Austin, TX Open Data Value Research Project

# Summary and Recommendations

## Introduction

Over a series of sprints in 2017, the City of Austin conducted a user research study into the needs of open data portal users. One of the top challenges that came out of that study, and prioritized by Austin's open data community, was the need for publishers to be able to measure the value their datasets deliver to open data consumers.

### Why is this important?

Open data publishers are often asked to publish more datasets while also improving the quality of their department's existing data on the open data portal. However, publishers are not receiving feedback from open data consumers on the utility and usefulness of their department's published data and the priority of their unpublished data. Publishing open data is often an extra task outside of staff's regularly assigned workload, and as a result, some publishers feel unmotivated to publish more data and improve what's already available because they don't see the value in it.

### How we did it

In September 2017, the *Open Data Value Team* was formed to identify and implement recommendations to address this challenge. Team members included: Shannon Halbrook, Paige Warner, Stefan Wray, Wei San Hui and Sara Smith.

In our first 45 day sprint, we performed a literature review on previous research conducted on the topic of open data value, reviewed and documented the site analytics currently available about City of Austin open data, and prepared a research plan for for the next sprint.

In our second sprint, we focused our attention on conducting Austin open data user research. We created two surveys, one for Austin open data publishers and the other for open data consumers. Our target audience for open data consumers included researchers, journalists, businesses, developers, and residents. We also conducted in-person and phone interviews with members from each target audience, including open data publishers and managers, as well as the City of San Francisco and the Sunlight Foundation to learn about the successes and challenges they've experienced or seen around increasing and communicating the value of open data.

In our third and final sprint, we performed statistical analysis on our quantitative survey data and synthesized our qualitative research findings from our user interviews and survey data. From there, we conducted several rounds of ideation workshops which resulted in a list of prioritized metrics/measurements and Lo-Fi prototypes to solve the challenges we uncovered in our research.

You can find our process and final report out slide deck online. We welcome feedback and encourage anyone interested in collaborating with us to reach out!

## What we discovered

Below are our key findings from our literature review and user research.

The literature review we conducted told us:

- Data quality, including adherence to standards, is integral to creating value because Value = surface qualities + public value.
- The current approach others use to measure the value of open data is through ad hoc and manual reporting and depends on users reaching out.
- Open data user groups include: citizens, business, researchers, developers, NGOs, and journalists.
- Value may be different to internal (i.e. City staff) and external users.
- Suggested metrics for measuring the value of open data include:
    - motivation for accessing data,
    - dataset views/downloads, and
    - application/visualization frequency.

However, after conducting our own user research through interviewing and surveying City of Austin open data users and peer programs, we discovered both alignment with and divergence from the literature review findings.

Below are our synthesized findings from our interview and survey data.
The * denotes level of priority/impact where *** = highest priority/greatest impact.

What would **increase the value** of open data?

- More context for both published and unpublished data ***
- User feedback
- Data visualization, geovisualization and dashboards
- Contact information for individual datasets
- A tool that better explains the metadata

What gives open data publishers **sense of accomplishment**?

- Seeing or hearing that others are using their data **
- Receiving words of appreciation or accolades regarding their data
- Being able to uncover important insights and constructing a narrative
- Knowing their data is accurate and trustworthy

What are characteristics of **high quality data**?

- Contains useful metadata **
- Follows standards *
- Includes a temporal component such as a change log or archive *
- Accurate
- Simple
- Trustworthy

What are characteristics of **high value data**?

