

Open Data in San Francisco:

Institutionalizing an Initiative



City and County of San Francisco
Mayor Edwin M. Lee

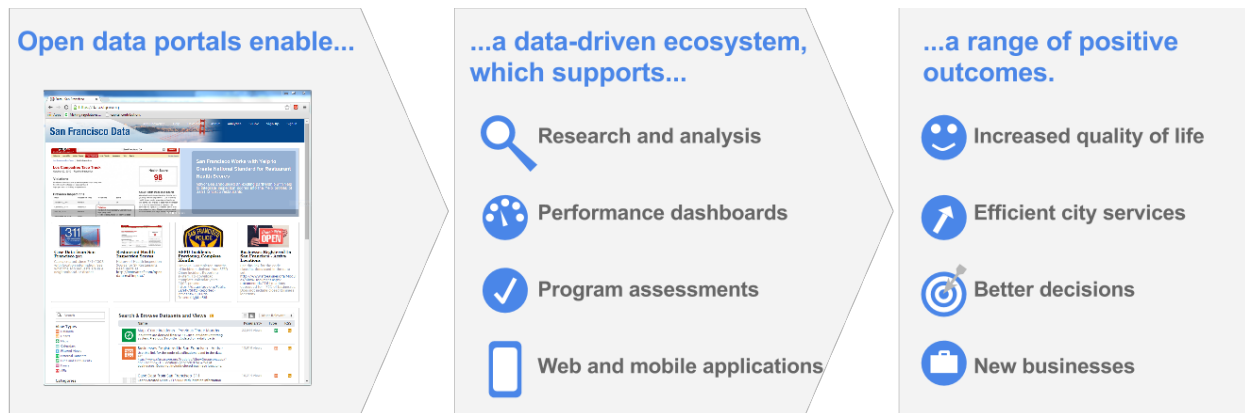
Joy Bonaguro, Chief Data Officer
July 14, 2014

Table of Contents

Executive Summary	3
1. The Case for Open Data	6
2. The State of Open Data in San Francisco	7
3. Engagement Methods in Developing this Plan	10
4. Mission and Vision	11
5. Goals and Strategies for Year 1	
Overview of Approach and Goals	12
Goal 1: Increase number and timeliness of datasets on DataSF	16
Goal 2: Improve the usability of DataSF	21
Goal 3: Improve the usability, quality, and consistency of our data	27
Goal 4: Enable use of confidential data, while appropriately protecting it	29
Goal 5: Support increased use of data in decision-making	31
Goal 6: Identify and foster innovations in open data and data use	35
6. Prioritization, Resource, and Risk Analysis	37
7. Conclusion	41
Appendices	
Appendix A. Engagement Methods	42
Appendix B. Cross walk between plan and Open Data Policy	43
Appendix C. Additional Details on Site Analytics	47
Appendix D. High-Level Timeline	48
Appendix E. Acknowledgements	48

Executive Summary

San Francisco has been a leader in open data. As one of the first cities with an open data policy, we helped fuel a movement that has spread across the country and the world. Open data can serve as a platform to 1) change how we use, share and consume government data - externally and internally; 2) transform data into services, and 3) foster continuous improvement in decision-making and the business of government.



Need to Evolve. But this plan demonstrates the need to evolve and mature our approach. Not only do we need to reinvigorate our program and release more data, we need to evolve our work to support the use of data in decision-making. To transform our initiative into a program, our strategic plan is designed to build the elements of an institutional approach to open data and data use more generally. The goals and strategies for year 1 lay out a framework for how we can grow, mature, and sustain our program to align our activities and resources with the expectations of the open data policy. In achieving these goals and strategies, we can fulfill our mission of enabling use of the City's data thereby fostering an ecosystem of data-enabled management, services, and decisions.

Timing and Resources. Our strategic plan is ambitious and reflects a vision of what we hope to accomplish over time. We do not expect to be able to deliver on all aspects of our strategic vision in year 1. However, by fully articulating our vision, we are better able to prioritize and allocate our resources and identify where we need key partners to help execute on our goals. Moreover, this plan recognizes that each of our goals are multi-year goals and that a great deal of work is already happening throughout the City. This plan helps us stitch together an overarching vision of how these efforts align, where the role of open data fits in, and how we can move forward to enable use of data.

While we expect our strategic goals to change over time, we believe the goals in this plan will be in place for the next three years. The Office of the Chief Data Officer (OCDO) will be focused on Goals 1, 2 and 6 in year 1, in conjunction with key partners and the departments themselves. In year 1, OCDO work on goals 3, 4, and 5 is focused on leveraging key partnerships or pursuing strategies that will inform future work. Section 6 and Appendix D include more details.

Goals

Supporting Strategies

Goal 1. Increase number and timeliness of datasets on DataSF

Why this matters. To enable the use of data we must first make it available. We need to ensure that we are publishing the City's data when allowed and in a timely way. We will:

1. Establish the role of data coordinators and support development of data catalogs—this will provide the basis of data governance and allow us to understand the scope of the City's data.*
2. Develop methods to inform the prioritization of datasets for publication—this will allow us to stagger publication based on resource availability.
3. Develop metrics to track and measure progress in publishing open data—this will provide the basic reporting for our data publication plans.
4. Develop our program to automate publication of data—this will increase efficiency and decrease department effort in publishing datasets.*
5. Develop an outreach and support program for data coordinators and data publishers—this will help departments be successful in publishing data.
6. Establish methods to ensure SF licensing and publication of data for new information systems—this will stem future open data costs by building open data into new system requirements.*

Goal 2. Improve the usability of DataSF

Why this matters. To ensure that our open data is readily accessible and used, we need to make sure that our data website and the means of accessing the data support the needs of users. We will:

1. Better leverage existing services and features from our data portal vendor, Socrata—this will help optimize our investment in our vendor.*
2. Partner closely with Socrata to inform the development of the portal—this will help ensure that the data platform evolves to meet our needs.
3. Redesign our web presence and supporting processes and materials to better meet the needs of our users—this will increase the impact of open data by easing access to more users.

Goal 3. Improve the usability, quality and consistency of our data

Why this matters. While Goals 1 and 2 help provide access to the City's data, the ultimate value of the data depends on its usability, quality, and consistency. We will:

1. Establish metadata standards for published data—this is key to increasing understanding and usability of our data.*
2. Establish mechanisms to elicit and track feedback and learnings from data users—this will help us flag data quality problems.
3. Explore the creation of data quality processes and measures—this will help inform how to support improved data quality over time.

Goal 4. Enable use of confidential data, while appropriately protecting it

Why this matters. While the City needs to appropriately protect confidential data, we also need to enable better access to and use of this data for cross-department data sharing. We will:

1. Create a data classification and sharing standard—this will help improve efficiency and consistency in sharing and protecting data.*
2. Create a process for accessing your individual data—this will help support data quality and privacy.*

Goal 5. Support increased use of data in decision-making

Why this matters. Once data is available, we need to use it. We need to match the availability of data with the capacity to use data, both in terms of people and technology. We will:

1. Establish a training curriculum to support increased use of data in decision-making—this will increase the capacity of City staff to use, analyze, and manage with data.*

2. Help establish department stat programs based on department readiness; codify lessons learned and materials for broader use—this will help increase effectiveness in using and leveraging “stat” programs.*
3. Continue to develop our portfolio of transparency tools and websites—this will help leverage open data to inform broader decision-making.*

Goal 6. Identify and foster innovations in open data and data use

Why this matters. The pace of change in the open data, analytics, and visualization spaces is breathtaking. We need to identify and nurture innovation in order to ensure that the City benefits. We will:

1. Develop and maintain a communications and engagement strategy—this will help ensure that we stay in touch with evolving stakeholder needs.
2. Conduct ongoing reviews of best practices and the changing technology landscape—this will help us stay ahead of the innovation curve.
3. Identify and enable targeted data-centric initiatives—this will help us uncover and foster new uses of data, internally and externally.*
4. Establish a data licensing framework and standard—this will help ensure that our data users can use our data in any way they see fit.*

*Indicates need for resources or activities beyond the OCDO (e.g. key partnerships, department effort, volunteers etc).

1. The Case for Open Data

The opening of government data is another stage in the evolution of how government interacts with its constituents - once again enabled by changes in technology. In the late 1990's and early 2000's, governments around the world pushed for "e-government". The emergence of the Internet enabled e-government by allowing governments to migrate services online. The term e-gov may now seem quaint - few agencies discuss their "e-gov" initiatives - they are expected to offer services online.

Similar to e-gov, distributing vast amounts of open data is now feasible and desirable given changes in technology. Changes in data extraction and distribution technologies ease the path to publication. The proliferation of mobile enabled devices drives demand for data-driven services anytime, anywhere. And new tools in data analysis and visualisation allow us to better explore, understand and communicate large datasets.

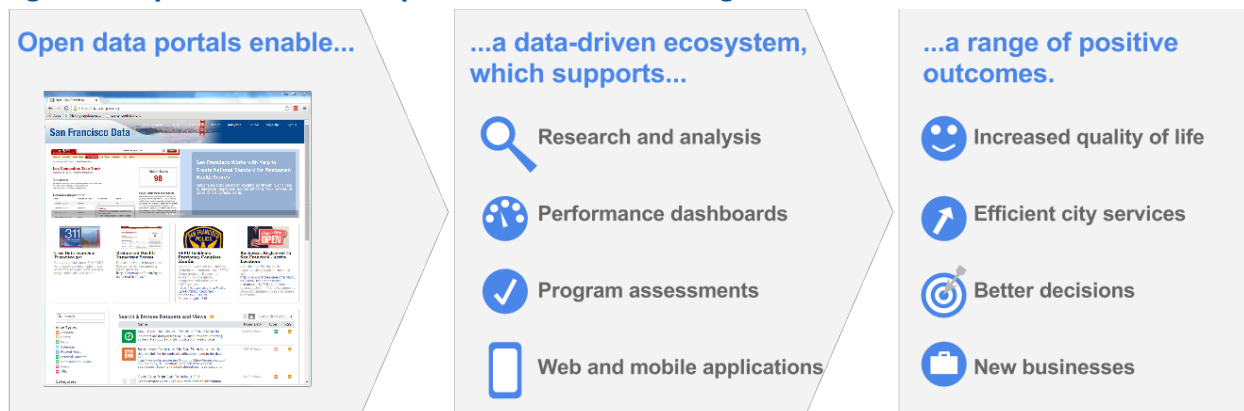
Much of the celebrated benefit of open data has been focused on consumer-based web and mobile applications. And while these have brought broad value (real-time access to transit data, for example), they are not the only source of value. Open data portals enable a wide range of data-driven work by creating a shared platform where users can readily access and use government data, whether developing applications, conducting research, or acting as a concerned citizen. It even eases access to data within the government entity itself.

Ultimately, this data-driven ecosystem should support a range of positive outcomes - from increased quality of life, more efficient city services, better decisions, as well as the business models it has already fostered. Figure 1 shows how open data is just an input to broader change.

What is Open Data?

Open data has a variety of definitions. [Open Definition](#) summarizes it as "A piece of data or content is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike." The U.S. Federal Government's Open Data Policy states that "open data refers to publicly available data structured in a way that enables the data to be fully discoverable and usable by end users." In practice, the open data movement has focused on the public release of government managed data.

Figure 1. Open Data as an Input into Broader Change



2. The State of Open Data in San Francisco

San Francisco has been a leader in the open data movement. We were one of the first cities in the country to establish an open data policy in 2009, which we then codified in the City's administrative code in 2010.¹ As of March, 2014 we were ranked as #1 in the U.S. Open Data Census that assessed 41 localities on the openness of 17 datasets.²

San Francisco hosts its open data on DataSF - our open data portal, data.sfgov.org. DataSF allows end users to find, visualize and use our data, whether developing novel applications or combining the data across multiple agencies to support new services. Figure 2 describes a handful of the many applications built using DataSF. Figure 3 describes how the San Francisco Ethics Commission uses DataSF to increase transparency by summarizing and creating visualizations related to ethics data and reports.

By being early to the open data table, San Francisco demonstrated the value to localities across the country and world. Beyond simply opening up our data, we've unleashed several innovations in the open data movement. For example, open data standards create a multiplier benefit to open data by codifying how data is structured, which allows applications created in one locality to be easily used by another locality or readily integrated into private applications. Figure 4 describes two data standards pioneered by San Francisco in partnership with key stakeholders.

While San Francisco has had many successes in open data - we need to do more. Recent changes to the open data legislation established the role of Chief Data Officer and Data Coordinators in each department. These changes reflect a need to codify and mature our approach to open data. These new roles combined with this plan create the institutional framework to grow, mature and sustain our open data program.



