Reinvent Albany

Open Data Year Three

New York City’s Open Data Law

October 27, 2014

Executive Summary

Three years into the New York City Open Data Law of 2012, the City's open data efforts are robust, healthy, and improving. The implementation of the Open Data Law is fundamentally strong and evolving nicely. With the city's open data initiative still in its infancy, there is tremendous potential for open data; we're still in the early days of open data.

The implementation of the Open Data Law is fundamentally strong for four specific reasons:

1. The City has five fully funded, dedicated open data staff. This is probably more than any state or local government, and gives the City the capacity to rapidly improve and expand its open data efforts.

2. Open data has strong support from the Mayor and City Council and has a synergistic relationship with the Mayor's Office of Data Analysis (MODA) which ensure it is sustainable.

3. Open data has an expanding group of public stakeholders, including businesses, academics, advocates and government who use the data.

4. The quantity and quality of the data available continues to expand, and with it, so does use of that data.

Despite the tremendous amount of open data already released to the public, the city is barely scratching the surface of the open data law's potential; open data as a civic institution is still so new that opportunity is everywhere. In the next year, we hope the de Blasio administration advances open data in four key ways:

First, the Mayor's Office of Operations needs to get agencies to put the most frequently FOILed and requested data on the Open Data Portal. Most agencies still do not understand that the Open Data Law is intended to help them reduce FOIL requests, reduce 311 requests for information, and help them get information to the public at a lower cost with less hassle.

Second, DOITT needs to create a public feedback process for the Open Data Portal which results in more and better city data being put online. When data errors are discovered and reported by the public, the responsible agency should correct those problems. DOITT should ensure that the public can track the progress of those corrections.

Third, the severe problems with the search function of the Open Data Portal have to be completely fixed. Despite repeated, emphatic, requests from the NYC Transparency Working Group and other many open data stakeholders, it took DOITT more than two years to correct crippling flaws with the search function – flaws that severely reduced the usefulness of the Open Data Platform, and undoubtedly kept people from using the data on it.

Four, DOITT needs to clearly show the status of data sets to be published, or which have been delayed or removed from the data publication plan. Overall, it hard for the public to tell if the City is meeting its own data release targets. For example, a dataset essentially disappears from public view if it was scheduled for release in the 2013 plan, delayed, and then not included in the 2014 plan.

Platform Usability

Full Text Search

While the search function on the open data portal received its largest-ever upgrade last week, the search function still needs improvements. Even technologically proficient users have found it difficult to locate specific data sets. The Socrata software powering the portal cannot search just the names of data sets, or just the descriptions of data sets. Any search is performed on the name, description, and entire contents of every data set simultaneously.

This means that a search for the 311 Service Requests data set turns up thousands of other data sets, including every data set where the street address begins with "311" (e.g. 311 East 3rd Street) or where a police precinct has recorded exactly 311 assaults, or where a school has exactly 311 students, or when parking signs mention that you can call 311 for more information, and so on.

After last week's update, the top results for a search for 311 are now data sets with 311 in the title, but the underlying problem still exists.

Metadata

Technologists have pointed out deficiencies in the metadata, (the descriptive information about each data set) which greatly reduces the usefulness of the search function. The data portal has no uniform practice for displaying metadata. Some data sets put the entire metadata in the dataset's description on the open data portal, some contain a link to download the metadata in a Microsoft Word file in another location, and some data sets contain no links to any metadata at all.

This is the sort of problem for which a robust feedback process would be extremely helpful.

Unofficial Datasets

The open data portal allows users to share "filtered views" of data sets; if a user is interested in 311 requests in their neighborhood, they can filter the other requests and view just the ones they care about. However, since its launch in 2011 until last week's update, each filtered view was presented alongside the official data set, and showed up in the same search results as the official data set.

There are nearly twice as many filtered views as official data sets on the open data portal. Some particularly popular data sets have dozens of separate filtered views, which often made it virtually impossible to find the actual (official) data set. These filtered views are created by users of the site, not by the city, and are actually subsets of the official data sets. The only indication that these are unofficial (and incomplete) data sets is a tiny purple icon on the edge of the page.

Displaying official and unofficial data sets alongside one another was a mistake we were happy to see resolved in the last update to the open data portal. Trying new ways of displaying information is a worthy goal, but the public comes to New York City's open data portal because they want New York City's data. It is not government's job to host transparency enthusiasts' data, and we are delighted to see that option on the portal.

Browsing

For years, the lack of a working search function made the open data portal difficult to navigate. Unfortunately, the portal was difficult to manually browse, because the portal could not show data sets published by a single agency. Last week, Socrata added the ability to filter data sets by agency: this is another feature we are happy to see added to the city's portal.