- Ability to inform government decisions and policy ***
- Enables change and improvements ***
- A topic someone can relate to **
- Ability to answer a question or solve a problem **
- The data tells a story *
- Trustworthy
- Raw and unaggregated
- Visualized on a map or chart
- Makes government transparent
- Reduces # of public information requests
- Contributes to public good
- Current and frequently updated

Examples of **value derived** from open data

- Community engagement such as civic technology projects and academic research **
- Improved government operations, decision making and services *
- Increased collaboration across City departments and government agencies
- Predictive analysis

**Pain points** related to open data

- Not knowing what data consumers want **
- City of Austin resource constraints, most notably staff time *
- Silos across City departments and other government agencies *
- Getting City leadership buy-in and engagement *
- Publishing data with known or potential inaccuracy
- Determining whether to publish sensitive data
- Fear of data being misinterpreted
- Lack of data quality related to metadata, naming conventions and data that's too aggregated to use

**Helpful metrics** to measure open data value

- <u>Standard analytics:</u> number of data views, # repeat visitors, comparison of views to other departments, number of internal vs. external views, update frequency, number of downloads

- Interaction with consumers: customer satisfaction, qualitative survey data, # dataset requests
- Government efficiency: % decrease in PIRs, usefulness in policy making, % decrease in questions from public
- Data quality: related to breadth of content, consistency, accuracy and understandability
- Use cases: number of embeds, number of derived views, API usage, # of stories/case studies about a dataset

# Recommendations

Based on our findings, we generated and prioritized 96 challenge statements and 165 potential solutions. We then prioritized that list and developed 21 lo-fi prototypes for the top ideas. Below are our top priority recommendations to guide the City of Austin's Open Data Program towards delivering more high-value datasets to consumers, tracking more comprehensive dataset-level value metrics and establishing a feedback loop to communicate those value metrics with publishers.

The recommendations are organized into the following categories:

1. How to Track Usage & Measure Value
2. How to Prioritize Unpublished Data
3. How to Increase the Value of a Dataset
4. Ideas to Further Explore

The project type identifies dependencies such as a need for external expertise that may include direct support from Socrata or contributions from community members such as our local Code for America brigade.

# How to Track Usage & Measure Value

## Open Data Partnership with Universities ⟳ In Progress

| | |
|---|---|
| Project Type: | Internal |
| Impact: | High |
| Level of Effort: | Medium |

**What it is?**

We know that some colleges and universities use City of Austin open data for both research and class projects; however, we don't have a way to track this use and communicate the datasets' value back to the data publishers. We also don't have a formal way to collect data needs from colleges and universities. The goal of this project is to create a formal data request and prioritization process with the university and college community.

An added benefit to this would be the insights and outcomes from these projects that could improve City services and inform City policy and decision making. Ideally, students and

researchers would present their work to City leadership and data owners at the end of each semester.

**Who leads? Who implements?**

The City of Austin Open Data Team.

## Citation/Attribution Process for Datasets ⟳ In Progress

| | |
|---|---|
| Project Type: | Internal |
| Impact: | High |
| Level of Effort: | Medium |

**What it is?**

Citations and attribution for data on the data portal are valuable because they allow us to more easily track published use cases. They clarify information so that users can learn about context, development, and quality, so requesting that people cite or attribute the data would be beneficial to both internal and external stakeholders.

**Who leads? Who implements?**

This item lends itself well to a sprint with the Open Data Team. The sprint team would create/provide an example citation/attribution and DOI (digital object identifier) to users and note that they are appropriate but not required.

**Risks / impediments**

The Law Department might need to be involved if the data licensing we currently have established needs to change.

## Dataset Analytics Partnership with Socrata ⟳ In Progress

| | |
|---|---|
| Project Type: | External |
| Impact: | High |
| Level of Effort: | Medium |

**What it is?**

The Open Data Value Sprint identified several metrics that would help to better understand how people are consuming datasets on the City's data portal. Help from Socrata is needed in order to incorporate these new ideas for metrics as well as to improve upon existing metrics. The Socrata Connect conference in Austin in May is an opportunity for us to meet with Socrata staff to better understand how Socrata analytics function and to express our ideas for new metrics and new metric features that would be helpful. There also appears to be discrepancies between what Google Analytics tells us and what Socrata Analytics tells us. It would be useful to better understand that aspect.