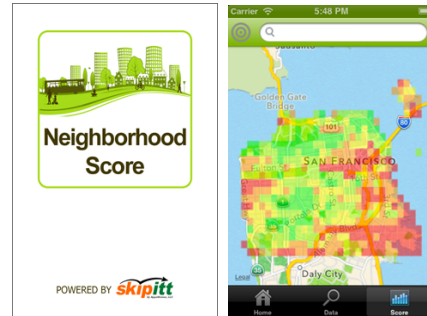
¹ Our open data code is available at <http://www.amlegal.com/library/ca/sfrancisco.shtml> under Administrative Code, Chapter 22D: Open Data Policy.

² Article describing the results of the open data census: <http://www.governing.com/news/headlines/san-francisco-is-the-best-city-for-open-aata.html>

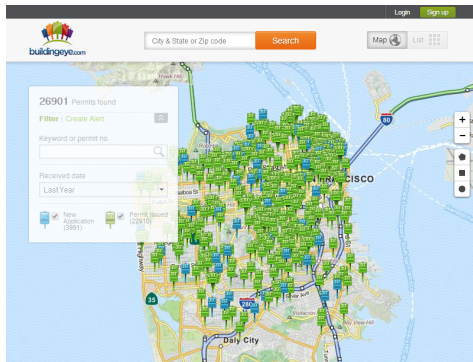
Figure 2. Examples of Applications Built Using DataSF

Below are just a handful of the applications built using DataSF. Our Applications Showcase features many more, apps.sfgov.org/showcase.

Neighborhood Score. Neighborhood Score is a mobile application designed to provide an overall health and sustainability score, block-by-block for neighborhoods in San Francisco. The app combines a variety of health-related data, including measures on mental health, safety, traffic and physical well being, to generate a neighborhood score. This app can be used to assess livability, identify success and failures in various communities, and advocate for a healthier city on a street-by-street basis – empowering residents and elected officials.



BuildingEye. Buildingeye.com makes building and planning information easier to find and understand by mapping what's happening in the city. Users can set up alerts to notify them about new construction in a geography of their choice.



SF Rec and Park App. The SF Rec and Park App provides users with a way to find locations (including Parks, Playgrounds, Dog Parks, Museums, Rec Centers, Picnic Tables, Gardens, Restrooms, News, Events, and other points of interest) based on their current location. Each location includes a description and pictures and can be viewed on a GPS enabled Mobile Map.

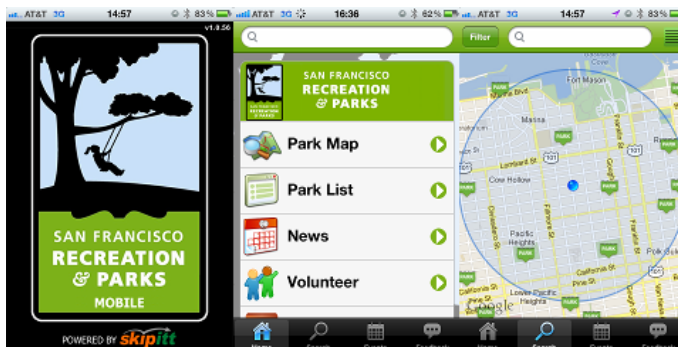


Figure 3. Use of DataSF to Increase Transparency in Ethics

DataSF is hosted on Socrata's proprietary platform, which allows end users to summarize and create charts of data and then embed those visualizations into existing websites. The Ethics Commission was in need of an easy and affordable way to present the campaign and ethics data that it collects. By publishing the data on DataSF and then creating summaries and charts of the data, the Ethics Commission was able to convert lengthy reports into easy to consume charts and information. In addition, since the charts update automatically when the data updates, the Ethics Commission has negligible overhead for updating the dashboards.

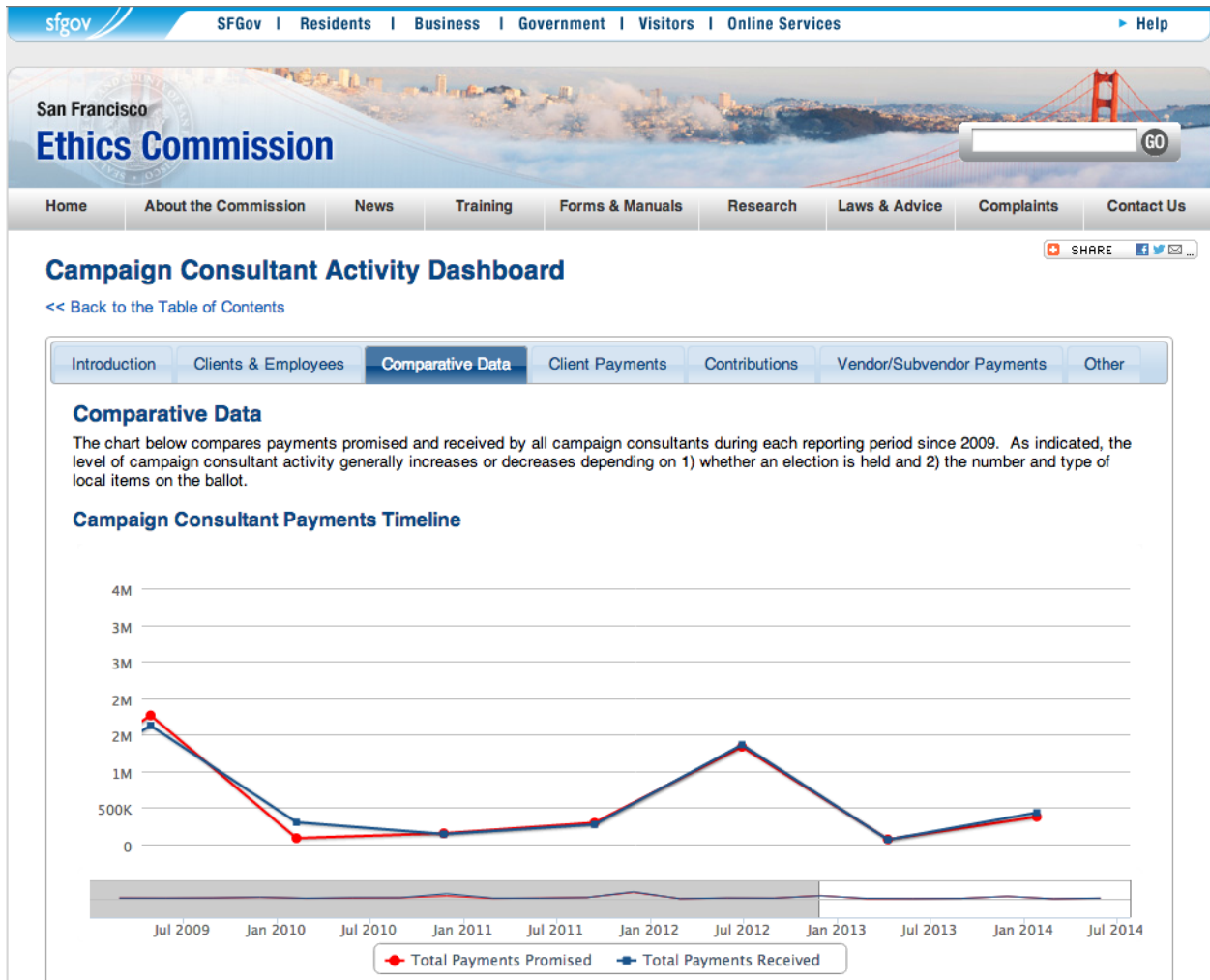


Figure 4. Open Data Standards

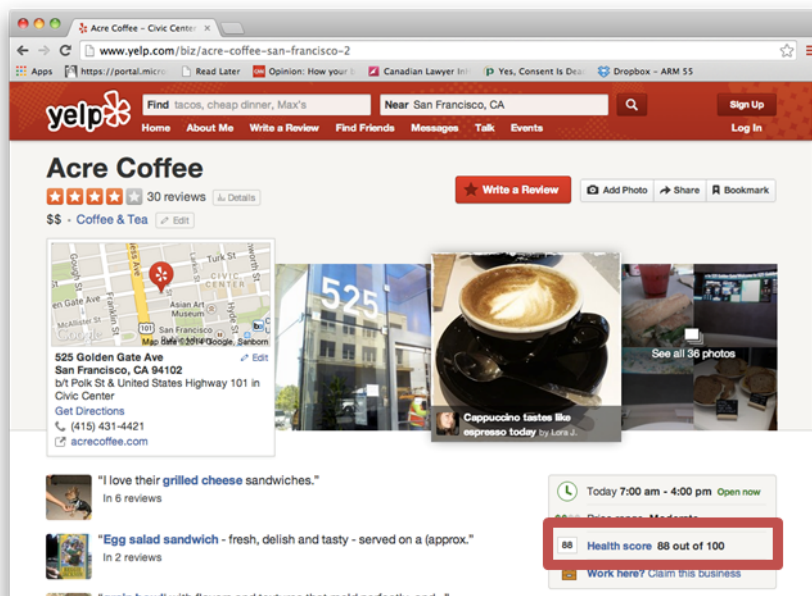


House Facts Standard. Developed in conjunction with [Code for America](#) and the Department of Public Health, the Standard will make the health and safety history for every residence in the City accessible to citizens in a “computer friendly” format. The House Facts Standard can improve public health by improving housing conditions and drive economic development with the generation of apps and other data-based tools.

Establishing a standard format for reporting data makes it easily replicable in other locations; six other cities including Las Vegas, NV and Kansas City, MO have signed on to test the standard, placing the potential civic impact of this project at the national scale.

LIVES Standard. Public availability of restaurant health scores has increased consumer confidence and led to improvements in health and safety practices of restaurants around the country. Working locally and nationally, Yelp, San Francisco, NYC and Philly worked together to create a national open data standard and partnered with Code For America and Yelp on a campaign to enroll more cities which already collect and publish this data.

Yelp’s engineering team, SF, NYC, and Philly technical staff designed the [Local Inspector Value-entry Specification \(LIVES\)](#), which enables local municipalities to accurately upload restaurant health inspection scores to Yelp’s database.



3. Engagement Methods in Developing this Plan

To maximize the chance that our plan reflects the voices of many stakeholders, we used a variety of engagement methods. However, this was at best a first pass, and we expect both our assessment and plan to be a living document that improves as both our understanding of the challenges and our ability to effectively engage our stakeholders increases. To ensure that we maintain a robust engagement strategy, we capture it under Goal 6 of our plan.

In addition to our engagement strategy, we reviewed not only the literature but existing open data plans and practices from NYC, Chicago, Philadelphia, Great Britain and many more.

Appendix A. summarizes our engagement strategies by stakeholder group.

4. Mission and Vision

Mission

Our mission is to enable use of the City's data to support a broad range of outcomes - from increasing government transparency and efficiency to unlocking new realms of economic value.

Vision Statement

The City's data is understood, documented, and of high quality. The data is published so that it is usable, timely, and accessible, which supports broad and unanticipated uses of our data.

This vision enables a range of outcomes for many stakeholders - from citizens, to nonprofits and community groups, to businesses and the City itself. The table below lists just a handful of outcomes that can be realized through open data.

Area	Potential outcomes
Transparency and Accountability.	Information about city activities, services and management is readily available to the public, which supports government accountability. This in turn could reduce accountability-centric legislation and attending administrative costs.
City Performance and Operations Management.	Using data that is readily available and of high quality: <ol style="list-style-type: none">1. City departments actively manage and improve their services and projects2. City departments effectively and efficiently monitor facilities, equipment, and property
Resident Engagement.	Using city data, residents, nonprofits, businesses, and the government collaboratively plan future city initiatives.
Quality of Life.	Applications put city information at our citizens fingertips, easing access to and increasing predictability of city services, while enhancing overall quality of life.
Program Planning and Evaluation.	City departments and non-profits use city data to conduct needs assessments, advocate for resources, evaluate and continuously improve programs, and demonstrate success.
The Data Economy.	City data provides the seed corn for new start-ups while existing businesses use city data to improve their business or identify new market opportunities.
Research Collaboration.	Researchers collaborate with city departments and use city data to develop key insights into government efficiency, service delivery, allocation of resources, and other areas.

5. Goals and Strategies for Year 1

Overview of Approach and Goals

Given the sheer range of stakeholders and latent value in open data, we will focus on goals that benefit the broadest set of stakeholders. By doing this, we pursue our mission as an enabler of data use. The work plan below also reflects a set of principles we will use to inform our strategies in year one:

- Say no to perfection - something is better than nothing
- Create infrastructure for future growth, while solving immediate problems and pain points
- Fail early and often - but learn from the experience
- Use long division - if a problem seems too big, break it into manageable bits
- Leverage existing tools where possible - plan for no big technology changes

Beyond our principles, an underlying goal in year one is to build the elements of an institutional approach to open data and data use more generally. After the Executive Order on open data, San Francisco released a large number of datasets and had good momentum. Eventually, the pace of publication slowed. In the last year or so, we realized the need to allocate dedicated resources to the program, and modified our code to create new roles (the Chief Data Officer and Data Coordinators). We are still in the midst of aligning our resources with the expectations of our policy. This plan helps identify what resources we need to be successful.

