Quarterly Report: Open Data Program

To: Mayor Edwin M. Lee

Cc: Kate Howard, Mayor's Budget Director, Naomi M. Kelly, City Administrator, Ben Rosenfield, Controller, Miguel Gamino, Chief Information Officer, Nancy Alfaro, 311 Customer Service Center Director, Kenneth Bukowski, Deputy City Administrator, and Jennifer Johnston, Deputy City Administrator

From: Joy Bonaguro, Chief Data Officer

Date: May 1, 2015

Executive Summary

As required by the City's Open Data policy, this memo summarizes our quarterly progress. Below we summarize highlights from the full report and our workplan.

1. Inventory and Publication Plan Progress. Our Open Data Policy calls for the creation of dataset inventories by department. Our deadline for completing the dataset inventories was March 31, 2015. Of 52 departments:
- 36 (69%) have completed the inventory
- 7 (13%) are working on Phase 2 of the inventory
- 2 (4%) have completed Phase 1 but not started on Phase 2
- 7 (13%) have yet to complete Phase 1 of the inventory

The status of each department is in section 1 of the full report below.

2. Publication Progress. Our current publication metrics are mostly descriptive and not tied to specific publishing goals. As we complete the inventory, we will develop specific goals. The full section notes metrics that we plan to use as key performance indicators as well as metrics we plan to develop over time.
- 419 original datasets (Note: this number is inflated and will decrease as part of our data reset strategy.)
- 41% of datasets are "natively" hosted (our goal is 100% when appropriate)

3. Users and Uses of DataSF. In the prior quarter we deployed Google Analytics which will allow for much more robust analysis of our web traffic. With only two quarters of data, we cannot track performance over time, however, our initial data suggests that we are meeting our target audience:
- 40% of our user sessions originated from the City of San Francisco
- 41% of our our users visited our website for more than 1 session

Workplan Progress. As part of our strategic plan, we created a timeline of quarterly milestones. You can view the full workplan timeline and tracker online. Highlights include:
- In partnership with SFPD, relaunched the crime dataset as a single dataset going back to 2003
- Completed dataset inventory for majority of departments
- Developed department prioritization guidance and delegated on rolling basis
- Rolled out new licensing across portal
- Conducted additional Data Academy trainings - all fully booked
- Launched phase 1 of new website
- Launched dataset reset strategy - in progress
- In partnership with the Controller's office and a ten departments, developed strategic plan to streamline internal confidential data sharing

We divided the full report into three sections. These sections include our initial metrics and these will evolve over time. A "*" indicates measures that we are considering as preliminary key performance metrics.

1. Inventory and Publication Plan Progress

Under the City's Open Data Policy, departments must create an inventory (or catalog) of their datasets as well as a timeline for publishing the datasets. The Office of the Chief Data Officer is supporting departments in this effort. Once this process is complete, we will no longer track this set of metrics.

1.1 Percent of departments that have completed the dataset inventory: 69%*

March 31, 2015 was the target date for all departments to complete the dataset inventory. As of March 31, 2015, 69% (or 36 out of 52) of departments have completed the inventory. We will accept department submissions on a monthly basis. The dataset inventory will be published in Q4.

1.2 Departments by stage in inventory and publication plan process*

To simplify the inventory process, we broke the dataset inventory into two phases. Phase 1 consists of identifying information sources, while phase 2 is the completion of the dataset inventory, which includes a detailed list of the datasets in the department.The chart below shows our overall progress in completing the dataset inventory. Data is as of March 31, 2015.

Departments who completed the inventory began work on publishing plans in Q3 with the goal of launching publishing plans in Summer 2015. Starting next quarter, we will track the status of publication plan creation.

Phase 1: Number of Departments by Status

Phase 2: Number of Departments by Status
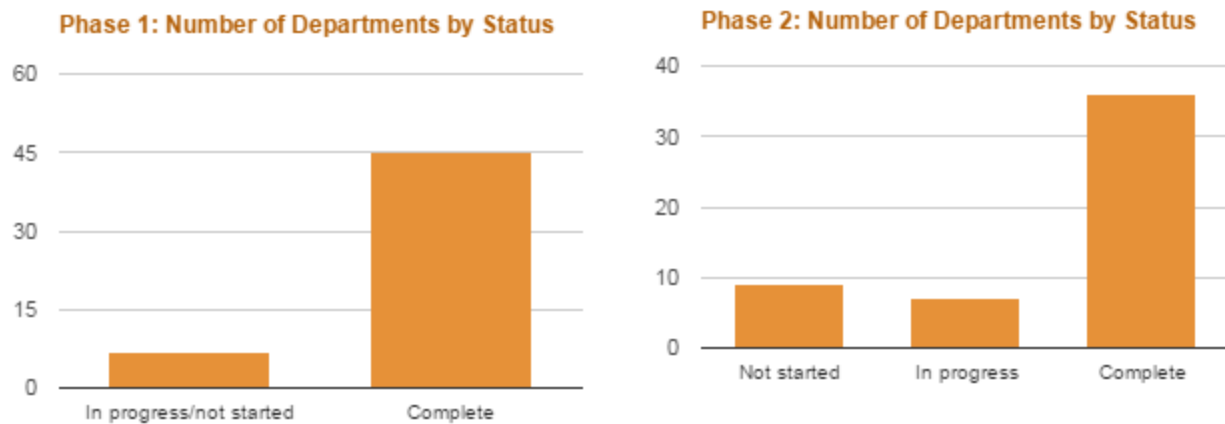
Department status by stage in inventory*

