Open Data in San Francisco

Joy Bonaguro
Chief Data Officer
City and County of San Francisco

Civic Tech and Smart Cities Workshop
SF State, June 8, 2015
Phases of Open Data

Phase 1: The Executive Edict

Phase 2: The Publishing Scramble

Phase 3: The Period of Stagnation

Phase 4: The Resource Reckoning

Phase 5: The Integration of the Initiative
CENTRAL HYPOTHESIS: DataSF can help solve our internal data challenges
How to Broaden the Value of Open Data?

Find, feel and fix the PAIN.
IGNORANCE IS THE CURSE OF GOD

KNOWLEDGE IS THE WING

WHEREWITH WE FLY TO HEAVEN.

“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”
<table>
<thead>
<tr>
<th>M</th>
<th>S</th>
<th>T</th>
<th>AA</th>
<th>AB</th>
<th>AC</th>
<th>AD</th>
<th>AE</th>
<th>AF</th>
<th>AG</th>
<th>AV</th>
<th>AW</th>
<th>AX</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td>M6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>clip</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Coordinators Portal
Resources for San Francisco's Data Coordinators

Overview

This portal is designed to help you, as a City Data Coordinator, orient to your role and make sure you have the resources you need to execute on your responsibilities.

The data coordinator is responsible for:
- Developing a catalog of datasets that can be made public (see tips for creating your inventory)
- Monitoring and responding to comments on public datasets
- Monitoring and tracking the rollout of public datasets on the open data portal
- Ensuring that appropriate processes are in place to exclude private, confidential or proprietary data

In addition to the responsibilities, data coordinators will be supported in the work through:
- Training and workshop opportunities to develop professionally

Tasks

Next Up

- Rolling deadlines through April
- Dataset prioritization - as data coordinators complete inventory, prioritization is assigned

In the future

- Through May
  - Publishing plans - assigned based on completion of all prior steps. None assigned yet.
<table>
<thead>
<tr>
<th>Inventory ID</th>
<th>Department or Division</th>
<th>Dataset Name</th>
<th>Dataset Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>311-0001</td>
<td>311</td>
<td>311 Cases by Channel</td>
<td>These are the service requests created by 311, categorized by channel.</td>
</tr>
<tr>
<td>311-0002</td>
<td>311</td>
<td>311 Call Metrics</td>
<td>Monthly metrics from the SF311 Call Center, including call data.</td>
</tr>
<tr>
<td>311-0003</td>
<td>311</td>
<td>311 Information Requests by Month</td>
<td>SF311 launched its current knowledge base in July 2012.</td>
</tr>
<tr>
<td>311-0004</td>
<td>311</td>
<td>Case Data from San Francisco 311</td>
<td>SF311 cases created since 7/1/2008 with location information.</td>
</tr>
<tr>
<td>311-0005</td>
<td>311</td>
<td>Centralized Commission Database</td>
<td>Centralized Commission Database, stores the key informants.</td>
</tr>
<tr>
<td>ADP-0001</td>
<td>Adult Probation</td>
<td>Weekly active clients (ctag)</td>
<td>Contains all active probationers, including case information.</td>
</tr>
<tr>
<td>ADP-0002</td>
<td>Adult Probation</td>
<td>Weekly assessment data (ctag)</td>
<td>Contains most recent assessment result for all active clients.</td>
</tr>
<tr>
<td>ADP-0003</td>
<td>Adult Probation</td>
<td>Active client address - monthly (ctag)</td>
<td>Contains most recent address entries, by address type (reformat).</td>
</tr>
<tr>
<td>ADP-0004</td>
<td>Adult Probation</td>
<td>Active client race - monthly (ctag)</td>
<td>Contains most recent race entries for each active client at Prototypes.</td>
</tr>
<tr>
<td>ADP-0005</td>
<td>Adult Probation</td>
<td>Active client language - annual (ctag)</td>
<td>Contains most recent primary and secondary language information.</td>
</tr>
<tr>
<td>ADP-0006</td>
<td>Adult Probation</td>
<td>COMPAS assessment data - as requested (northpoint)</td>
<td>Contains subscales, need scores, and assessment data.</td>
</tr>
<tr>
<td>ADP-0007</td>
<td>Adult Probation</td>
<td>COMPAS PSI data - as requested</td>
<td>Contains all information in Pre-Sentence Report submitted to Prototypes.</td>
</tr>
<tr>
<td>ADP-0008</td>
<td>Adult Probation</td>
<td>COMPAS family impact statement data - as requested</td>
<td>Contains information included in the Family Impact Statement Report.</td>
</tr>
<tr>
<td>ADP-0009</td>
<td>Adult Probation</td>
<td>PRCS spreadsheet - active clients</td>
<td>Contains all active PRCS clients and pending releases to facility.</td>
</tr>
<tr>
<td>ADP-0010</td>
<td>Adult Probation</td>
<td>PRCS spreadsheet - completions</td>
<td>Contains all completed PRCS clients, including case information.</td>
</tr>
<tr>
<td>ADP-0011</td>
<td>Adult Probation</td>
<td>1170 spreadsheet - active clients</td>
<td>Contains all active 1170 clients and pending releases to 1170 clients.</td>
</tr>
<tr>
<td>ADP-0012</td>
<td>Adult Probation</td>
<td>1170 spreadsheet - completions</td>
<td>Contains all completed 1170 clients, including case information.</td>
</tr>
<tr>
<td>ADP-0013</td>
<td>Adult Probation</td>
<td>PRCS violations</td>
<td>Contains all violations for PRCS clients, including charge and disposition.</td>
</tr>
<tr>
<td>ADP-0014</td>
<td>Adult Probation</td>
<td>PRCS warrants</td>
<td>Contains all warrants issued for PRCS clients, including charging and disposition.</td>
</tr>
<tr>
<td>ADP-0015</td>
<td>Adult Probation</td>
<td>1170 rearrest spreadsheet</td>
<td>Contains all arrests of 1170 clients, including charge and disposition.</td>
</tr>
<tr>
<td>ADP-0016</td>
<td>Adult Probation</td>
<td>Reentry pod class enrollments</td>
<td>Contains all class enrollment information for participants.</td>
</tr>
<tr>
<td>ADP-0017</td>
<td>Adult Probation</td>
<td>Reentry pod participants</td>
<td>Contains information regarding participants in the Reentry Program.</td>
</tr>
</tbody>
</table>
Welcome to this San Francisco Data API, powered by Socrata
For more information on how to access this dataset via the Socrata Open Data API, read on. Additional information on query functionality, code samples, and libraries can be accessed using the navigation above. Happy hacking!