**Who leads? Who implements?**

The Measuring Value Sprint team can lead, but may need assistance making the connections with appropriate Socrata staff.

## Socrata Open Data Application (SODA) Index ⟳ In Progress

| | |
|---|---|
| Project Type: | External |
| Impact: | High |
| Level of Effort: | Medium |

**What it is?**

Application developers currently register their applications with Socrata and receive an App Token. And for that Socrata maintains a record of all App Tokens and the associated Application name and other information.

If Socrata were to identify dataset URLs used in association with the App Tokens, then Socrata would be able to join their App Token registration records with dataset URLs and generate a new table with records showing which datasets are associated with APIs.

In this way Socrata would be able to create a new dataset with the following fields:

- Application Name
- Description
- Organization
- Website
- Date App Token Registered
- Dataset URL
- Portal Location (i.e. City, County, State)

Socrata would then be able to have a dashboard with aggregate information for all the data portals it has as customers.

- Total # App Tokens in use
- Avg # App Tokens per Portal
- Total # Applications connected to App Tokens (probably same number)
- Avg # Applications per Portal

With this it would be possible to gain insight into the total universe of applications connected to and using datasets on Socrata platforms.

This would be useful to analyze and characterize datasets that lend themselves to being more apt to be used by developers.

Austin could compare its App Token and Application numbers to the complete set of all information about this to see how we compare to the SODA Index.

**Who leads? Who implements?**

Socrata would need to lead. There needs to be internal buy-in.

## List of use cases and how to populate it ⟳ In Progress

| | |
|---|---|
| Project Type: | Internal initially; open for contributions by external |
| Impact: | High |
| Level of Effort: | Medium |

**What it is?**

This would be a running list of use cases for how individual datasets are used, open for contributions by COA staff and external stakeholders. Initially, the list could be updated manually, but the goal would be to have an automated process by which projects using a dataset are discovered and added to the list.

**Who leads? Who implements?**

The open data team might start the list and then post it in a Google Group or other prominent forum to invite external user contribution.

**Risks / impediments**

It could be difficult to get participation and track down uses and then manually maintain it until the process is automated (which is also difficult).

## How to get people to use the Star Rating on datasets

| | |
|---|---|
| Project Type: | Internal |
| Impact: | Medium |
| Level of Effort: | Low |

**What it is?**

The Star Rating is a feature already available on the Socrata platform however it is rarely used. The Star Rating has the potential to communicate the use value of a dataset to the data owner, the department open data liaison and consumers. Our recommendation includes research into how we might encourage consumers to start using the star rating as well as implementing marketing strategies.

**Who leads? Who implements?**

A City of Austin open data sprint team.

## Value Calculator for Published Datasets

| | |
|---|---|
| Project Type: | Internal |
| Impact: | Medium |
| Level of Effort: | Low |

**What it is?**

The value calculator for published datasets would be used to evaluate datasets after the fact, i.e. after the dataset is published and how it's doing in terms of generating value for consumers. It's a good way for publishers to validate initial predictions for unpublished datasets. After time, we should be able to compare results from the two calculators (for published and unpublished data), and improve our evaluation criteria and predictions for unpublished data.

**Who leads? Who implements?**

This would be done at the departmental level by each individual data liaison or publisher.

**Risks / impediments**

It's time consuming and requires cooperation from all city departments. But we will be rewarded with a potentially higher usage of our datasets, especially the ones that we have vetted using this scoring system. Also, it will take us a couple of trial runs to refine our scoring criteria and improve on its accuracy.

## Dashboard Creation and Integration for Dataset Metrics

| | |
|---|---|
| Project Type: | External |
| Impact: | Medium |
| Level of Effort: | High |

**What it is?**

This dashboard would pull in analytics from our proposed citation script, Socrata analytics, Google analytics, PIR analytics, metadata analytics and proposed university use tracking.