We will then use our resources to integrate open data into the institution. We will develop a shared understanding of roles and responsibilities as open data touches every department and involves a range of technology. We also need to create and then continuously improve operational support and processes (e.g. publishing and automating data, troubleshooting, handling data requests etc). Lastly, we need to establish the data governance frameworks and standards that will facilitate broad improvements in data sharing, metadata, and quality. Figure 5 provides a conceptual overview of the phases of open data - we find ourselves in Phase 4.

Figure 5. Phases of Open Data



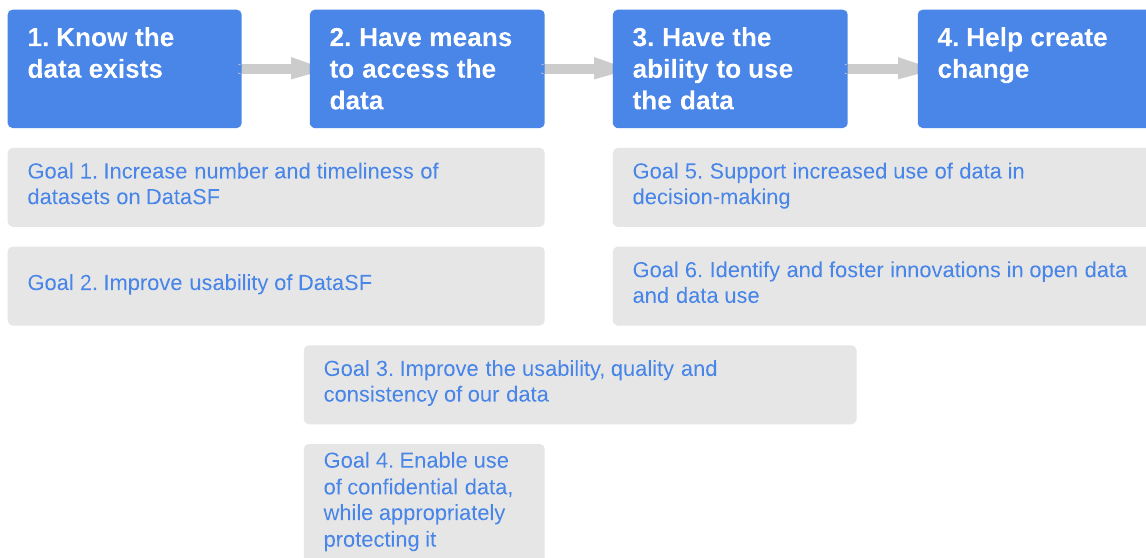
Summary of Year 1 Goals

To mature and integrate our open data program, we've identified six core goals for Year 1:

Goal 1.	Increase number and timeliness of datasets on DataSF	A precursor to our mission of enabling use of data is to make it available. In the near term we need to ensure that we are publishing or plan to publish the City's data when allowed. We should also publish the data at a frequency that matches the rate of data change. For example, datasets that change daily should be refreshed daily.
Goal 2.	Improve the usability of DataSF	Our data portal, DataSF, is a key part of our open data policy as it creates a single point of entry to our data. To ensure that our open data is readily accessible and used, we need to make sure that the website and the means of accessing the data support the needs of users.
Goal 3.	Improve the usability, quality and consistency of our data	While Goals 1 and 2 help provide access to the City's data, the ultimate value of the data depends on its usability, quality, and consistency. Usability helps us understand the data - what is it, how is it collected, when is it published - the basic documentation that supports use of the data. Quality speaks to how reliable and complete the data is - can we trust the conclusions or decisions we make based on the data? Consistency helps us combine data from different systems, by using consistent definitions across datasets, whether it's race, service categories, target populations, location etc.
Goal 4.	Enable use of confidential data, while appropriately protecting it	The City collects a tremendous volume of individually identifiable information, some of which is heavily regulated in the case of health or justice-related data. The City also has proprietary or confidential data (e.g. account level water consumption). While the City needs to appropriately protect this data, we also need to enable better access to and use of this data - whether it's cross-department data sharing or means to summarize and publish the data.
Goal 5.	Support increased use of data in decision-making	Once data is available, we need to use it. Open data already supports a broad range of external uses. But it can also better support internal use of data in decision-making - both by creating access to data and enabling new means of displaying and communicating data. We need to match the availability of data with the capacity to use data, both in terms of people and technology.
Goal 6.	Identify and foster innovations in open data and data use	The pace of change in the open data, analytics, and visualization spaces is breathtaking. We need to not only ensure we are aware of innovations, but we need to selectively identify and nurture innovation in order to ensure that the City and our stakeholders benefit from changes in technology and the experiences of others.

In structuring these goals, we hope to solve a number of challenges related to enabling use of data. The first challenge is knowing what data we have and then second, having a means to access that data. While open data's focus has been on external users of data, we've found that our internal staff also lacks information about the scope of the City's data and an efficient means to access that data.³ Once you have access to data, you then need the right set of skills and technology to use it effectively and the quality of the data must support the use - whether it's analysis or the creation of new services. If you can solve these challenges, you can enable use of data to create change. Figure 6 shows the alignment between our goals and our key challenges in using data.

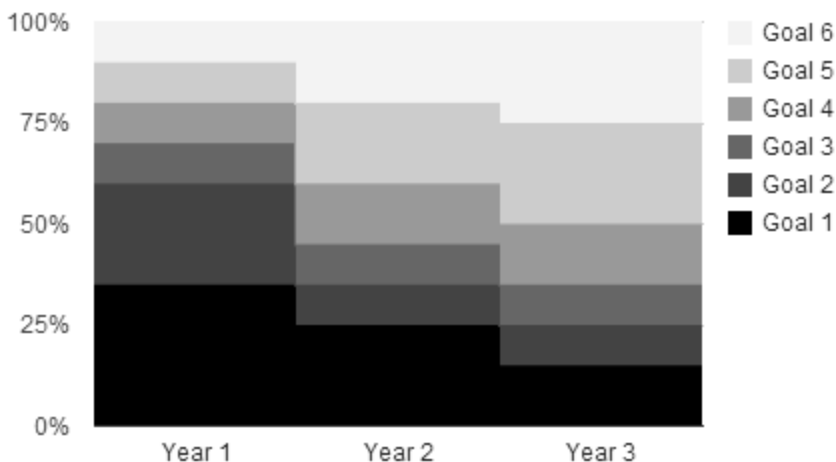
Figure 6. Alignment of Goals with Key Challenges in Data Use



While we expect these goals to change over time, we believe these core goals will be in place for the next three years. Our focus in Year 1 is on Goals 1 and 2 and we expect these to take up the bulk of time for Open Data staff. Our work in these areas will help us establish the fundamentals of a mature program. Goal 3 will be an ongoing need, however, much of the work activity will be in the departments themselves and we do not expect this to take up the bulk of our time. For Goal 4, we expect to conduct background work but then expect this goal to take up more time in years 2 and 3. Eventually, we expect the bulk of our work to be in Goals 5 and 6 but we expect this to happen in the context of key partnerships. Figure 7 provides a visual of how we expect the allocation of our effort to change over time. The allocation below is at best a good guess - it is only meant to convey the relative weight of our work.

³ For example, we found that knowledge of existing datasets was a common barrier to data use and that when accessing data outside of departments, many analysts relied on personal relationships or contacts.

Figure 7. Expected Allocation of Open Data Effort Over Time



In the rest of this section, we describe the current state for each goal and identify a set of strategies to support the goal. For strategies that support multiple goals, we placed it under what we feel is its primary goal.

Appendix B. provides a crosswalk that demonstrates how our goals and strategies meet the Open Data legislation.

In achieving these goals, we hope to partner closely with existing groups in the City, including but not limited to:

- The Office of the Controller - which does a great deal of complementary work, including its series of transparency websites, performance management, and the work of the office of the City Services Auditor.
- The Library - which houses a great deal of expertise in information management and standards.
- The Department of Technology - which provides expertise in technology, in particular extracting data from existing information systems.
- 311 - which has been managing DataSF and the supporting processes.

Goal 1.

Increase number and timeliness of datasets on DataSF

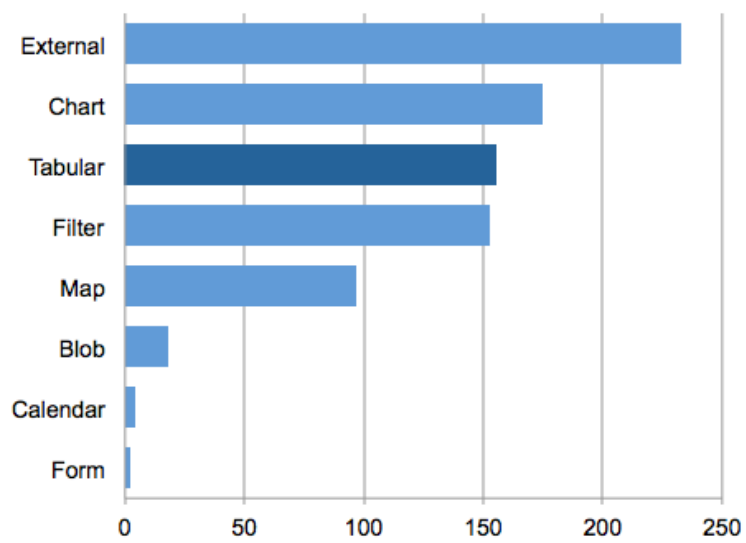
A precursor to our mission of enabling use of data is to make it available. In the near term we need to ensure that we are publishing or plan to publish the City's data when allowed. We should also publish the data at a frequency that matches the rate of data change. For example, datasets that change daily should be refreshed daily.

Current State

While the City had a strong initial push in releasing data, the pipeline has narrowed. An analysis of DataSF indicates that much of the data on the site (847 items in the catalog) is derivative of a smaller set of datasets. The data platform allows users to create views of data (using filters, maps, charts, etc). Each new view is counted as a new item in the portal. In addition, more than 200 datasets are pointers to data outside of the system. For example, an external dataset for Census data refers users to the US Census bureau website. Other external datasets point to actual datasets, including zipped .csv files. It may be the case that not all data will be hosted natively on DataSF, however, native hosting of the City's data should be a starting goal, with exceptions used only when needed. This allows us to fully realize the tools on DataSF.

When we control for either derived items or external pointers, DataSF is natively hosting approximately 150 datasets in a machine readable format, denoted as "Tabular" in Figure 8.

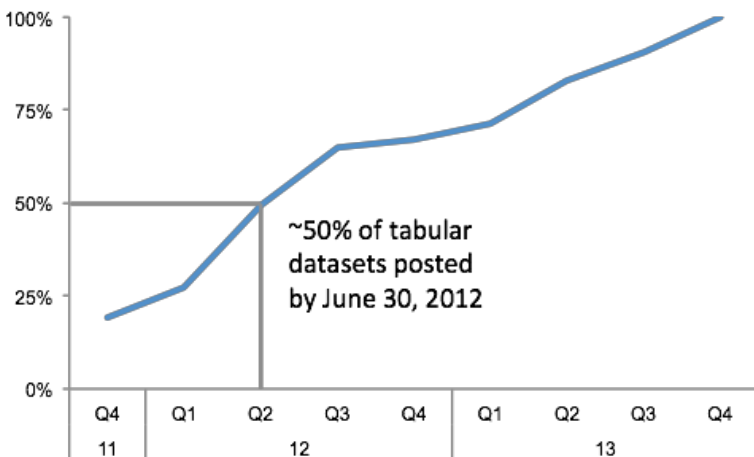
Figure 8. Datasets on DataSF by Publication Format.



Source: Analysis of Data Catalog data as of February 17, 2014. Data available at <https://data.sfgov.org/Other/Data-Catalog/h4ui-ubbu>.

Moreover, the pace of publication has slowed. If we analyze just the tabular datasets, 50% of them were published by summer of 2012. In the last six months of 2013, 25 datasets were published. See Figure 9.

Figure 9. Tabular Datasets on DataSF Posted by Quarter (cumulative percent).



Source: Analysis of Data Catalog data as of February 17, 2014. Data available at <https://data.sfgov.org/Other/Data-Catalog/h4ui-ubbu>.

Strategies for Goal 1

To increase the number of original datasets on DataSF, we will pursue the following strategies in year one.