Feedback

One area with room for improvement is the public feedback cycle. Currently, users who report errors in data sets or request new data sets get no response from DOITT for months, if ever. With thousands of datasets, each serving as its own self-contained conversation on the open data portal, it's difficult to tell what feedback DOITT is reading, reacting, and responding to. There's no easy way to know how many comments has DOITT received, what the most common requests are, and how DOITT has responded, if at all. It's unclear if DOITT has this ability, either.

Among data sets with the most comments, it appears that DOITT has not provided many comments in the last 8 months. Many comments are users voicing concerns that the data hasn't been recently updated. Most of the other questions are about data quality, usage, or other resources on the portal. Users can provide feedback, point out errors in the data sets, and ask questions of DOITT, but it will be difficult to foster a community on the open data portal without more visible and consistent engagement from the city.

Evaluation of 2014 Open Data Plan

The July 2014 update of the Open Data Compliance plan scheduled 345 data sets for release.[1] The schedule is available in a machine-readable format from the open data portal, which we again applaud as an excellent application of the portal. This update includes the scheduled release dates for datasets from July 15, 2014 through December 21, 2018.

Our preliminary analysis found that of the 60 data sets scheduled for release between the issuance of this plan in July and October, only 10 have been uploaded. Perhaps the remaining data sets have been uploaded with different names than the one in the open data plan. Perhaps agencies are having trouble meeting their self-imposed schedule for publishing data. Either way, transparency advocates understand that there will always reasonable delays in releasing some datasets; it's important that those delays are acknowledged and accounted for, so an open discussion can be fostered about how to fix these problems.

[1] NYC Open Data Dashboard: https://data.cityofnewyork.us/dashboard

Of the ten data sets which have been published since July, five of them are incomplete; for example, one dataset updates with a monthly instead of a weekly frequency, and another dataset does not update at all, but is a spreadsheet from 2012. Three other datasets do not match the title given in the Open Data Plan Update, and so it is not clear if these were published according to the plan or not. (See table on Page 6.)

High Priority Data

Summary

Since the passage of New York City's Open Data Law in 2012, thousands of data sets have been published on the open data portal. This is one of the biggest, if not the biggest open data repositories in the country. Agencies have done their best to publish high-value data sets. However, agencies' ideas of high-value aren't necessarily the same as the public's ideas of high-value.

Agencies don't need to guess what the public thinks is high-value data. The public is telling government exactly what high-value data we want online, via Freedom of Information Law requests. Agencies should be examining their FOIL logs to determine what data is most frequently requested, and publish that data to the portal.

Portal's Low Participation

FOIL is where the public tells agencies what they want, but agencies are listening to the comments form on the open data portal. The portal accepts public requests for new data sets via a built-in suggestion form. In the three years since its launch in November 2011, the open data portal has had 109 suggestions via the site. Of those suggestions, DOITT has published 5 data sets.

To put those 109 suggestions in context, New York City agencies have received over 120,000 FOIL requests since November 2011. This is according to Public Advocate di Blasio's 2013 report on FOIL, which found that city agencies receive over 40,000 FOIL requests a year.[2]

[2] Public Advocate Bill de Blasio, *Breaking Through Bureaucracy*, April 22, 2013: http://advocate.nyc.gov/sites/advocate.nyc.gov/files/deBlasioFOILReport_0.pdf

With 109 requests in 3 years, there just aren't enough comments on the built-in suggestion form to provide a meaningful impression of high-value data. If this is the metric agencies are using to gauge public interest, they're missing a tremendous opportunity for hundreds of thousands of requests for what to publish next.

Table 1: Datasets Published after July 15, Per Open Data Plan

Agency	Dataset	Update Frequency	Planned Release Date	Comments
Department of Information Technology and Telecommunications (DoITT)	Public Pay Telephone locations and Wi-Fi pilot usage.	Weekly	07/15/2014	Data is updated monthly, not weekly
Department of Transportation (DOT)	CitiBike System Data	Monthly	08/01/2014	External link to citibike.com
Office of the Mayor (OTM)	City GHG Reduction projects	Quarterly	08/01/2014	Complete
Office of the Mayor (OTM)	Sustainability Indicators	Annually	08/01/2014	Yes - but, only for 2012 and it is a static spreadsheet.
Department of Homeless Services (DHS)	Daily Report	Daily	08/23/2014	Complete
Department of Cultural Affairs (DCLA)	Cultural Institutions	Annually	09/08/2014	May be the wrong data set.
Department of Cultural Affairs (DCLA)	Capital Projects	Annually	09/08/2014	Related to Capital Grants Funding?
Department of Environmental Protection (DEP)	CCR Annual Report	Annually	09/30/2014	May be the wrong data set.
Department of Health and Mental Hygiene (DOHMH)	Restaurant Inspection Data	Daily	09/30/2014	Complete
Department of Investigation (DOI)	Mayor's Management Report Performance Statistics	Annually	10/01/2014	Complete

Listening to FOIL Works

New Yorkers have relied on Freedom of Information Law requests to get information from their government for the last 40 years. These FOIL requests tell us what information is most often requested, and who the agency's largest (or loudest) constituencies are. While FOIL requests must, by definition, ask for records, many of those requested records are in fact digital data accessed by FOIL officers responding to requests.

