Open Data Alliance

GUIDEBOOK

Leslie Denning & Kevin McCraney
OPEN DATA LITERACY PROJECT | UNIVERSITY OF WASHINGTON

OPEN DATA ALLIANCE GUIDEBOOK: EXECUTIVE SUMMARY

The Open Data Alliance Guidebook was formulated to support and formalize a coalition of professionals concerned with Open Data projects and focuses on Public Records Requests as a commonality across jurisdictions, divided into five (5) parts:

Agenda's and Guide

- To support the timing and topics of the first formal meeting and activities as well as a guide for future meetings.

Interview Consolidation: A Thematic Distillation

- To establish a consensus about Public Records Requests and Open Data for the alliance.

Open Data Alliance Principles

- To ensure all members have the same approach regarding open data projects.

Identifying Valuable Datasets for Open Data

- Tools for identifying Public Records for Open Data publication.

Growth & Outreach

- Tutorials and guides that promote community engagement and data literacy.

OPEN DATA ALLIANCE AGENDA

Date: TBD (October 16th, 17th, or 18th?)
Time: TBD (Depends on the ACCIS Schedule)
Facilitator: TBD (Will Saunders or Sean Krier)

Time	Item	Speaker
	Introductions & the Open Data Literacy Project	
5mins	Introduction from Kevin & Leslie at the Open Data Literacy Project	Kevin & Leslie
40mins	Presentation and Q&A – Open Data Alliance Summer Internship	Kevin & Leslie
20mins	Discussion: Member introduction & expectations for the Alliance	
10mins	Break and Unconference set-up	
	Mini Unconference – Open Data Topics – 1 hour	
5mins	Explanation of Unconference instructions	
15mins	Create session topics that are focused on open data	
10mins	Organize session topics and assign participants to lead	
30mins	Begin first session topic and discuss (suggested group discussion)	
5mins	Break and Unconference tear-down	
	Wrap Up – 15 minutes	
10mins	Schedule next meeting and identify topic for the next session -- post materials to GitHub	
5mins	Final words – Reminder about GitHub resources	

OPEN DATA ALLIANCE AGENDA

Date: TBD
Time: TBD
Facilitator: TBD

Time	Item	Speaker
	Reconvene & Introductions of New Members	
5mins	Recap of last meeting and introduction of new members	
10mins	Quick discussion of any topics unrelated to the topic sessions of the day	
	Open Data Topic Sessions – Learning and Discussion	
3mins	Reiterate the foundations of useful discussion behaviors for new and returning members	
45mins	Begin first session topic (show and tell, group discussion, or learn how to do X)	
5mins	Break	
45mins	Begin second session topic (show and tell, group discussion, or learn how to do X)	
5mins	Break	
	Wrap Up	
10mins	Discussion of the day's events and suggestions for steps forward	
10mins	Schedule next meeting and identify topics for the next session – post materials to GitHub	
5mins	Final words – Reminder about GitHub resources	

OPEN DATA ALLIANCE MEETING GUIDE

Overall Meeting Guide

Open Data Alliance meetings are an open space for collaboration and learning. To be successfully open, all members must be willing to share their thoughts and ideas without fear of embarrassment or repercussions outside of this meeting. All members must respect the time of other members and to be prepared with what is expected at each meeting.

Administrative Reminders:

Sign-ups for future meetings and spots to put topics of the mini-unconference sections are available on GitHub to which all members have access and can edit.

Be sure to look at the "Growth & Outreach" document for resources to help with questions about open data, sources for learning or supplementing technical skills, and opportunities for community engagement activities.

Roster of members for outreach are on the ODA GitHub (OpenDataAlliance2018) and updateable for future communications.

Mini Unconference Instructions

Materials:

- Note cards (or pieces of paper)
- Markers (any writing instrument)
- Camera or notetaker (for taking pictures of the cards/recording the sessions/notes)

Instructions:

Each person will be given several cards and will write out open data related topics that could be up for discussion and identified as possible projects. Next, provide a space where the cards created can be arranged and consolidated if necessary.

At this point, choose the topic for the first session (suggested as a group discussion) and the facilitator. You may also plan the next few meetings with several topics for each meeting and assign facilitators for each session. It is recommended, depending on the topics, that the sessions differ in their delivery. For example, a show and tell session, a group discussion, or a demonstration of a tool and then an interactive learning session with the tool.

If at any point you become aware of any outside resources that could be useful to your topics, allow time for a small presentation or demonstration of their information and use that to build on throughout the session. Be sure to make any new resources available on GitHub.