Strategy 1.1. Establish the role of data coordinators and support development of data catalogs. The open data legislation calls for the role of data coordinators and data catalogs (essentially, an inventory of the data in the department). Before we can increase the number of datasets available, we need to first understand what data exists. The data coordinators and catalogs will serve as a key point of understanding the scope of the City's data. The data coordinators will:

- Inventory department data sets and establish a plan and timeline for publishing them
- Serve as a key point of accountability for timelines and questions about data sets
- Implement privacy, data licensing, metadata and other standards and practices
- Provide quarterly reports on progress in implementing the open data plan

In creating the data catalogs, we will try to capture information to support prioritization of data publication, including but not limited to existing output methods, public information requests, availability of historic data, and frequency of data changes.

We will also work to capture concerns in publishing data, either because it is confidential, raises security concerns, or is otherwise regulated. This will also help inform Strategy 4.1 as well as

identify any common misconceptions or misunderstandings related to what can be public or not.

Strategy 1.2. Develop methods to inform the prioritization of datasets for publication.

Once the data catalogs are created, we will likely need to stagger publication based on resource constraints (in particular for smaller departments), technical challenges associated with extracting data from old systems, and desirability. To prioritize we will engage a number of key stakeholders both internally and externally. Currently, our main method of identifying data priorities is the data request feature on DataSF, which serves as some indicator of desirability. However, one method is not sufficient or strategic and may bias our priorities.

In addition to identifying additional methods to prioritize datasets, we'll also explore gradual approaches to publishing datasets that raise concerns with respect to data quality or discomfort with making it public. For example, we could start by simply listing summary data and then add more detail and dates over time.

One tactic we will explore is to publish datasets in clusters or groups of data related to a critical policy area. Ideally, we will pair the release of issue-specific data in combination with means of presenting and visualizing the data, thereby increasing the value of the data release.

Strategy 1.3. Develop metrics to track and measure progress in publishing open data.

“Number of datasets published” is a blunt metric. While we need to initially increase the number of datasets, we also need to explore measuring publication of high value data, increasing the frequency of updates, responding to data requests or automation of publication. In addition, the creation of data catalogs can provide the basis to create comparable measures between departments, e.g. percent of public datasets that are published. We will also normalize and define what constitutes a dataset. For example, if one organization publishes data by year while another uses year as an attribute, the first organization will appear to be a more prolific publisher.

While the metrics discussed above are important - ultimately they are process measures. Over time, we also need to identify ways of measuring the outcomes and impacts of our open data initiative. In the meantime, the process metrics will provide the basis for an outcomes based evaluation for open data.

Strategy 1.4. Develop our program to automate publication of data. One of the key challenges in opening data is extracting it from legacy systems and then preparing it for broader consumption. Older systems were not designed with data exporting or sharing in mind. Proprietary data formats need to be converted into modern, open formats, or the data may need to be reorganized or structured in a way that supports public distribution. Lastly, the processes that extract, transform and load data should be automated, such that after the initial configuration, we have little to no overhead other than monitoring the ongoing process. In sum, our automation program (activities summarized as extract, transform and load - ETL) is a critical part of our overall program as it will support the key processes that ensure our data is extracted appropriately and published in a timely manner on DataSF.

While a handful of departments may have the resources for ETL, most departments do not. The

City already has a variety of tools to support ETL, including tools that are part of our existing platform, DataSF, and additional licenses related to geographic data. We will identify any gaps in ETL tools and licensing but also identify a services strategy to support ETL. We will allocate our ETL resources based on the priority levels identified under Strategy 1.2.

An important sub-component of our ETL program is how we treat geographic data. The City has a vast store of geographic data in a variety of formats. Some of this data is pointed to via the “external” dataset feature on DataSF. Part of our program will be to determine the best way to disseminate geographic data and how to deploy ETL services to support that method.

Strategy 1.5. Develop an outreach and support program for data coordinators and other data publishers. Through our engagement strategy, we identified a knowledge gap in terms of what resources were available, how to publish data, and who to contact for help. Other localities, including Philadelphia and the state of New York have developed open data guidebooks to facilitate the publication of data. We will leverage and tailor these existing resources and identify additional services as needed to better support the publication of data.

This strategy will be initiated during the establishment of the Data Coordinators and the supporting work to develop data catalogs.

Strategy 1.6. Establish methods to ensure SF licensing and publication of data for new information systems. While extracting data from legacy systems is painful, new systems should be built with open data as a standard output. Any new information system should be required to have automated outputs to support broader publication and dissemination of the city’s data, while retaining the appropriate licensing. We will create and then incorporate contract provisions into the City’s standard buying processes. This will not only stem future ETL costs, but will help normalize and then institutionalize open data.

Other supportive strategies. We believe the strategies detailed above are necessary to achieve the goal of increasing the number and timeliness of datasets on DataSF. Resource permitting, we will also pursue the following strategies.

Assess the value of incorporating or federating non-city data, such as US Census and other national or state survey data. Incorporating non-city data could increase the value of city data by contextualizing and extending it. It could also serve to increase use of DataSF by serving as a central data repository for high value data from diverse sources.

Identify processes to transition uploaded files or externally linked data to machine readable data on DataSF. Our analysis of the data catalog suggests that more than 230 datasets are “external”, meaning they point to other websites that publish or host datasets. Ideally this data will be hosted natively on DataSF, allowing us to realize the full benefits of the platform. The reason this is not a key strategy is because it may be addressed via the data catalogs and prioritization method as well as our approach to managing geographic data (a large contributor to the “external” category).

Assess and develop services to address paper to digital conversion. Some departments have large stores of data locked in paper or even microfiche format. We do not yet know how large a challenge this is. New services that combine character recognition with digitization could help make these datasets accessible.

Identify legislative barriers and recommend changes. Legislation can unwittingly work against the goals of open data. A notable example is wet ink signatures. Requiring wet ink signatures artificially constrains important datasets to paper, increasing publication and dissemination costs. The City in the past has successfully pursued legislative changes (regarding ethics reports) for wet ink signatures, but there may be other opportunities.

Goal 2.

Improve the usability of DataSF

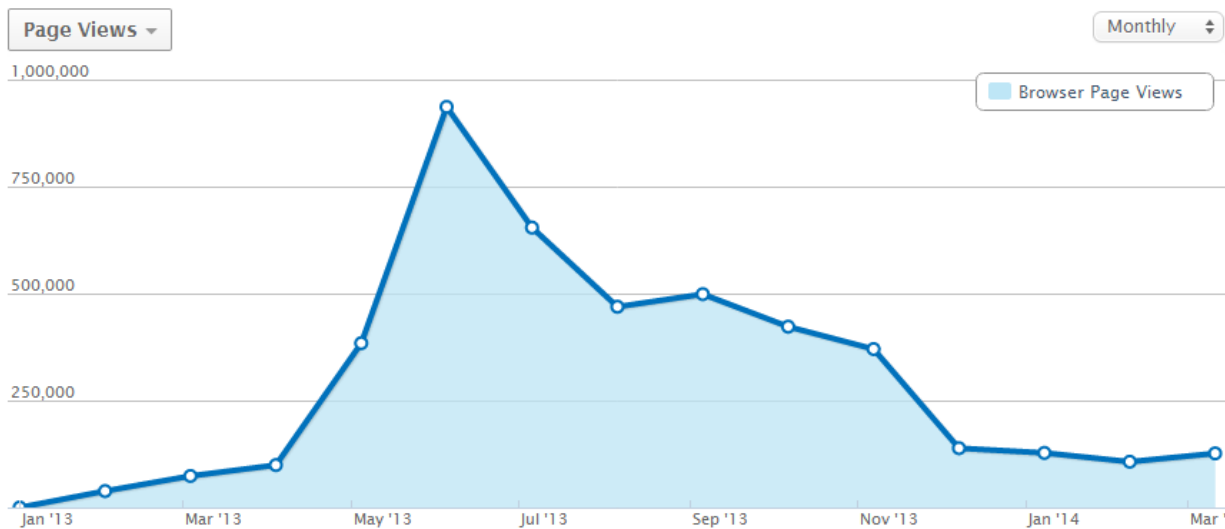
Our data portal, DataSF, is a key part of our open data policy as it creates a single point of entry to our data. To ensure that our open data is readily accessible and used, we need to make sure that the website and the means of accessing the data support the needs of users.

Current State

Summary of Site Traffic

Our data portal is hosted on the Socrata platform, a proprietary vendor. It was launched on March 14, 2012, and site analytics became available in spring of 2013. Data on page views (see Figure 10) shows a large traffic spike in June of 2013⁴, with an elevated page view rate through November. (Note that this does not indicate unique users). Since December of 2013, page views have hovered around approximately 125,000 per month. This indicates continuing but lagging interest in the data portal - the general trend is flat.

Figure 10. Page views by month



We also see that our top datasets from January 1, 2013 - March 31, 2014 are crime, registered businesses, film locations, city lots, and 311 case data (see Figure 11). And the interest in these datasets has been steady over time. Other popular data include building footprints, motorcycle parking, and the neighborhood groups map. Appendix C. includes our top 35 datasets.

⁴ This spike was due to increased visits during the United States Conference of Mayors, Summer 2013 Annual Meeting.

Figure 11. Top datasets by number of views received, January 1, 2013 - March 31, 2014

Top Datasets

Name	
Map: Crime Incidents - Previous Three Months	18,330
Businesses Registered in San Francisco - Active	13,184
Film Locations in San Francisco	9,534
City Lots (Zipped Shapefile Format)	6,499
Case Data from San Francisco 311	6,164

However, our top search term used on our site by a huge margin is “public health” with 95,946 requests - our second top term was “parking” with 4,525 queries. The disconnect between search queries and top datasets may reflect the fact that most of the search results for “public health” return “external” versus natively hosted datasets.

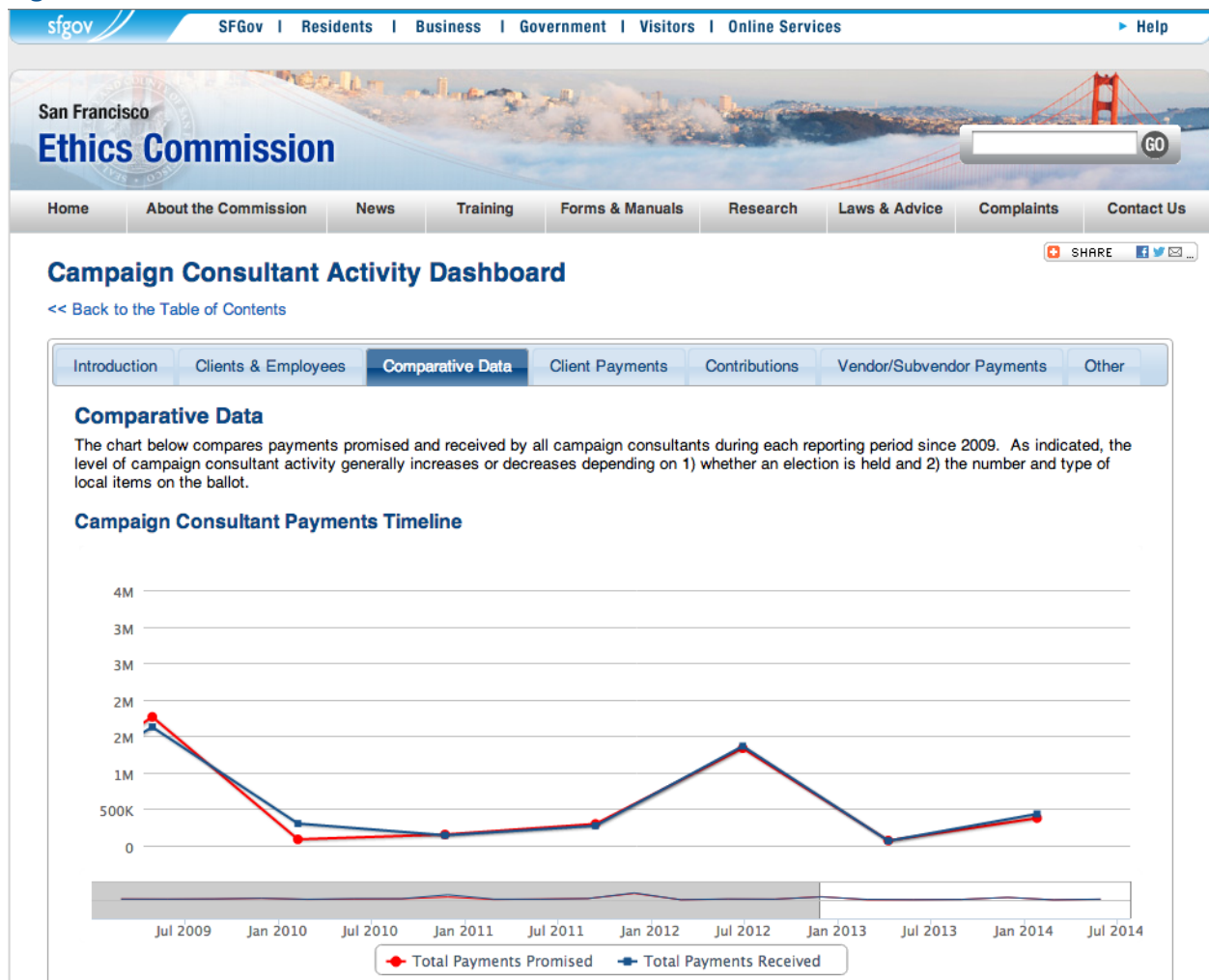
When we look at top referrers (Figure 12 on next page) we see that two variations of DataSF account for more than a third of our top five referrals. This indicates strong branding for DataSF, suggesting that we should retain our domain. Our top referrer is the Ethics Commission, see Figure 12, which has made extensive use of DataSF - not only has a publishing platform but as a means to create dashboards and visualizations on its own site. See Figure 13 on the next page for a screenshot showing how the Ethics Commission creates visualizations using the DataSF platform and then embeds the visualizations into a web page. This makes them the top embedders, i.e. the top data visualizations that have been viewed within an external website.