The table below shows the status of individual departments in the inventory phases. Data is as of March 31, 2015, which was the deadline for all departments to complete their inventory. Going forward, we will accept department submissions on a rolling basis.

Department	Inventory Phase 1	Inventory Phase 2
311	Complete	Complete
Academy of Sciences	In progress/not started	Not started
Adult Probation	Complete	Complete
Airport	Complete	Complete
Arts Commission	Complete	Complete
Asian Art Museum	Complete	Complete
Assessor-Recorder	Complete	In progress
Board of Appeals	Complete	Complete
Board of Supervisors	Complete	Complete
Building Inspection	Complete	Not started
Child Support Services	Complete	Complete
Children and Families Commission	Complete	Complete
Children, Youth & Their Families	Complete	Complete
City Attorney	In progress/not started	Not started
Civil Service Commission	Complete	Not started
Community Investment and Infrastructure	Complete	Complete
Controller	Complete	Complete
District Attorney	Complete	Complete
Economic & Workforce Development	Complete	Complete
Elections	Complete	Complete
Emergency Management	Complete	Complete
Environment	Complete	Complete
Ethics Commission	Complete	Complete
Film Commission	In progress/not started	In progress
Fine Arts Museums	In progress/not started	Not started
Fire Department	Complete	Complete
GSA – City Administrator's Office	Complete	In progress
GSA – Public Works	Complete	Complete
GSA – Technology	Complete	Complete
Health Service System	Complete	In progress
Human Resources	Complete	Complete
Human Rights Commission	Complete	Complete

Human Services Agency	Complete	In progress
Juvenile Probation	Complete	Complete
MOHCD	Complete	Complete
Municipal Transportation Agency	Complete	Complete
Office of Citizen Complaints	Complete	Not started
Office of Early Care and Education	In progress/not started	Not started
Planning	Complete	Complete
Police Department	Complete	Complete
Port	Complete	Complete
Public Defender	In progress/not started	Not started
Public Health	Complete	In progress
Public Library	Complete	Complete
Public Utilities Commission	Complete	Complete
Recreation and Parks	Complete	Complete
Rent Arbitration Board	Complete	Complete
Retirement System	Complete	In progress
Sheriff	In progress/not started	Not started
Status of Women	Complete	Complete
Treasurer-Tax Collector	Complete	Complete
War Memorial	Complete	Complete

2. Publication Performance

The metrics below are largely based on availability and are mostly descriptive. We are working to create additional metrics and the supporting data that are tied to programmatic goal per the table below.

Program Goal	Metrics Under Consideration for Public Dashboard
Increase timeliness of data publication	● Percent of datasets updated on time ● Percent of eligible datasets without a publishing lag
Increase the number of datasets published	For both the program and by department: ● Percent of datasets published against target and/or per publication plans ● Percent of inventoried datasets that are published ● Percent of inventoried datasets published by priority level ● Percent of inventoried datasets published by dataset classification

Expected revisions per dataset reset strategy. In the meantime, we continue to monitor overall publishing trends. However, we expect these numbers to continue to revise as part of our dataset "reset" strategy that we launched this past quarter to clean or remove low quality datasets, consolidate datasets, and overall standardize what we consider to be a dataset. If the dataset reset is not complete by next quarterly report, we will add a status tracker to this report.

