



Testimony of

John Kaehny, Co-Chair NYC Transparency Working Group

Executive Director, Reinvent Albany

New York City Council Committee on Technology

Open Data Oversight and Int 1707-2017, Int 1528-2017

September 20, 2017

Good afternoon Chair Vacca and members of the Technology Committee. I am John Kaehny, Executive Director of Reinvent Albany and co-chair of the NYC Transparency Working Group.

Thank you, Chairman Vacca, for your energetic advocacy for open data. This is probably your last open data hearing, and, on behalf of the Transparency Working Group, I want to thank your great work and hope we can send you off in style.

This hearing is both an oversight hearing of ongoing Open Data Law mandates and a chance to comment on two bills, Int 1707-2017, Int 1528-2017.

Oversight

First, oversight --- what's working and not working. The good news is that New York City has an open data culture, a public expectation that government data will be online

in a usable form. In 2012, New York City passed the first Open Data Law in the world, and thankfully, the Mayor and Council support this and fund a smart Open Data Team. City Council cares and does effective oversight. City agencies have published about 1,600 datasets, including 200 automated datasets. Let me flag that. We think automating data sets is the single most important thing that the Open Data Team can do to make Open Data sustainable. (“Automated” datasets automatically update on the Open Data Portal from a City computer.) Every dataset that can be automated should be, and the Open Data Team should have an automation plan and public targets.

Another key to open data is getting City agencies to use it and understand how it makes their work easier. Most experts believe that ultimately, by far the biggest consumer of government open data will be government. We expect City agencies to save tens of millions a year from open data as agency staff spend less time looking for information for their own everyday use and responding to Freedom of Information requests.

We are seeing the insights and information provided by Open Data in everyday public life. Data sets like 311, NYPD crime data and DOT’s traffic crash data are a regular feature of neighborhood and community board meetings, City Council meetings and news articles. Digital businesses are using Department of Building data to instantly inform building owners of violations, and transportation planners are using TLC trip data sets to map neighborhood activity. Just yesterday, the national blog, Five Thirty Eight, ran a great story about the lingering effects of Hurricane Sandy based on calls to 311 pulled from the Open Data Portal.

Specific Oversight Issues

Last year, the Council passed seven amendments to the Open Data Law. To track whether a dataset is complying with these mandates, one has to look at four separate spreadsheets. We have not merged the spreadsheets or done detailed analysis, but we find this form of reporting extremely cumbersome and time consuming. We strongly recommend that information about all public datasets, whether they are published or unpublished, be listed in a single dataset and all information about the status of that dataset be included data fields. Otherwise, monitoring agency compliance is impractical.

Compliance Data

- Data Dictionaries: 1648 listed. 615 data sets have data dictionaries, 1033 do not.
- Address / geospatial Standardization: 350 listed. 206 have standardized.

Publishing

The core of the 2012 Open Data Law is a requirement that agencies publish all of their “public datasets” by the end of 2018. The agencies are supposed to propose an annual publishing schedule and then report on their progress. Unfortunately, the City’s reporting is very confusing. The public can see what is actually on the portal, but it is hugely time consuming to try and track whether agencies are meeting their publishing targets. We looked at six datasets that agencies said they would publish in 2014. Parks Department *Street Tree* (two data sets), FDNY Fire *Incident Information*, School Construction Authority *Funded Capacity Seats* and HRA *Cash Assistance Engagement*.

Four of those appeared in later publishing plans, but then never appeared in the Open Data Portal. What happened to this data?

Agency Procrastination

It is fairly obvious that many agencies are procrastinating and delaying the scheduled publication of their public data in the hope that day will never arrive. Agencies have a whopping 102 data sets scheduled for publication in the second half of 2018, including 70 scheduled for December 2018. This is baloney. It's like a highschool kid saying they will turn in all of their homework on the last of school.

Intro 1528-2017 (Open Data/FOIL) and Intro 1707-2017 (Open Data Update)

Intro 1528-2017

We strongly support Intro 1528-2017 which amends Local Law 7 of 2016 to add the name of the public dataset agencies use when they reply to a Freedom of Information Law request for tabular data.

Intro 1528-2017

We have extensive comments on Intro 1528-2017, some of which we conveyed to Chair Vacca and the administration prior to this hearing. We believe this legislation should have three main goals:

1. Clarify that the Open Data Law will continue past 2018.
2. Strengthen the mandate for agencies to continue to publish public data sets.

3. Foster a sustainable open data process by promoting the automation of data sets.

There is universal interest in ensuring that the Open Data Law continues past the December 2018 publishing deadline. The 2012 Open Data Law established a six-year publishing schedule that has helped to prod agencies. While it made sense initially, we believe that a three year extension of the publishing schedule to 2021 will give an excuse for agencies to procrastinate on publishing their public data even more than they already have. We suggest the following:

Recommendations:

- Keep 2018 deadline. Delete any mention of 2021 extension.
- Draft new bill language that makes it clear that the 2018 deadline is still in place and that public data sets created after that time must be published within twelve months* of their creation. (*Twelve months seems reasonable, but this would best be determined in consultation with the administration.)
- Require an annual update / compliance plan that summarizes status of published, scheduled to be published and delayed public data sets, including compliance with mandates created by the Open Data Law amendments.

Strengthen existing mandate to publish data sets by requiring a new “*Status of All Public Data Sets*” dataset on the Open Data Portal:

Because there is no public right of action in the Open Data Law, the public relies on Council oversight and the “naming and shaming” of laggard agencies to get public data

sets published. Unfortunately, there is no single dataset of public data sets that lists all agency data sets and their publishing status or compliance with open data mandates. (Data dictionaries, geospatial etc.) There is no complete list, or easy way to track datasets that have been delayed for multiple years. This lack of transparency, reduces accountability and reduces pressure on agencies to comply with the Open Data Law.

Recommendations re: new “Status of All Public Data Sets” dataset:

- Add bill language that mandates the creation of a single, *Status of All Public Data Sets*, dataset. This Status dataset should include all datasets that have been classified as public datasets and data fields that include all information about that dataset. There should NOT be multiple datasets tracking compliance with the Open Data Law Amendments and Annual Plans. This one dataset should list all information about publishing status and mandate compliance for every dataset as fields in a master list. The *Status* dataset should include all information about all public datasets including the following fields for each dataset:
 1. Original scheduled publication.
 2. Current scheduled publication.
 3. Publication date (if published).
 4. Current status of dataset if delayed and reason for delay.
 5. Reason for unpublishing from portal (if unpublished.)
 6. Compliance with mandates established under the Open Data Law Amendments including data dictionaries, geospatial data etc.

Foster sustainability of Open Data Initiative by automating datasets

We believe it is hugely important to automate as many data sets as possible.

Recommendation: Add bill language that requires the following fields in the *Status of All Datasets* dataset:

1. Whether dataset is automated.
2. Whether dataset can be automated.
3. When dataset is scheduled for automation (if it can be.)
4. Reason for not automating if dataset can be automated.

Comments on other items included in the bill

We support --- review of Technical Standards Manual every two years

We oppose --- extending publishing deadline to 2021 (see above).

We support --- change of compliance plan deadline from July to September.

We support --- Designating agency open data coordinator.

We support --- publishing web portal site analytics

Public Right of Action

Reinvent Albany and our colleagues in the NYC Transparency Working Group strongly support as private right of action be included in the Open Data Law. When the Open Data Law was originally proposed in 2011/2012 it included a private right of action, which was removed after adamant opposition from the Law Department. We again note that the Open Data Law is one of a small handful of NYC laws that does include a private right of action.