This isn't just a theory. The Federal Environmental Protection Agency has repeatedly testified to Congress (as early as 2010) that, in order to cut down the number of FOIL requests they have to process, they publish frequently requested records on their web site.[3] In New York State, in June 2014, the Department of Environmental Conservation provided Reinvent Albany with a spreadsheet listing the 3,977 FOIL requests it received in 2013. A majority of FOIL requests to the DEC are actually for records from the same few data sets over and over again.

Our straightforward analysis of DEC FOIL logs show that agencies can and should use FOIL to identify which data sets the public is most interested in. Putting these data sets online in an open data format would reduce the number of FOIL requests to DEC by 2,200, a reduction of over 55% of requests in 2013.[4] Listening to FOIL works at the federal level, it works at the state level, and it will work here in New York City.

We recommend that agencies and DOITT's open data team undertake the following steps to use FOIL to power open data:

1. Identify which agencies keep FOIL logs identifying the topic of the request.
2. Select the agency with the most FOIL requests and analyze its FOIL log.
3. Publish that agency's public data based on the FOIL log analysis.
4. Write guidance to agencies explaining how they can analyze FOIL logs to help them determine what data to put on the open data portal.

[3] Government Executive, *Posting Information Online Could Preempt FOIA Requests*, March 18, 2010: http://www.govexec.com/oversight/2010/03/posting-information-online-could-preempt-foia-requests/31089/

[4] Reinvent Albany, *Listening to FOIL,* July 23, 2014: http://reinventalbany.org/wp-content/uploads/2014/07/Final-DEC-FOIL-Analysis.pdf

Cost Effective

Performing this kind of analysis on FOIL logs is not expensive. Agency FOIL officers are already deeply immersed in the records and data sets that the public asks for, and will probably be able to name a handful of frequently FOILed records without referring to the logs. This is no replacement for a comprehensive analysis, but it will start the process moving forward. For overburdened FOIL officers, cutting their workload in half is an enormous benefit. For the public, getting high value data published on the open data portal sooner is a huge win.

Specific Examples

While DOITT and agencies collaborate to find the most frequently FOILed data sets, there are a few high-value data sets which our groups would like to see added to the Open Data Plan.

There are many data sets relating to agency performance, such as AgencyStat data, which are still not published on the open data portal. The Bloomberg administration worked hard to collect agency performance data and put it to work improving city operations. However, very little of that highly-refined data has been made available to the public or City Council via the open data portal.

For instance, the Mayor's Office has gathered and mapped all of the data on sewer backups, a big issue in Queens in particular; this map was cited early this year in a hearing on funding for the Mayor's Office of Data Analytics, but it's not in the open data portal.

Additionally, AgencyStat data such as HousingStat, NYCHA's performance management process (which includes real-time updates of broken door locks and broken elevators in public housing) is not available to either the public or council. This data is some of the most highly sought after, and the Open Data Law cannot succeed without data like this being published.

Lastly, the city should publish its Vendor Information Exchange System (VENDEX) database, which the city uses to evaluate potential vendors for city agencies, as well as the Doing Business Database, which lists the vendors and lobbyists that do business with the city. Both of these are available to the public, but are not on the open data portal, significantly hampering their usefulness. They are not accessible via APIs, and the VENDEX database is only accessible in person at the Mayor's Office of Contract Services.

Data Quality

New York City has cultivated a robust civic technology community around open data well before the Open Data Law was even passed, thanks to the efforts of the NYC Economic Development Corporation, DOITT, and the Mayor's Office of Media and Entertainment.

BetaNYC, one of the largest such civic technology groups and member of the NYC Transparency Working Group, has held hundreds of weekly events where civic hackers meet to collaborate on projects built with NYC's open data. They have outlined a number of specific concerns, which we summarize below.

The open data portal has numerous data sets that are described as "updated monthly," but which instead are apparently abandoned. The data set of emergency alerts dispatched via NotifyNYC, for example, is updated every month, but the last alerts to be added to the data set are from June.[5]

There are smaller but still significant omissions and errors in many data sets, including extremely high-profile and high-value data. The NYPD's crash data has over 250,000 crashes with an "unspecified" cause, and another 55,000 with an incomplete or missing location. This is data for the di Blasio administration's Vision Zero effort, and either the NYPD is working with bad data or their publishing bad data to the portal.