As of March 31, 2015, the City had 419 datasets published. This is down from 434 last quarter per our dataset reset strategy. We expect this to greatly decrease in the next 2 quarters per the reset and once we feel we have a revised baseline, will track against it per our proposed measures.
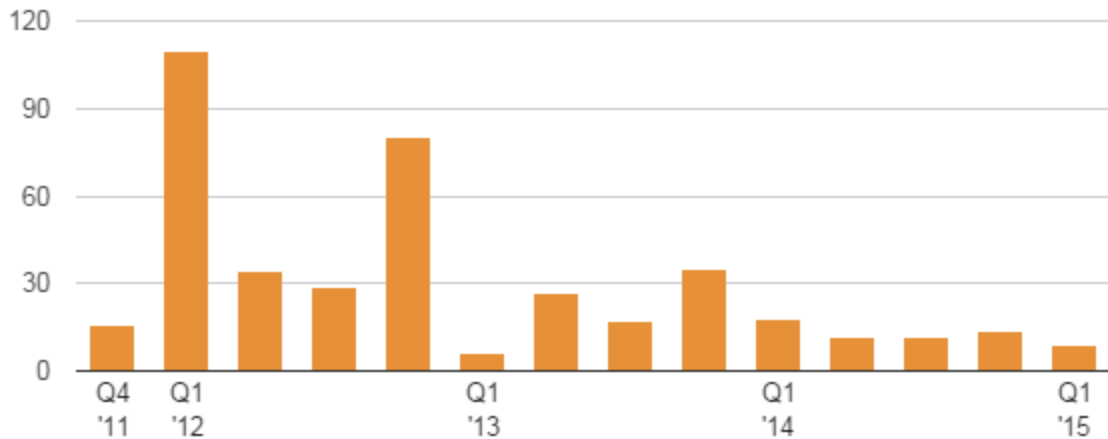
2.1 Datasets published by calendar year

The charts below shows the number of datasets published by calendar year. The year of publication indicates when the data was first published. As we review our data and in many cases work to republish or retire datasets, the historic numbers will change as the data will be based on only published datasets. The numbers below only include original datasets - not datasets or views that are built on original datasets as this distorts our actual publishing totals. More than half of our datasets were published in 2012.
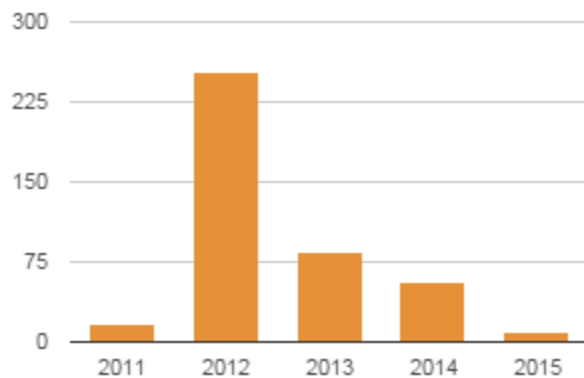
2.2. Datasets published by quarter

The quarterly numbers show that our rate of publication continues to slow. We expect this to continue until

the inventory and publication plans are in place. In addition, as we "clean" up our datasets, the historic numbers will revise downwards as some of the datasets are duplicative or should not be counted as a single dataset.

2.3 Number of datasets published by category

The chart below shows the number of datasets published by category. We track this in order to see what are the most popular categories. The chart may also indicate if our categories are too broad, which may make it more difficult to navigate our site, which uses category as a key navigation method. City Management and Ethics continues to be our largest category due to the high number of datasets published by the Ethics Commission.


2.4 Percent of published datasets that are natively hosted: 41%*

Datasets can be natively hosted on our open data platform or externally linked. When natively hosted, datasets can leverage the API, visualizations, user derived summaries and visuals, OData connections and more. Our goal is for all datasets to be natively hosted unless the external hosting provides comparable and additional features.

2.5 Datasets published by format

The chart below shows the number of datasets published by format. Both external links and files are non-native ways of publishing the data. Datasets using this format are candidates to migrate to a native hosting method.
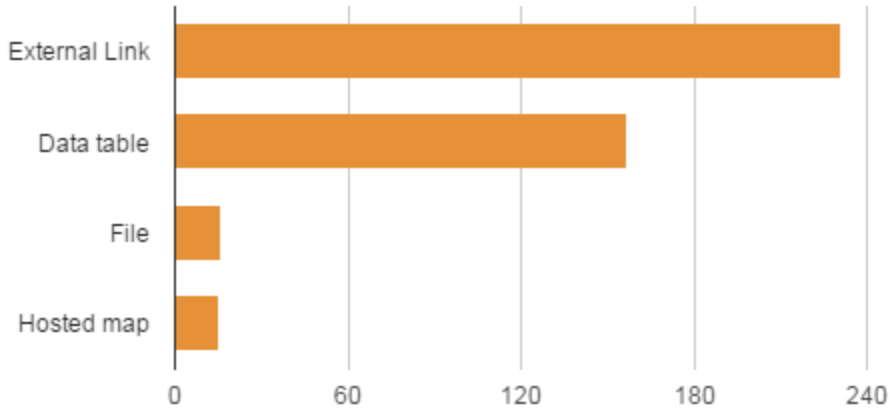
2.6 Top publishers of non-native datasets