Figure 12. Top referrers, January 1, 2013 - March 31, 2014

Top Referrers

Name	Referrals
http://www.sfethics.org	62,229
http://www.datasf.org	42,474
https://www.google.com	36,245
http://datasf.org	18,098
http://www.google.com	16,718

Figure 13. Screenshot of the Ethics Commission’s use of DataSF charts on their website



Source: <http://www.sfethics.org/ethics/2013/07/campaign-consultant-activity-dashboard.html>

Usability and User Experience

The experience people have when arriving at our site and their ability to understand what they can do and what data they can access is key to our open data initiative. If our users are intimidated or cannot find what they need, we are not meeting the vision of *creating access to* much less *enabling use of* data. Posting our information is one step in public access and transparency. Making our information easy to use and access is an essential next step.

Our current data portal drops users into an intimidating place that features the site’s analytics but provides no orientation to the nature of the site (see Figure 14). Naive users may be quickly overwhelmed. If they reach for the Help page they are admonished to not use it unless they have a question about functionality. Instead they are told to search the “knowledge base” with no additional help (see Figure 15). We need to do a better job of supporting our users - in particular our new or less technical users.

Figure 14. Screenshot of landing page on DataSF

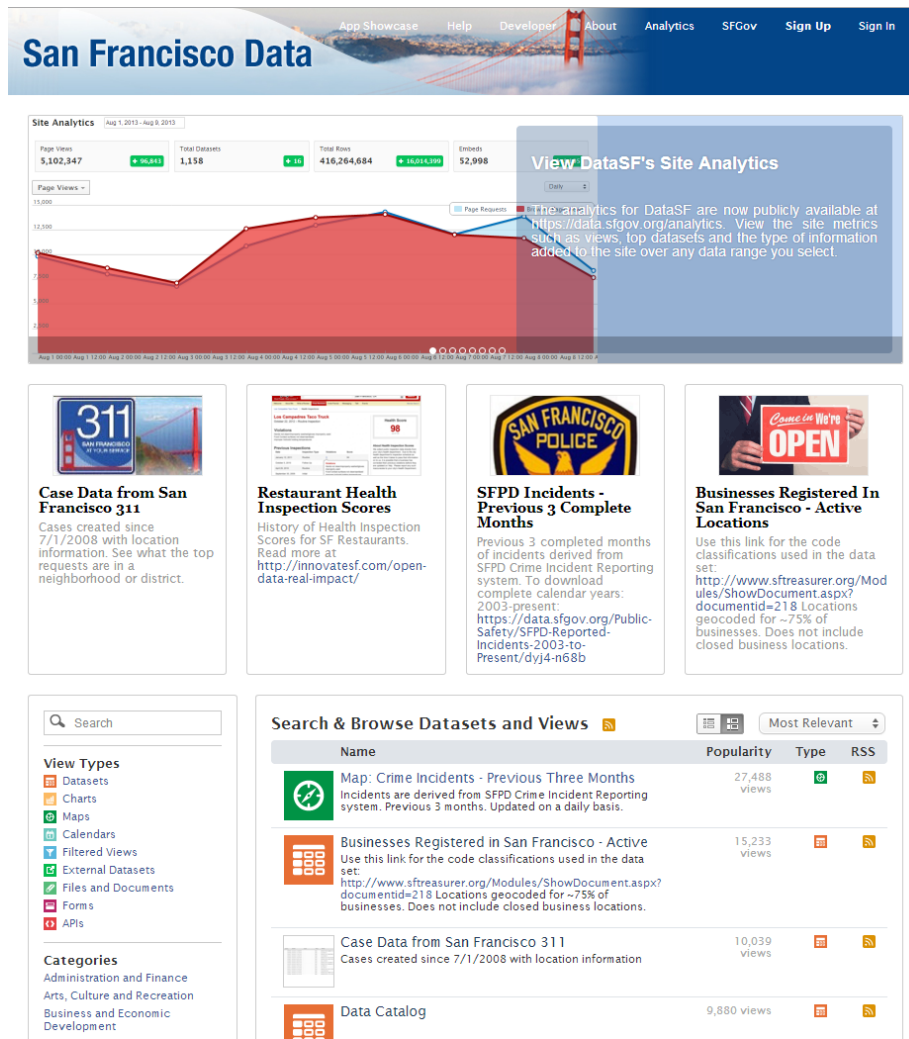
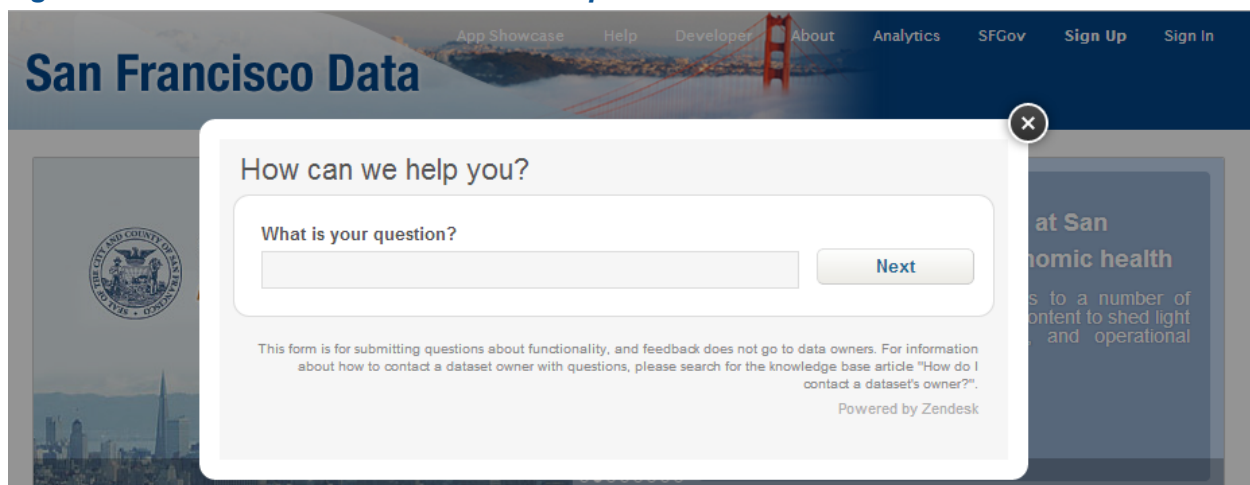


Figure 15. Screenshot of result from "Help" link on DataSF



Strategies for Goal 2

Before we ask departments to upload more data, we must first make it easier for both publishers and consumer to use the portal. While the strategies below assume that we will continue using the Socrata platform, we will continue to assess the technology landscape (see Strategy 6.2) and will make technology changes when it makes sense. To improve the usability and accessibility of DataSF in the near term, we will pursue the following strategies in year one.

Strategy 2.1. Better leverage existing services and features from Socrata. DataSF resides on a proprietary platform provided by Socrata. We were an early adopter of the Socrata platform and after our initial setup, have not made many changes. But in the meantime, Socrata has extended and broadened not only the available features but added significant new functionality. For example, new dashboarding and programming management features could help extend or even replace some of our other open data work.

We need to take a fresh look at what features are available, how we can better leverage them, and then implement accordingly.

Strategy 2.2. Partner closely with Socrata to inform the development of the portal. The Socrata platform is very powerful and has been key to our success in disseminating data. While the previous strategy will help us take better advantage of the existing feature set, we also want to ensure that the platform evolves to meet not only our needs, but our stakeholders and the broader open data community. Some of our key goals in partnering with Socrata are to:


- Improve the overall usability and end user experience.
- Increase the findability of base or original datasets.
- Better integrate timely training and documentation into the portal.
- Better support content management services.

Strategy 2.3. Redesign our web presence and supporting processes and materials to better meet the needs of our users. While the previous two strategies will inform the technical execution of the data platform, we need to do more. Our long term vision is for DataSF to serve as a true multi-sided platform, where data publishers, consumers, and citizens derive broad value from an ecosystem of data, supporting content, visualizations, and services.

To better understand our user's needs we'll leverage usability techniques, including the creation of personas that represent our primary user populations - both existing and desired. Personas are a method of grouping users that represent similar behaviors, goals, or demographics. Figure 16 describes the initial personas that we are considering.

We'll use our personas to identify additional materials that could include technical assistance, demos and how tos, getting started tips and guidance, or non-web based outreach.

Figure 16. Draft Personas for DataSF

	Persona	Description
	Citizen Programmers	Citizen Programmers are comfortable with a broad range of advanced technology, including hadoop, python and R. They want access to Government data to create mobile and web applications to help their City or even create new businesses. This group has been the traditional focus of open data.
	Savvy Analysts	Savvy Analysts are experts at using data within desktop software, such as Excel, Tableau, and ArcView. They want the download button so they can start crunching the numbers. Savvy Analysts come from think tanks, the media, academia and the City itself.
	Community Organizers	Community Organizers belong to community groups and nonprofits and may not have access to the best technical resources or software. However, they are savvy and resourceful, and could benefit from additional information on City data and resources.
	Decision-Makers and Elected Officials	Decision-Makers are looking for information that is presented in a digestible format to help inform or explain their decisions. They hope that public information can foster critical dialogue regarding choices that we make as a City.

Goal 3.

Improve the usability, quality and consistency of our data

While Goals 1 and 2 help provide access to the City's data, the ultimate value of the data depends on its usability, quality, and consistency. Usability helps us understand the data - what is it, how is it collected, when is it published - the basic documentation that supports use of the data. Quality speaks to how reliable and complete the data is - can we trust the conclusions or decisions we make based on the data? Consistency helps us combine data from different systems, by using consistent definitions across datasets, whether it's race, service categories, target populations, location, etc.

Current State

The City's data exists across a vast network of systems and interconnections from mainframe systems to cloud deployments. Input methods, validation and maintenance standards vary across these systems. Moreover, different departments use different categories and definitions for similar data (e.g. race or service type). In some or even many of these cases, differences may have value given the different missions and goals of the departments.

More subtly, the data quality demands of a department may be different from those of an external user. Departments know their data intimately and may use shortcuts and nomenclature that facilitates the efficacy of their own operations. In some cases, it's only when external parties use that same data that questions of data quality and consistency arise. Differences in quality expectation exist even within the City. For example, one department we spoke with has a general need for block level city addresses versus the high granularity and precision needed for another department. When the departments attempt to use the same address set, it often cannot meet both of their needs simultaneously.

Strategies for Goal 3

Given these challenges, this goal is really a multi-year, city-wide goal. City departments and the Department of Technology are key partners in this endeavor. We can play a key role in terms of elevating and framing data quality questions that emerge from data users and establishing metadata standards and data quality measures.

Strategy 3.1. Establish metadata standards for published data. Metadata is data about your data - when was it published, what are the field definitions, who owns the data, etc. Requiring a common set of metadata for our published data will be a first step in increasing data quality but a core step in increasing usability. At a minimum it will increase understanding of what is published, allowing users to more effectively leverage the data. It will also facilitate findability by

having a standard set of fields to search for each published dataset.

It will also provide a common set of information that is consistent across departments allowing us to better track publishing performance, identify areas of inconsistency in data terms or definitions and identify common challenges in data quality.

Strategy 3.2. Establish mechanisms to elicit and track feedback and learnings from data users. Our data users are in an excellent position to provide feedback that can flag or elevate data concerns - whether usability and understanding or quality. While DataSF provides some means for doing this via community ratings and discussion, we may need to extend this or provide additional methods. The new role of data coordinator will probably be key in establishing a successful and sustainable cycle of continuous data improvement.

