL applicants, whose requests must be responded to.

I want to emphasize that DoITT is doing their best to resolve questions and concerns. Due to deficiencies in the open data portal, keeping track of comments is difficult. Without a clear mandate to fix problems at the source, oftentimes there is nothing more that they can do than tell the user that the agency has been informed.

Thank you for your time, and please do not hesitate to contact me for more information at john@thegovlab.org.

John Krauss

Appendix

Below are specific modifications we would recommend to the proposed legislation.

Int. 914:
• Comments that merit a data fix should require one, or an appropriate response explaining why a fix is impossible.
• All comments and requests, as well as their responses and resolutions, should be made available in a machine-readable format.

Int. 890:
• Large amounts of data will still be lost if archival is only required when “a majority of the data contained” changes or is removed.
• Many datasets are updated monthly or annually. Even in cases where only a small percentage of data changes, archival copies of all these should be kept.
• For datasets updated more frequently, a changelog of removed data should be maintained using free software such as dat\textsuperscript{3}. This would allow reconstruction of older data without keeping a prohibitive number of redundant copies.

\textsuperscript{3} http://dat-data.com/