Departments with large numbers of externally linked datasets are candidates for native hosting. We will work with departments to re-publish or consolidate external datasets as part of our data "reset" strategy.
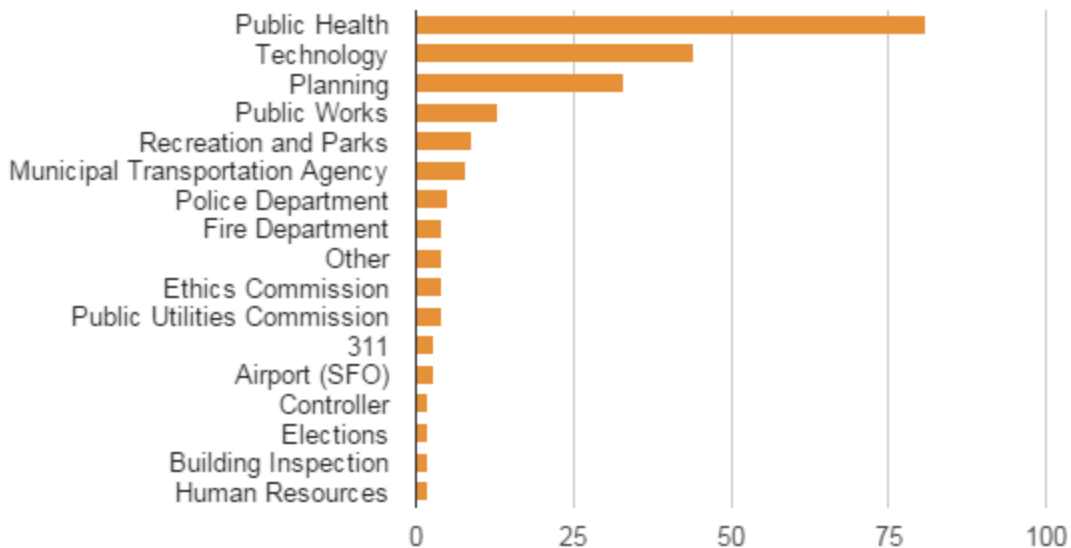
3. Users and Use of DataSF

3.1 User Analytics: January - March 31, 2015

For this quarter and last, we were able to obtain additional detail about our users with the addition of google analytics. This allows us to have a richer sense of who is using our website. While we are just starting to collect this data, our 2nd full quarter of data indicates that we are attracting our target local audience (40% of sessions come from San Francisco) and 41% of our users are returning users. As we collect more of this data, we will create a consistent set of measures to track, including tracking data over time or in response to specific campaigns. We will also identify differences between google analytics and our Socrata analytics and reconcile if possible.

In addition, these metrics represent an undercount due to a technology misconfiguration by our vendor Socrata, with no data collected January 1-14.

	Metric	What it means
39,851	User sessions (visits)	Number of site visits where a user interacted with our website within a given time frame (30 minutes)
24,259	Users	Number of unique visitors to our website - a visitor can have multiple visits
41%	Returning users	% of our our users who visited our website for more than 1 session
40%	Sessions in San Francisco	% of our user sessions that originated somewhere in the City of San Francisco
3:33	Average session duration	Average number of minutes spent on the site
13.46	Pages per session	Average number of pages per session
.52%	Bounce rate	Percent of visitors that come to the site with no interaction

3.2 Top 10 Datasets

We saw an increase in the use of our crime dataset, which we relaunched as a single dataset with data going back to 2003, which greatly improved the usability of the dataset.

Name	
SFPD Incidents - from 1 January 2003	4,628
Map: Crime Incidents - from 1 Jan 2003	4,285
Registered Business Locations - San Francisco	3,547
HSA 90 day emergency shelter waitlist	3,054
City Lots (Zipped Shapefile Format)	2,199
Case Data from San Francisco 311 (SF311)	1,648
Street Names	1,560
Film Locations in San Francisco	1,515
Development Pipeline Q3 2014	1,259
Building Footprints (Zipped Shapefile Format)	1,241

3.3 Top Embedders

Embedders are users who create visuals on DataSF and then embed or place those visuals in their own website. This indicates that DataSF is providing a useful service. The numbers below indicate page views not the number of embedders and who are they. We are working on our analytics to create a more useful version of this report.

Name	Embeds
http://www.sfethics.org	4,632
http://www.sf311.org	3,114
http://sf-planning.org	967
https://www.google.com	868
http://sf311.org	541
http://www.sf-planning.org	212
http://spotcrime.com	168
http://vertigoliving.com	128
http://www.socrata.com	122
http://www.sfmta.com	95

3.4 Creation of Derived Views Over Time

Derived views indicate engagement and use of the website. As we re-deploy our dataset interface and increase the rate of publishing, we would want to see this increase.