Strategy 3.3. Explore the creation of data quality processes and measures. Data quality issues arise both in how we input data into our systems and how we maintain it over time. Data quality processes could help standardize how we approach data quality by providing processes for assessing the level of data quality, cleaning data, matching data from different systems, or identifying how we need to change our data input or maintenance processes. Quality measures could help us determine what level of data quality (e.g. level of completeness, adherence to standard rules) we desire and track our progress towards meeting that goal. Given the sheer breadth and depth of City data, measures may need to vary based on department or maturity level. This area requires additional research and assessment to determine the best path forward.

Goal 4.

Enable use of confidential data, while appropriately protecting it

The City collects a tremendous volume of individually identifiable information, some of which is heavily regulated in the case of health or justice-related data. The City also has proprietary or confidential data (e.g. account level water consumption). While the City needs to appropriately protect this data, we also need to enable better access to and use of this data - whether it's cross-department data sharing or means to summarize and publish the data.

Current State

As an example, if the Human Services Agency wants to access individual data from the Department of Public Health, they must engage in a negotiation that if successful, results into a one-to-one agreement using a Memorandum of Understanding. This process can be time-consuming with high administrative costs but is also subject to individual interpretations on what data is appropriate to share and how it should be best protected. Moreover, the agreement may be subject to a point in time sharing of the data, which does not support ongoing use. Ultimately, this limits our ability to share and then use data that could be essential in terms of improving individual outcomes - whether it's tracking use of city services across departments, making referrals, or even collaborating outside of the city governance. For example, the HOPE SF initiative, a project to both revitalize public housing and provide resident services, would like to track outcomes of students that use HOPE SF services. Services are provided by several departments and a key outcome is school performance data, which is held by SF Unified School District.

Moreover, while individual departments have a strong understanding of which of their data must be protected, we have less of a shared understanding across the City of what data is private, requiring enhanced controls, versus what data could be shared more broadly. This runs the risk of both over and under protecting confidential data. Lastly, City residents have little visibility into what data is collected about them and no ready means to access or correct it.

Strategies for Goal 4

The following strategies will help us better manage but also leverage our private data.

Strategy 4.1. Create a data classification and sharing standard. While we need to continue to ensure the confidentiality and integrity of personally identifiable information, we must also find a more effective and efficient means for sharing that data when appropriate and within the boundaries of law. Our goal is to have one framework for efficiently and consistently protecting

and sharing private data. A classification and sharing standard would allow us to classify our data a priori by risk level. Classifying data will allow us to develop a shared understanding of the relative risks posed by diverse data. Based on this classification, we can then create a standard set of controls and rules for not only protecting the data, but also sharing and even publishing if appropriate using aggregation, anonymization or other means. This framework and process will fill a process gap in how we classify and share private data, and reduce individual and organization risk in sharing data.

This strategy may also support Goal 1 by increasing the amount of data we could publish and Goal 5 by enabling more efficient access to private data for analysis.

Strategy 4.2. Create a process for accessing your individual data. A process for accessing data that the City holds about you will increase transparency and may help improve data quality to support Goal 3, which is supported by the metadata standards strategy. Given the distributed nature of individual data, we expect this to be a complex undertaking and we will focus on background research and planning in year one.

Goal 5.

Support increased use of data in decision-making

Once data is available, we need to use it. Open data already supports a broad range of external uses. But it can also better support internal use of data in decision-making - both by creating access to data and enabling new means of displaying and communicating data. We need to match the availability of data with the capacity to use data, both in terms of people and technology.

Current State

The City's charter calls for a performance based budget and the City has used performance metrics to inform funding and decision-making. But there is always room to better use data in decision-making, and both new technologies and types of data demand that we continually evolve our efforts to manage using data.

Beyond performance based budgets, "SFStat" was a management initiative created in March 2004, modeled on the "Citistat" program in Baltimore, MD. The program reviewed service, human resources, and budget data from major departments on a regular basis during public meetings. Subsequent to SFStat, individual departments have pursued their own approach.

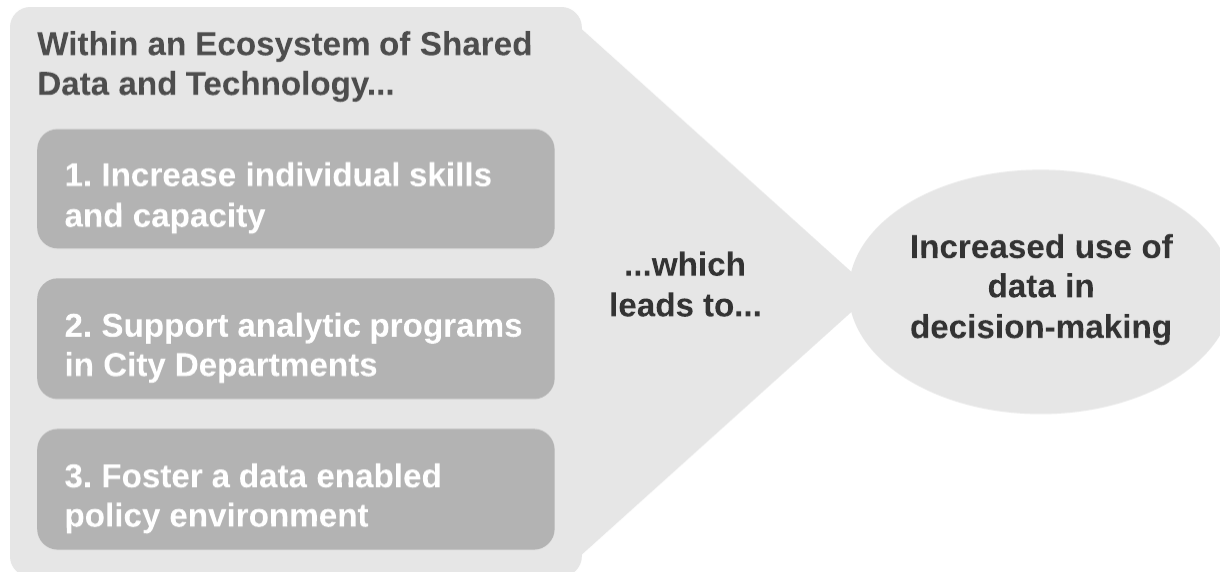
Strategies for Goal 5

Figure 17 provides a schematic of what we believe are the key inputs to use of data in decision-making. At our base, we need to have a set of shared data as well as technology to support data extraction, publishing and dissemination, analysis, and visualization. Much of the rest of our open data work and supporting processes can support creating this base. But we also need 1) analysts who can use, analyse, visualize and communicate about data and 2) managers that know what questions to ask and what data to expect. Next we need engaged departments that use data as a key input into management and decision-making. Lastly, our broader policy environment needs to be able to leverage data in policy discussions among elected officials, City residents, and our many stakeholder groups. Our strategies target each of these inputs, respectively.

Note that at the department level, the strategies below seek to enable and support increased data in decision-making while being agnostic about the particular forms and approaches Departments adopt. Earlier experiences, including SFStat, demonstrate that use of data in decision-making is an act of continuous learning that may vary based on department maturity or capacity, technological constraints, or even priority. So the strategies below focus on meeting

departments where they are and in a way that supports their evolution over time, while simultaneously leveraging broader institutional experience and resources.

Figure 17. Inputs to Support Increased Use of Data in Decision-Making



Strategy 5.1. Establish a training curriculum to support increased use of data in decision-making.⁵ The City is fortunate in that it has a great deal of existing analytical capability and existing networks for sharing experiences (e.g. the City Analysts Network). In addition, the City Services Auditor (CSA) in the Controller’s office has been providing a variety of trainings to promote use of new analytical tools. Partnering with CSA to extend current activities into a broader curriculum could help to support better use of data in decision making. Components of the curriculum should include not only data analysis and tools, but visualization, information design, and communication. As needed, we may also include training and support in using ETL tools and services under Strategy 1.4. One of our findings was that departments, in particular smaller departments, struggle to access data in their backend systems. In addition to using ETL to support open data, we can also leverage the tools to support department access and use of data. This will allow us to more cost-effectively leverage our existing ETL tools and processes.

Lastly, use of data in decision-making requires different approaches to management and so a well-rounded curriculum should cover decision-making in management and executive leadership. An expected outcome of the trainings is to help foster existing and create new learning networks. Over time, we may incorporate elements into the Department of Human Resources existing trainings.

Strategy 5.2. Help establish department stat programs based on department readiness; codify lessons learned and materials for broader use.⁶ Several departments have taken a lead in establishing analytics programs, including the Department of Public Works, Police

⁵ Pending discussions and planning with CSA.

⁶ Ibid.

Department, and MTA, while others are looking to jumpstart analytical efforts. Over the next year, a handful of departments are looking to establish their own programs and have reached out to the City Services Auditor (CSA) for consulting and support services. Based on these projects, the CSA, in collaboration with the CDO, will also develop enduring materials that can be used by other departments in the future.

This strategy will allow us to simultaneously move select departments forward, while codifying lessons from those experiences. This will allow us to better leverage City efforts, identify opportunities for further City-wide support, including technology licensing, and provide a framework for departments to start their own efforts. The materials developed could include guidance on assessing department readiness for a data analytics program, planning templates and documents, case studies, and guidance on technology planning and choices. The projects will also serve as a forum to identify additional policy or standard requirements for not only open data but data management and governance broadly.

Strategy 5.3. Continue to develop our portfolio of transparency tools and websites. The Controller's office provides a variety of transparency tools that track the City's finances, performance, and economy. The tools go beyond simply publishing data to transforming the data into information that can be consumed and understood by the general population. [SFOpenBook](#) allows users to drill into the details of our spending and revenue, as well as our budget, over multiple years. The [Government Barometer](#) tracks performance across a wide variety of departments, from public safety and health and human services, to environment and energy and reports them using friendly, interactive visuals. The [Economic Barometer](#) similarly provides a quarterly summary of key economic indicators, including employment, real estate values, population data and tourism. Figure 18 shows a screenshot from the Economic Barometer.

Each of these tools provides policy makers and the public with ready access to City data contextualized and presented in a way that informs decision-making. In partnership with the Controller's Office, we should continue to develop these but also identify new areas to deploy transparency or dashboarding tools. In some cases this work may overlap with or complement Strategy 6.3.

Figure 18. Employment Summary on the Economic Barometer

- Home
- Quarterly Summary
- Economy-Wide
- Real Estate
- Visitors
- Demographics
- Build Your Own

Employment Summary

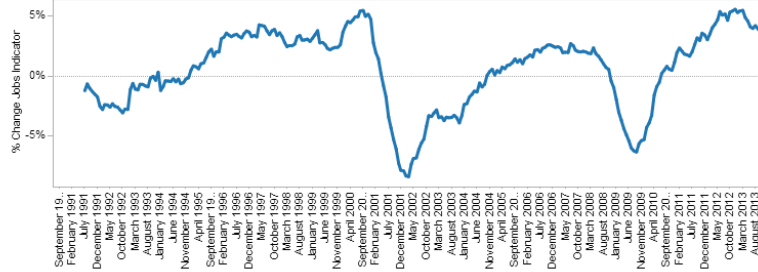
The Employment Summary gives us a monthly look at Metro Division and County-level jobs and unemployment. The latest available data for SF County Employment is 6-9 months behind the latest available data for the Metro Division.

[Data Export Guide](#)
[Economic Indicator Sources](#)

Total Employment MD (use Jobs Indicator drop-down menu to select indicator)



Annual Percentage Change



Share



Download

1,198 views

[See more by this author](#)



Goal 6.

Identify and foster innovations in open data and data use

The pace of change in the open data, analytics, and visualization spaces is breathtaking. We need to not only ensure we are aware of innovations, but we need to selectively identify and nurture innovation in order to ensure that the City and our stakeholders benefit from changes in technology and the experiences of others.

Current State

As discussed in the first section, San Francisco has been a leader in the open data movement. And our leadership has not been limited to simply having an open data policy and program. Work to create standards in housing and health inspections have broadly benefited the open data community. We've also been early to adopt new applications that have then been leveraged in other localities. However, we need to consciously engage and dedicate time and resources to maintain the pace of innovation.

Strategies for Goal 6

The following strategies will help us continue to identify and foster innovation - not only in open data but in use of open data.

Strategy 6.1. Develop and maintain a communications and engagement strategy. A robust communications and engagement strategy will help us identify new ideas and approaches to using the City's data. But we also need to do a better job of communicating our activities, our plans, and our struggles in order to broaden the benefit of our experience.