Dataset Inventory

⚠️ Heads up! A new, improved version of this API is available for your use. Go check it out, or read on for details on how to migrate.

Getting Started

All communication with the API is done through HTTPS, and errors are communicated through HTTP response codes. Available response types include JSON, XML, and CSV, which are selectable by the "extension" (.json, etc.) on the API endpoint or through content-negotiation with HTTP Accepts headers.

Following RESTful paradigms, we return standard HTTP Status Codes to denote success or failure. For example, a successful request will get a 200 OK, while a bad query will get a 400 Bad Request. Errors are also accompanied by an error response including human-readable error messages.

This documentation also includes inline, runnable examples. Click on any link that contains a gear symbol next to it to run that example live against the Dataset Inventory API. If you just want to grab the API endpoint and go, you’ll find it below.

About this dataset

The dataset inventory provides a list of data maintained by departments that are candidates for open data publishing or have already been published and is collected in accordance with Chapter 22D of the Administrative Code. The inventory will be used in conjunction with

Show more ▼

Total Rows: 657
Source Domain: data.sfgov.org
Created: 3/26/2015, 11:22:56 AM
Last Updated: 5/14/2015, 5:44:45 PM
Knowledge ACCESS Ability
Welcome to SF OpenData! SF OpenData is the central clearinghouse for data published by the City and County of San Francisco and is part of the broader open data program, DataSF. Explore, view, and download our data. Developers - check out our developer page for tips on API access and use. Read more on our About page.
Task: Obtain Data

Hey! You got this data?

Maybe...

Hello ???

Can I have it?

Here it is...

Thanks! Some ?s...

Task: Update Data

See previous
Publication Automation

Extract, Transform & Load
The Balancing Act

Privacy

Transparency & Open Data
# Data Inventory and Raw Classification

## Raw Classification

<table>
<thead>
<tr>
<th>Raw Classification</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public</td>
<td>This data could be publicly disseminated without any concerns</td>
</tr>
<tr>
<td>Protected</td>
<td>This data is protected by law or regulation and can only be shared or accessed internally and per organizational procedures; OR this information includes individually identified information</td>
</tr>
<tr>
<td>Sensitive</td>
<td>In its raw form, this data poses security concerns, could be misused to target individuals or poses other concerns.</td>
</tr>
</tbody>
</table>
## Survey of Applicable Laws