Location data is by far the most frequent data quality issue on the open data portal. There are countless data sets which contain addresses or Borough/Block/Lot information but which are not "geocoded" – transcribed to standard latitude/longitude coordinates. Data which is geocoded can be easily "mashed up" and mapped with other data, increasing its usefulness.

Metadata, that is data about the data is often difficult to locate, and is just plain missing for many data sets. This lowers the usefulness of even high-quality data; users shouldn't have to guess what each column tracks, or what a certain abbreviation means.

[5] See: https://data.cityofnewyork.us/Public-Safety/Emergency-Notifications/5gke-dir7

Many members of the open data community feel that DOITT has slowed or stopped responding to user comments on the portal over the last 6-10 months. Most questions are not answered, so participation in the discussion has likewise slowed. The same is true for suggestions or requests for new data sets.

Measuring Usage and Success

The NYC Open Data portal publishes its site analytics on a page with statistics broken down into hourly and daily visits, and there is a running tally of total views of datasets, total rows (data points) in all data sets, and the number of times all charts have been embedded on other web sites.[6] This should be considered a best practice for all city agency web sites: using web site views to determine what information the public is most interested in.

The front page of the portal also has a compilation of "data stories," which are examples of how the Open Data has been developed into useful maps and applications. These data stories act as showcases for the platform, but they are not accompanied by anecdotes or statistics on how various members of the public are finding uses for these apps. We encourage DOITT and agencies to share more success stories about open data.

Perhaps it is too early to tell how the access to open data has impacted the public in NYC, as the portal is still new and many people still do not yet know of its existence. Agencies can help speed the portal's adoption by linking to the open data portal from their home pages and wherever they keep data on their pages. It is also unclear if the ease of access to data has been of use to various city or state agencies, especially agencies that did not always communicate with one another. At the moment, the city has not produced any reports on how the site is used and by whom.

Accountability

Currently, DOITT maintains an index of data sets which have been published on the open data portal. This data set of data sets provides the name, description, update frequency, and link to every data set published to the portal. In addition, it

[6] See https://nycopendata.socrata.com/analytics

has two columns, one for indicating if the data set is "Available" (i.e. published), and one to indicate if the data set is "Behind Schedule" per the Open Data Plan. It does not describe when the data set was supposed to be published, nor when it was actually published.

This is somewhat puzzling, as the Open Data Plan has a separate list of data sets which are scheduled to be published. Unlike the "index" data set, the "scheduled" data set displays the date by which each data set was supposed to be published. However, it does not display when the data set was actually published. Nor does it contain either the "Available" or the "Behind Schedule" Column, which is only available for data which has already been published.

Based on the fields in the "published data sets" data sets, it appears that DOITT tracks this information. We strongly encourage DOITT to merge these data sets and provide information about when each data set was published and when a data set is behind schedule.

Recommendations

Data federation

Big agencies like NYPD and DOE should be posting their own data on their own portals, and DOITT should federate that data into the NYC Open Data portal. Socrata has built-in federation capability, so it can share data across two Socrata-powered portals in minutes. CKAN also has this functionality, this is how the Federal data portal, data.gov, operates. called "harvesting" — the European Union's portal, publicdata.eu, uses this.

Continue Improving Search

From its launch in November 2011 until last week's update, the search capability on the open data portal was essentially useless. Now, it's adequate, but there are still improvements to be made. The biggest problem with the portal now is that there's no way to search just the title or description of a data set.

Link Maps to the Portal

New York City's agencies have created numerous maps, for everything from parks to rat sightings to water fountains. While these maps are useful, the real value lies in the data powering these maps. For every interactive map on an agency's web site, there should be a link to download the underlying data set from the portal.

Prioritize Data Sets with FOIL

Prioritize agencies' high-value data sets for publication first. Examine FOIL logs and look for large numbers of requests for information from a handful of data sets. Publish those data sets on the open data portal.

Create a Public Feedback Loop

When users of the open data portal leave comments or pose questions, they should get a prompt response, even if the response is something as simple as "we're looking into it." Currently, users hear nothing about their feedback for months at a time.

Stop Selling Data

The Open Data Law requires agencies to publish their data for reuse without restriction in an open format. However, the Rent Guidelines Board is still selling Housing Rents Markets and Trends online. They have data from 2005, 2008, 2011, 2012, 2013, and 2014 for sale. PLUTO and ACRIS were formerly sold on the City Store, but are now provided in open formats on the portal. The rest of the City Store data should follow the same process.