We'll leverage multiple channels, including the portal and our broader web presence, to feature new data ideas or pilots happening in the City. We'll also contribute to learning networks in the government, non-profit and private sectors. For learning networks within the city, we will work to foster a culture of data use and collaboration and highlight the work that is already happening. And we'll participate in events not only in the open data and civic programmer community, but the evolving Chief Data Officer communities. Part of this engagement strategy will be to have regular interactions with our data users (and publishers), both internally and externally.

Strategy 6.2. Conduct ongoing reviews of best practices and the changing technology landscape. To ensure that San Francisco maintains its leadership position in open data, we have to stay abreast of emerging best practices and changes in technology that can better support or even transform our program. In part, this will be a natural result of our

communications and engagement strategy, but retaining it as a specific strategy will help ensure that we are making regular and conscious efforts to assess the rapidly changing landscape.

Strategy 6.3. Identify and enable targeted data-centric initiatives. Through our engagement strategy and ongoing reviews we hope to identify opportunities for targeted data initiatives. These might range from identifying a new application to cross department collaborations on data sharing or analysis to creating new data standards. We can then leverage pilots and demonstrations using internal resources, leveraging hackathons and datapaloozas, or issuing challenges.⁷ We will also explore private/public partnerships for pooling technical or other resources. Given our limited resources, we will select initiatives based on criteria that may include:

- Does it involve the publication of a new dataset?
- Does it address a pressing problem or information need?
- Is it easily achievable? Or do we have the right resources in place?
- Does it have cross-department benefits?
- Does it create broad value for the open data movement?

Note that in some cases, initiatives in support of this strategy will overlap with our transparency initiatives under Strategy 5.3.

Strategy 6.4. Establish a data licensing framework and standard. Part of releasing our data is to ensure that it can be bent, folded, and remade for uses that we did not imagine. A key part of fostering known and unknown future uses is to have the correct licensing framework and the most efficient means to deploy that framework.

⁷ Hackathons and datapaloozas are events that can be hosted by an government entity, a nonprofit or other group. The events bring together a range of people from technologists and data scientists to community groups and government. The goal is leverage talent inside and outside of government to brainstorm ideas or even create solutions to solve a public problem using technology and government data. Challenges are similar to hackathons and datapaloozas in terms of objectives and types of participants. However, the participants “compete” to provide solutions to the problem. Depending on how the challenge is conducted, it could include prize money (usually provided by a foundation) or name recognition or other non-monetary awards.

6. Prioritization, Resource, and Risk Analysis

Our year 1 strategic plan is ambitious and reflects a vision of what we hope to accomplish. Due to limited resources, we may not be able to deliver on all aspects of our strategic vision in year 1. However, by fully articulating our vision, we are better able to prioritize and allocate the resources we do have. In the sections below, we prioritize our proposed strategies, conduct a gap analysis and contingency plan. We also identify major risks and key mitigations.

6.1 Priority, Resource Gap Analysis, and Contingency Plan for Proposed Strategies

The Open Data Ordinance mandates some of our year one activities, while others are either in the critical path for broader work or a key part of setting a platform for future success. As a result, we prioritized our strategies using the MoSCoW method in the context of what we feel must happen in Year 1 (M=Must, S=Should, C=Could).⁸ This does not mean that certain activities will not become “musts” or “shoulds” after the first year or even first six months.

We then identified resource gaps as follows:

- No - no resource gap
- Yes - we do not believe we can be successful with existing resources
- Partial - the strategy can be supported at a minimal level with current resources, but should be supplemented to ensure success

We then characterized the gap and described the general resource strategy for each gap type:

Gap Type	Description	General resource strategy
Y1 Gap & Ongoing Need	Indicates areas of greatest resource gaps because we have immediate needs, and expect long term resource demands.	<ul style="list-style-type: none"> • Prioritize using near term resources (interns or fellows, borrowed FTEs, key partnerships). • Seek dedicated FTE support over time.
Y1 Gap & Maintenance	Indicates a resource gap in an activity that we feel we need to complete in Year 1, but we do not expect to have high and ongoing resource demands to support the strategy.	<ul style="list-style-type: none"> • Prioritize using near term resources (interns or fellows, borrowed FTEs, key partnerships). • Do not seek dedicated FTE support.
Gap & Ongoing Need	Indicates a resource gap in an activity that we could delay if we cannot resource appropriately. However, we expect to have an ongoing need to resource this activity.	<ul style="list-style-type: none"> • After higher priority strategies are on track, address using near term resources (interns or fellows, borrowed FTEs, key partnerships). • Seek dedicated FTE support over time.

⁸ MoSCoW prioritization is traditionally used in software development to determine what requirements you Must have, Should have, Could have, and Won't have. In our case, we used it to prioritize our activities.

Gap & Maintenance	Indicates a resource gap in an activity that we could delay if we cannot resource appropriately. We do not expect an ongoing need to resource this activity.	<ul style="list-style-type: none"> • After higher priority strategies are on track, address using near term resources (interns or fellows, borrowed FTEs, key partnerships). • Do not seek dedicated FTE support.
-------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Lastly, the table also includes a contingency approach if we are unable to close the resource gap.

Table: Prioritization, Gap Analysis and Contingency Plan

Strategy	M	S	C	Resource Gap	Type of Gap	Contingency Plan if Unable to Close Gap
Strategy 1.1. Establish the role of data coordinators and support development of data catalogs.	X			Partial	Y1 Gap & Maintenance	Scale down support and outreach, e.g. provide only guidance documents on data catalogs versus consulting/advising
Strategy 1.2. Develop methods to inform the prioritization of datasets for publication.	X			Partial	Y1 Gap & Maintenance	Rely on surveys and less resource intensive methods
Strategy 1.3. Develop metrics to track and measure progress in publishing open data.		X		Partial	Gap & Maintenance	Use metrics based on the data catalogs; Defer research on broader evaluation metrics
Strategy 1.4. Develop our program to automate publication of data.	X			No ⁹	No Gap* (projected)	
Strategy 1.5. Develop an outreach and support program for data coordinators and other data publishers.	X			Partial	Y1 Gap & Ongoing Need	Limit support to open data guidebook; Minimize in person outreach and support
Strategy 1.6. Establish methods to ensure SF licensing and publication of data for new information systems.	X			No	No Gap*	
Strategy 2.1. Better leverage existing services and features from Socrata.		X		Partial	Y1 Gap & Maintenance	Slow down adoption of new services; limit expansion to cost free services
Strategy 2.2. Partner closely with Socrata to inform the development of the portal.		X		No	No Gap	
Strategy 2.3. Redesign our web presence and supporting processes and materials to better	X			Partial	Y1 Gap & Ongoing Need	Slow down implementation, rely on web-based materials, leverage city web support

⁹ Our preliminary analysis of both our technology licensing and staffing suggest that we have sufficient resources in year 1. We expect the prioritization of datasets to help regulate the volume of ETL service demand.

meet the needs of our users.						
Strategy 3.1. Establish metadata standards for published data.	X		No	Y1 Gap & Maintenance*	Slow down implementation	
Strategy 3.2. Establish mechanisms to elicit and track feedback and learnings from data users.			X Partial	Gap & Ongoing Need	Rely on current feedback mechanisms and/or implement simple web-based forms	
Strategy 3.3. Explore the creation of data quality processes and measures.			X Yes	Gap & Ongoing Need	Slow down research; Defer to year 2	
Strategy 4.1. Create a data classification and sharing standard.		X	Yes	Y1 Gap & Maintenance*	Identify key partner; Slow down implementation; Defer to year 2	
Strategy 4.2. Create a process for accessing your individual data.	X		No ¹⁰	No Gap*		
Strategy 5.1. Establish a training curriculum to support increased use of data in decision-making.		X	Yes	Gap & Ongoing Need	Identify key partner; Defer development and possibly create small set of internally developed classes	
Strategy 5.2. Help establish department stat programs based on department readiness; codify lessons learned and materials for broader use		X	Yes	Gap & Ongoing Need	Identify key partner; Restrict effort to fostering learning group (e.g. listserv and possibly ongoing group meetings)	
Strategy 5.3. Continue to develop our portfolio of transparency tools and websites.		X	Yes	Gap & Ongoing Need	Identify key partner; Defer development of new tools; Select only technically trivial projects with low data effort	
Strategy 6.1. Develop and maintain a communications and engagement strategy.		X	Partial	Y1 Gap & Ongoing Need	Decrease number and types of engagement	
Strategy 6.2. Conduct ongoing reviews of best practices and the changing technology landscape.		X	No	No Gap		
Strategy 6.3. Identify and enable targeted data-centric initiatives.		X	Yes	Y1 Gap & Ongoing Need	Identify key partners; Minimize complexity of selected projects; Highly limit type and number	
Strategy 6.4. Establish a data licensing framework and standard.	X		No	No Gap*		

* For these strategies, we assume that we will be able to create working groups or teams using existing City staff and resources. However, our timeline will be dependent on staff and resource availability.

¹⁰ In this case, we are only referring to resources to create the process. We expect there may be significant gaps in executing the process.

6.2 Major Risks

In addition to the risks posed by resource constraints, our strategic plan faces a handful of major risks. In the table below we describe each major risk and how we are attempting to mitigate it. Some proposed mitigations rely on closing resource gaps discussed in the previous section. Where we cannot close resource gaps, we have unmitigated program risk.

Major Risk	Description	Discussion and Key Mitigations
Dependence on Department Data Coordinators	The role of Data Coordinators is key to making open data work but is also an unfunded extension of current duties. Not all departments will be able to resource this role adequately.	Provide extensive support per Strategy 1.1 and 1.5.* Fund ETL activities via Strategy 1.4.* Slow down rate of publication where appropriate. Demonstrate internal value and build executive support via Strategy 6.1.*
Dependence on early stage vendor	The vendor hosting our DataSF platform is an early stage company in the technology sector, which is notoriously darwinian.	While the commercial success of Socrata is not under our control, recent venture capital funding (\$18 million on June 26, 2013) for Socrata indicates viability. In addition, Socrata's client base continues to grow. To further mitigate our risk, we will continue to monitor Socrata's viability but also work to include contract provisions that would allow for a controlled exit from Socrata with respect to our data.
Competing priorities	In the face of tangible and critical challenges such as homelessness, housing affordability, children's services and many many more, open data can sometimes feel a bit lofty and esoteric - even though it can provide value and support to each of these urgent issues.	Demonstrate value and build executive support via Strategy 6.1.* Use high priority initiatives and topics to inform the selection of activities to support Strategies 5.3 and 6.3.* Note, however, that each of these activities are "should" versus "must" do in year 1.
Inability to meet demand for ETL services	As discussed in Strategy 1.4, our ability to extract data from systems and release it is key to open data. While our preliminary analysis suggests that we have sufficient ETL resources in Year 1, unexpected demand could slow down our open data rollout.	Our goal is to resource ETL per Strategy 1.4 in a way that does not depend on smaller or under resourced departments "paying" for ETL. However, we do expect larger departments to conduct most of their own ETL work with only the central service only providing consulting. This should help moderate any increased demand.

*Indicates key dependency on closing resource gap discussed in section 6.1.

7. Conclusion

San Francisco has been a leader and innovator in the open data movement. And we are uniquely positioned to continue to do so. From our geographic location at the center of technology to our vast stores of in house analytical capability to the world class research and nonprofit institutions sitting in our backyard - San Francisco can leverage its work for national and world wide innovation in open data. We are also discovering that open data has an important role to play in terms of enabling greater use of our own data.

While this plan is ambitious, it is built on a foundation of our existing work in open data and data initiatives within the City. By integrating our disparate activities, this plan creates a cohesive vision of how we can achieve our goals. All that remains is for us to commit to it.