**Who leads? Who implements?**

A civic technology group such as Open Austin would be ideal for leading this effort or a consultant with the necessary technical expertise.

## Google Analytics ⟳ In Progress

| | |
|---|---|
| Project Type: | Internal |
| Impact: | Low |
| Level of Effort: | High |

**What it is?**

Google Analytics allow us to capture information on user behavior, both at the site level and dataset level. For the Open Data Portal as a whole, it provides information about the audience, demographics, behavior, and acquisition patterns. For individual datasets, it gives analytics such as the page title, page views, unique page views, average time on page, entrances, bounce rate, and the percent exit. It also provides these metrics in terms of content categories.

**Who leads? Who Implements?**

City of Austin staff.

## Dataset User Survey

| | |
|---|---|
| Project Type: | External |
| Impact: | Low |
| Level of Effort: | High |

**What is it?**

This solution would collect use case information from consumers by asking them to choose from a pre-populated list of uses of what they plan to use the data for. This would be a pop-up window that appears randomly/on occasion to collect a sample size from our consumer user base.

**Who leads? Who implements?**

This solution would require Socrata to build and implement this survey tool.

# How to Prioritize Unpublished Data

## Advisory Board with Community Members

| | |
|---|---|
| Project Type: | Internal |
| Impact: | High |
| Level of Effort: | Medium |

**What it is?**

An advisory board, comprised of community members, would be formed to review the research conducted by the Measuring Value Sprint team, along with the key findings and recommendations, and would then provide additional insight into what constitutes valuable data.

**Who leads? Who implements?**

This would be initiated by the Open Data team, with the hope that groups like Open Austin would take leadership roles in sustaining.

## Unpublished dataset list for consumers to vote on and prioritize

| | |
|---|---|
| Project Type: | Internal |
| Impact: | High |
| Level of Effort: | High |

**What it is?**

A list of all city datasets that are not published on the open data portal. The goal would be to make the list available to all stakeholders to solicit community feedback as to which datasets should be prioritized for publication.

**Who leads? Who implements?**

A City of Austin open data sprint team.

**Risks / impediments**
This is a time-consuming and unfamiliar process for data owners and getting buy-in and participation from staff will be difficult.

## Require Staff to Publish Data Presented at Council Meetings

| | |
|---|---|
| Project Type: | Internal |
| Impact: | High |
| Level of Effort: | High |

**What it is?**

City staff regularly deliver presentations to City Council. Often, those presentations contain visualizations of data or some type of data product. Any data products presented at council meetings can, by default, be said to have a public interest. Such data products have already been vetted and are available for public consumption. Therefore, the underlying raw data used to create these data products are candidates for the open data portal.

The proposal is that if a City of Austin staff person plans to deliver a presentation to council that includes a data product, then a data steward responsible for the underlying raw data would post that data to the portal. The ideal would be that the backup material associated with the respective agenda item would include a link to the published dataset. This may not, however, always be practical to have a posted dataset prior to the council agenda.

An alternative to this idea, and one that could be a way to get started, would be to establish a way to capture details and keep a log of data products presented at council that would include pertinent related information.

**Who leads? Who implements?**

A City of Austin open data sprint team could lead, however we would need buy-in and support from the Office of the City Clerk. This could also be an initiative supported the City Manager's office and/or Council.

**Risks / impediments**

The risk is that neither the City Clerk nor anyone at City Hall is interested in adding another layer of work on top of what already exists.

## Increase dataset request voting and improve engagement

| | |
|---|---|
| Project Type: | Internal |
| Impact: | Medium |
| Level of Effort: | Low |

**What it is?**

Currently dataset requests can be submitted, viewed on the data portal, and voted on. However the feature is infrequently used. We want to encourage users to submit and vote as much as possible to gather information about what data the community is interested in.

**Who leads? Who implements?**

A City of Austin open data sprint team.