### Table of Contents

- Private Information: Health and Human Services .................................................. 4
- Federal law: Health Insurance Portability and Accountability Act (HIPAA) ....... 5
- Box 1: HIPAA Enforcement ....................................................................................... 7
- CA law: Confidentiality of Medical Information Act (CMIA) ............................... 8
- Box 2: Reconciling HIPAA and the CMIA – when HIPAA governs ...................... 10
- Federal law: substance abuse records ..................................................................... 12
- Box 3: Reconciling HIPAA’s Privacy Rule and Part 2 regulations .......................... 13
- CA law: restrictions on substance abuse records .................................................... 14
- CA law: restrictions on mental health records ......................................................... 15
- CA law: restrictions on psychotherapy notes .......................................................... 16
- CA law: restrictions on sensitive data about HIV/AIDS ......................................... 19
- Box 4: Reconciling HIPAA and CA laws, where CA laws are more stringent ...... 19
- CA and local law: social services privacy laws ....................................................... 21
- Local law : identity of 911 call takers ...................................................................... 33
- Local law: confidentiality of City employee medical information .......................... 26
- Other HHS-related confidentiality regulations: Federal rules ......................... 27
- Private Information: Education ................................................................................ 30

### Raw Classification | Description
--- | ---
Public | This data could be publicly disseminated without any concerns
**Protected** | This data is protected by law or regulation and can only be shared or accessed internally and per organizational procedures; OR this information includes individually identified information
Sensitive | In its raw form, this data poses security concerns, could be misused to target individuals or poses other concerns.
More Process & Control Points

Process: Prepublication
Scope: All datasets

Fast releases

Thorough process

Existing automation? Yes

Meets criteria to maintain ETL? Yes

Review on case by case basis to determine migration to Hybrid model

What should be our criteria:
- ETL quality checks
- Staff/sourcing to support
- Well documented

Predefined PROCESS ETL Creation

No

No
Defining your methods

<table>
<thead>
<tr>
<th>Method</th>
<th>What it is</th>
<th>Best for</th>
</tr>
</thead>
<tbody>
<tr>
<td>De-identification:</td>
<td>The simplest way to avoid any privacy issues, is to simply not publish the columns that include private data. For example, if a dataset is a list of users and includes their name, address or other information, you can simply remove those columns from the dataset.</td>
<td>Datasets that include individual information that is not necessary for consuming and understanding the data.</td>
</tr>
<tr>
<td>Aggregation:</td>
<td>Summarize the data based on a data property.</td>
<td>Datasets where the individual records are aggregated either by geography or some other factor such as a category in the dataset.</td>
</tr>
<tr>
<td>Obfuscation:</td>
<td>Obfuscation can happen in a number of ways but a common case is with address data. Sometimes we want to retain a proxy of the address without aggregating the data.</td>
<td></td>
</tr>
</tbody>
</table>

Consistent technical implementation
Joy’s Holiday Wishlist

★ Plug and play regardless of backend
★ Effortless probabilistic matching with expert curation
★ Role based access control down to the column level
★ Legally aware

(cheap too, please!! ;-)

©TheHolidaySpot.com
Knowledge Access ABILITY
Within an EcoSystem of Shared Data and Technology

1. Increase individual skills & capacity
2. Support analytics in departments
3. Foster a data-enabled policy environment

Which leads to...

Increased use of data in decision-making
Data Academy

A training program for City and County of San Francisco staff

Welcome to the City and County of San Francisco’s Data Academy! Our goal is to help analysts citywide explore, refine, and enhance their skills in data analysis and visualization. We also offer classes for those hoping to increase their comfort working with data. We offer a series of classroom based courses detailed below and have also curated a list of courses offered by other institutions for free or at cost.

Intro to Tableau

Tableau is a data visualization tool that allows for fast, drag-and-drop creation of charts, maps, and dashboards that provide interactive filter, highlight, and drill down capabilities. This training will give you basic tools to start using Tableau productively in your department.

Intermediate Tableau One

This training is intended for beginner to intermediate Tableau users or someone who took our Introduction to Tableau course. We will cover dual-axis charts, heat maps, histograms, bullet graphs, Gantt charts, Pareto diagrams, sparklines, as well as other standard charts.

The Data Academy is produced in partnership with the Controller’s Office.

Sign up to learn more

You must sign up with an email from an official SFGov email domain.
Schedule friendly
Demand driven
Free ;-}
What...does the city do about affordable housing?

Where...do you go for reliable data about affordable housing?

The answers are the problem...

The solution...Housing Data Hub

Single place:
- Housing overview
- Description of policies and programs
- Supporting data & viz
enduring sources of truth & unbroken data lineages
Strategic releases help democratize how governments disseminate their data for a local audience that may be focused on issues (not just apps).
Questions?

THANK YOU
@datasf | datasf.org | datasf.org/blog