“...the true work of innovation is not coming up with something big and new, but instead recombining things that already exist.” --Erik Brynjolfsson & Andrew McFee

Appendices

Appendix A. Engagement methods

Engagement methods for City Staff

Group	Engagement Strategy(ies)
Department and division heads	One on one and department meetings
Data managers and champions of internal and external data initiatives	One on one meetings and in-depth discussions; reviews of existing initiatives and supporting assessments where available
City Analysts (e.g. the City Analyst Network, GIS Network)	Surveys, brown bags and targeted one on one meetings
Public Information Officers	Group meeting

Engagement methods for external groups and partners

Group	Engagement Strategy(ies)
Citizens of San Francisco	Survey on DataSF (in progress)
Community and neighborhood groups	Survey on DataSF (in progress)
Non-profits organizations serving the City	Survey on DataSF (in progress), working to establish ongoing engagement strategy
Civic Hackers and Programmers	Survey on DataSF (in progress), one on one meetings, group presentation and discussion, hack nights and hackathon
Technology Sector	Survey on DataSF (in progress) and one on one meetings
National open data organizations	Review of sites and resources, phone calls
Peers in other localities	Phone calls, meetings, conferences

Appendix B. Cross walk between plan and Open data ordinance

Sec. 22D.2. Chief Data Officer and City Departments

(a) Chief Data Officer

#	Clause	Implementation
(a)	Chief Data Officer. In order to coordinate implementation, compliance, and expansion of the City's Open Data Policy, the Mayor shall appoint a Chief Data Officer (CDO) for the City and County of San Francisco. The CDO shall be responsible for drafting rules and technical standards to implement the open data policy, and determining within the boundaries of law which data sets are appropriate for public disclosure. In making this determination, the CDO shall balance the benefits of open data set forth in Section 22D.1, with the need to protect from disclosure information that is proprietary or confidential and that may be protected from disclosure in accordance with law. Nothing in the rules and technical standards shall compel or authorize the disclosure of privileged information, law enforcement information, national security information, personal information, unless required by law. Nothing in the rules or technical standards shall compel or authorize the disclosure of information which is prohibited by law.	This document serves to meet the general expectations. Strategy 4.1 will protect proprietary or confidential information.
(b)	The CDO's duties shall include, but are not limited to the following:	-
(b)(1)	Draft rules and technical standards to implement the open data policy ensuring the policy incorporates the following principles:	
(b)(1)(A)	(A) Data prioritized for publication should be of likely interest to the public;	See Strategy 1.2
(b)(1)(B)	(B) Data sets should be free of charge to the public through the web portal;	Existing practice
(b)(1)(C)	(C) Data sets shall not include privileged or confidential information, law enforcement information, national security information, personal information, proprietary information or information the disclosure of which is prohibited by law; and	See Strategy 4.1
(b)(1)(D)	(D) Data sets shall include, to the extent possible, metadata descriptions, API documentation, and the description of licensing requirements. Common core metadata shall, at a minimum, include fields for every dataset's title, description, tags, last update, publisher, contact information, unique identifier, and public access level as defined by the CDO.	See Strategy 3.1
(b)(2)	(2) Coordinate, maintain, and update the City's Open Data website, currently known as "DataSF";	See Goal 2
(b)(3)	(3) Present the Open Data rules and technical standards to the Committee on Information Technology (COIT) for adoption;	COIT will be the forum used to pass rules and technical standards.
(b)(4)	(4) Provide education and analytic tools for City departments to improve and assist with the release of open data to the public;	See Strategies 1.1, 1.5, 2.3
(b)(5)	(5) Assist departments by collecting and reviewing each department's open data implementation plans and creating a template for the departmental quarterly progress reports;	See Strategies 1.1, 1.5

(b)(6)	(6) Present an annual citywide implementation plan to COIT, the Mayor, and Board of Supervisors and respond, as necessary, to inquiries regarding the implementation of the open data policy and the compliance of departments with the deadlines established in this section.	This plan will be presented to all of these groups.
(b)(7)	(7) Help establish data standards within and outside the City through collaboration with external organizations;	See Strategies 3.1, 4.1, 6.1
(b)(8)	(8) Assist City departments with analysis of City data sets to improve decision making;	See Goal 5 and Strategy 6.3
(b)(9)	(9) Establish a process for providing citizens with secure access to their private data held by the City;	See Strategy 4.2
(b)(10)	(10) Establish guidelines for licensing open data sets released by the City and evaluate the merits and feasibility of making City data sets available pursuant to a generic license, such as those offered by "Creative Commons." Such a license could grant any user the right to copy, distribute, display and create derivative works at no cost and with a minimum level of conditions placed on the use; and,	See Strategy 6.4
(b)(11)	(11) Prior to issuing universally significant and substantial changes to rules and standards, solicit comments from the public, including from individuals and firms who have successfully developed applications using open data sets.	See Strategy 6.1; Rules and standards will also be presented to COIT, a public forum

(b) City Departments

#	Clause	Implementation
(b)	Each City department, board, commission, and agency ("Department") shall:	-
(b)(1)	Make reasonable efforts to make publicly available all data sets under the Department's control, provided however, that such disclosure shall be consistent with the rules and technical standards drafted by the CDO and adopted by COIT and with applicable law, including laws related to privacy;	Supported by Strategy 1.1, 1.5, 3.1, 4.1
(b)(2)	Review department data sets for potential inclusion on DataSF and ensure they comply with the rules and technical standards adopted by COIT;	Supported by Strategy 1.1, 1.5, 3.1, 4.1
(b)(3)	Designate a Data Coordinator (DC) no later than three months after the effective date of Ordinance No. _____, who will oversee implementation and compliance with the Open Data Policy within his/her respective department. Each DC shall work with the CDO to implement the City's open data policies and standards. The DC shall prepare an Open Data plan for the Department which shall include:	Supported by Strategy 1.1, 1.5; See Timeline
(b)(3)(A)	A timeline for the publication of the Department's open data and a summary of open data efforts planned and/or underway in the Department;	See Strategy 1.1, 1.2
(b)(3)(B)	A summary description of all data sets under the control of each Department (including data contained in already-operating information technology systems);	See Strategy 1.1

(b)(3)(C)	All public data sets proposed for inclusion on DataSF;	See Strategy 1.1
(b)(3)(D)	Quarterly updates of data sets available for publication.	See Strategy 1.1, 1.2, 1.3
(b)(4)	The DC's duties shall include, but are not limited to the following:	
(b)(4)(A)	No later than six months after the effective date of Ordinance No. _____, publish on DataSF, a catalogue of the Department's data that can be made public, including both raw data sets and application programming interfaces ("API's").	See Strategy 1.1 and Appendix D. for estimated timelines
(b)(4)(B)	Appear before COIT and respond to questions regarding the Department's compliance with the City's Open Data policies and standards;	Will be done as needed
(b)(4)(C)	Conspicuously display his/her contact information (including name, phone number or email address) on DataSF with his/her department's data sets;	Supported by Strategy 1.1
(b)(4)(D)	Monitor comments and public feedback on the Department's data sets on a timely basis and provide a prompt response;	Supported by Strategy 1.2, 3.2
(b)(4)(E)	Notify the Department of Technology upon publication of any updates or corrective action;	Existing practice
(b)(4)(F)	Work with the CDO to provide citizens with secure access to their own private data by outlining the types of relevant information that can be made available to individuals who request such information;	See Strategy 4.2
(b)(4)(G)	Implement the privacy protection guidelines established by the CDO and hold primary responsibility for ensuring that each published data set does not include information that is private, confidential, or proprietary; and	Supported by Strategy 1.1, 1.5; See Strategy 4.1
(b)(4)(H)	Make reasonable efforts to minimize restrictions or license-related barriers on the reuse of published open data.	See Strategy 6.4

(c) Department of Technology

#	Clause	Implementation
(c)	The Department of Technology (DT) shall provide and manage a single Internet site (web portal) for the City's public data sets (http://data.sfgov.org or successor site), called "DataSF." In managing the site, DT shall:	Current practice - Note that 311 has been managing DataSF
(c)(1)	Publish data sets with reasonable, user-friendly registration requirements, license requirements, or restrictions that comply with the rules and technical standards drafted by the CDO and adopted by COIT;	Current practice
(c)(2)	Provide mechanisms for departments to indicate data sets that have been recently updated;	Current practice, though we want to strengthen this per Strategy 6.1
(c)(3)	Include an on-line forum to solicit feedback from the public and to encourage public discussion on Open Data policies and public data set availability;	Current practice though we plan to improve this via Strategy 3.2
(c)(4)	Forward open data requests to the assigned DC; and,	Current practice, though this is done by 311 which has been managing the open data portal.
(c)(5)	Take measures to ensure access to public data sets while protecting	Current practice, though in

	DataSF from unlawful abuse or attempts to damage or impair use of the website.	practice this is managed by our vendor, Socrata
--	--------------------------------------------------------------------------------	-------------------------------------------------

Sec. 22D.3. Standards and Compliance

#	Clause	Implementation
(a)	The CDO and COIT shall work with the Purchaser to develop contract provisions to promote Open Data policies. The provisions shall include rules for including open data requirements in applicable City contracts and standard contract provisions that promote the City's open data policies, including, where appropriate, provisions to ensure that the City retains ownership of City data and the ability to post the data on data.sfgov.org or make it available through other means.	See Strategy 1.6
(b)	The following Open Data Policy deadlines are measured from effective date of Ordinance No. _____:	During the passage of this policy, the deadlines were made dependent on the CDO hire.
(b)(1)	Within three months, department heads designate Department Data Coordinators to oversee implementation and compliance with the Open Data Policy within his/her respective department;	See timeline in Appendix D - we expect most Data Coordinators to be appointed by May 30.
(b)(2)	Within six months, each Department shall begin conducting quarterly reviews of their progress on providing access to data sets requested by the public through the designated web portal;	The initial data catalog will serve as this review and maintenance of the data catalog will serve as the key review input.
(b)(3)	Within six months, each Department shall publish on DataSF a catalogue of their Department's data that can be made public, including both raw datasets and APIs; and	As permitted under (b)(5) below, we are extending this timeline to allow for six months for the data catalog. See timeline in appendix D.
(b)(4)	Within one year, the CDO shall present updated citywide Open Data implementation plan to COIT, the Mayor and Board of Supervisors.	The Open Data plan will be presented annually starting in late Spring/early summer of 2014.
(b)(5)	The CDO may propose a modification, for adoption by COIT, of the timelines set forth in the legislation.	See the Timeline in Appendix D.

Appendix C. Additional Detail on Site Analytics

Top 35 Datasets

Name	
Map: Crime Incidents - Previous Three Months	18,330
Businesses Registered in San Francisco - Active	13,184
Film Locations in San Francisco	9,534
City Lots (Zipped Shapefile Format)	6,499
Case Data from San Francisco 311	6,164
Data Catalog	6,161
SFPD Incidents - Previous Three Months	4,888
Building Footprints (Zipped Shapefile Format)	4,748
Motorcycle Parking Map	4,467
Neighborhood Groups Map	4,399
Restaurant Scores	3,507
SFPD Reported Incidents - 2003 to Present	3,428
San Francisco Pipeline Map Fourth Quarter 2012	3,359
Data Catalog for 311 Web	3,266
Street Names	3,063
Parking meters	2,855
Bicycle Parking (Public)	2,794
Streets of San Francisco (Zipped Shapefile Format)	2,124
SFGIS Data Catalog - Internal	1,966
Total Contributions for all Candidate Controlled Committees - November 5, 2013 Election	1,877
HSA 90 day emergency shelter waitlist	1,822
Zoning Districts	1,634
Elevation Contours (Zipped Shapefile Format)	1,616
Mobile Food Permit Map	1,514
Salary Ranges by Job Classification	1,397
Third-Party Spending in Support or Opposition of Candidates - November 6, 2012 Election	1,381
Map of Development Pipeline Second Quarter 2013	1,368
Mobile Food Facility Permit	1,348
Campaign Finance - Cash Balance of Active Campaign Committees (No Table)	1,329
Campaign Finance - Active Committees with Outstanding Debts (No Table)	1,302
Local Business Enterprise Directory	1,194
Campaign Finance - SFEC 1.126 Notification of Contract Approval Filings	1,148

San Francisco Basemap Street Centerlines (Zipped Shapefile Format)	1,084
Bay Area - General (Zipped Shapefile Format)	1,074
Registered Business Map	1,011

Appendix D. High-Level Timeline

For each of our strategies, we outline a high-level working timeline and expected resources. Adjustments to the timeline may occur based on resources or other factors as discussed in Section 6. The timeline includes Year 2 as some activities start in Year 1 but extend to Year 2. You can view [the timeline in a google spreadsheet](#).

Appendix E. Acknowledgements

A number of people, too numerous to list, helped inform the development of this plan. Many thanks to everyone I spoke with and for your many insights. Thank you to all the individuals who took the time to respond to my survey questions and participate in learning lunches and meetings.

A big shout out to the local brigade, Code for San Francisco, and Code for America for their insights and input and for being so welcoming. And many thanks to colleagues working in other places and on similar challenges - Dianna Anderson, Barbara Cohn, Millie Crossland, Ali Farahani, Brett Goldstein, Mark Headd, Laura Meixell, Jonathan Reichental, Tom Schenk, and Kel Wetherbee.

The Open Data Internal Advisory Group provided guidance and strategic direction. A big thank you to the members of this group - Carmen Chu, Luis Herrera, Kate Howard, Ed Reiskin, Ben Rosenfield, and Marc Touitou.

Special thanks go to Alexandra Bidot, Cyndy Comerford, Mathias Gibson, Kate Howard, Chanda Ikeda, Jeff Johnson, Lani Kent, Jason Lally, Sherman Luk, Andy Maimoni, Ephrem Naizghi, Jay Nath, and Tajel Shah for providing detailed input, feedback and insight into forming this plan.