## Value Calculator for Unpublished Datasets

| | |
|---|---|
| Project Type: | Internal |
| Impact: | Medium |
| Level of Effort: | Low |

**What it is?**

This is an aggregate scoring mechanism that we came up with that takes into account several key value metrics like number of PIR requests, number of vote or request from customers, any linkage to current event or strategic outcome etc. Each of these criteria will generate a score that will be added up to give the overall score. The higher the number, the more priority should be given to try and get this dataset out the door and published.

**Who leads? Who Implements?**

This has to be done at the individual dataset level where each department liaison would have to go through the exercise of using the value calculator to project the potential value of a dataset and use that to help in their final decision-making process (of whether to publish or put in the effort to collect the data)

**Risks / Impediments**
It's time consuming and requires cooperation from all city departments. It will also take several iterations to refine our scoring criteria and improve on its accuracy.

# How to Increase the Value of a Dataset

## Austin Open Data Forum

| | |
|---|---|
| Project Type: | Internal |
| Impact: | High |
| Level of Effort: | Low |

**What it is?**

A more informal forum, such as a Google Group, provides a place for people to post dataset requests, update requests, and ask questions. This forum would help improve communication with consumers and provide a platform for a community around open data.

**Who leads? Who implements?**

A City of Austin open data sprint team.

**Risks / impediments**

We already have communication platforms such as Bloomfire and the Socrata Portal; however, they are not heavily used. Research prior to implementation will be necessary to see why these current platforms do not have high engagement.

## Dataset Contact Information Updates and Culture Change

| | |
|---|---|
| Project Type: | Internal |
| Impact: | Medium |
| Level of Effort: | Medium |

**What it is?**

Contact info is very important because the city's responsiveness to users about a given dataset adds to its value. Sometimes datasets' contact info is not kept up-to-date when city staff leave or responsibilities change. We want to encourage city staff to review this information and help ensure it is current.

**Who leads? Who implements?**

A City of Austin open data sprint team.

**Risks / impediments**

It is difficult to change behavior to encourage a small, seemingly innocuous task that can easily fall off the radar. But contact info only needs to be changed when the responsibility has been reassigned.

## Life of a Dataset

| | |
|---|---|
| Project Type: | External |
| Impact: | Medium |
| Level of Effort: | High |

**What is it?**

This is a one-stop shop to get all that you need or ever want to know about a dataset quickly without going to 10 different places. It includes origin of the data (how it's collected), date it was created, who created it, any user comments posted to date, voting history, links to use cases, PIR requests, research papers, articles, presentation in conferences.

**Who leads? Who implements?**

A City of Austin open data sprint team can lead but support from Socrata or a consultant will be required.

## Marketing ⟳ In Progress

| | |
|---|---|
| Project Type: | Internal |
| Impact: | Low |
| Level of Effort: | Medium |

**What it is?**

The City of Austin needs a way to reach diverse users and promote the use of the City's open data portal. Many people, even within the City, do not know about the open data program. However, with increased awareness and excitement surrounding the program, these people could become valuable producers or consumers of our data. Marketing activities could include social events, improving the brand, community outreach, and internal communication and collaboration to strengthen buy-in and increase participation.

**Who leads? Who implements?**

We envision this will eventually evolve into a role or incorporated into someone's duties and responsibilities either in the Open Data Team or Public Information Office.

**Risks / impediments**

This might require budgeting for a new FTE.

# Ideas to Further Explore

## Google Tags

| | |
|---|---|
| Project Type: | External |
| Impact: | Low |
| Level of Effort: | High |

**What is it?**

Tags are snippets of code that are inserted into a site's source code to send data from that site to a third party site -- in this case, Google Analytics. It provides the functionality of Google Analytics on a more specific level than for pages.

**Who leads? Who implements?**

The use of Google Tags would require someone to set up tags for each dataset, including new ones as they arise. Someone would also have to regularly check and interpret the analytics.

**Risks / impediments**

Socrata does not currently support the use of Google Tags in the Socrata platform.