INTERVIEW CONSOLIDATION: A THEMATIC DISTILLATION

Introduction

As our interviews crossed many government agencies and jurisdictions, we encountered several distinct perspectives from people concerned with public records and open data. Whether they were Public Disclosure Officers or open data officials, our interviewees helped us identify themes that cut across all agencies and jurisdictions.

Overall, organizations are hesitant to use public records as a source for proactive publication of open data. The tendency to be “risk averse” is a relevant tension that must be considered when undertaking an open data project. However, the participants did acknowledge that if such a proactive publication occurred, it could be possible that the frequency of public records requests would lessen and in turn, help reduce the cost of fulfilling such requests, whether that cost stems from FTE or potential risk.

We asked our interviewee's several questions about the possibility of participating in a regional Open Data Alliance. It is apparent that there is an overwhelming desire for a collaborative environment, an opportunity to better understand what each department or jurisdiction is doing related to open data and sharing best practices along the way that reinforces such a collaboration.

Themes

Public Records & Open Data

Skepticism about public engagement.

There is some skepticism about whether the public will engage with published datasets on an open data portal and if they are knowledgeable enough to work with the data in the form provided. Although this skepticism is valid, there is an inherent underestimation of the curiosity and skills among the public.

Data is not intrinsically useful by itself—one must engage with it to produce structured information, and later insight. There will always be some students or data scientists, hobbyist or community activists, that will find a way to use the datasets. For the parts of the public who might not know that such data is available or do not know how to work with the data directly, they can find aid with the help of their public library if they offer such services. However, this concern is equally an opportunity to educate the public and provide resources to support their initial curiosity.

Insights: Create partnerships between libraries in order to foster awareness about Open Data/Open Access materials produced by government.

INTERVIEW CONSOLIDATION: A THEMATIC DISTILLATION

Institutional embarrassment.

There is a hesitation to make particular datasets available because of the possibility that the data will not reflect well on the institution from which it came, and that "institutional embarrassment" will invite further public scrutiny for the institution. There is a constant drive to control the narrative of any situation, and this is heightened when dealing with projects that have a government-public interaction.

Potential embarrassment derived from the data, or even the intentional misrepresentation of the data is a possible risk associated with open data. An institution can preempt this by providing sufficient metadata to the datasets when publishing in the form of a data dictionary, explaining the source, collection methods, and so on. Understanding the context of the data collection can help mitigate the risks involved with potentially embarrassing insights.

Insights: Institutional embarrassment is a big barrier to open data. Focus on ways to alleviate these feelings by highlighting the good that open data can do.

Directing requesters to a dataset that is not often updated.

One apprehension about publishing frequently requested public records as datasets originate from a scenario where a Public Disclosure Officer instructs the public to the dataset in place of a formal request. If the data set is not updated regularly, there could be an issue where the requester did not receive all the documents available, and the organization would be liable legally. As such, it is crucial to make it clear how often the datasets are updated and by who and is integral to assuaging fears about such an issue.

Insights: If directing requesters to datasets in lieu of traditional public records request fulfillment, be sure that the data has a clear description of what it is and when it is updated.

Risks of inadvertently releasing personally identifiable information.

The mistake of publishing a dataset with personally identifiable information is of great concern. In particular, a case where one is automatically publishing a dataset, and there is not the opportunity to catch such an error until it is too late. Although the purely accidental release of personally identifiable information is possible, the occurrence of such an error is typically preventable by the department's privacy policies in combination with the public records officers who understand the exempt types of information.

One other possible way to prevent such an issue and to allow for the automatic publication of data comes from taking advantage of newly implemented data systems that replace legacy systems. For example, if a data collecting system in the office is receiving an overhaul or a rebuild, there can be a

moment of opportunity to restructure it in a way that will allow for a simple extraction of data that avoids all personal information.

Insights: When implementing or rebuilding systems that collect data, be sure to focus on ways that allow for easy data extraction and avoid tedious redaction.

Understanding the stewardship of legacy datasets.

In a government office, datasets are manifold. They are kept in many formats, housed in many locations, and are updated by employees, but it is not clear who owns them. For datasets to proactively publish or be published and continually updated, there must be an understanding of who owns the data and who is the person updating. While individual departments have a better understanding of their data stewards, others might not be on the same level. It can sometimes be common to find data for possible publication only to find that the person compiling the information has left or retired. There is still an opportunity to begin the stewardship anew and to attempt to discern the information as thoroughly as possible by compiling extensive metadata before publishing to an open data portal.

Insights: Tracking down data owners and data creators can be time consuming to do retroactively. Start from one point and continue to document so that future people working with or looking for the data will have some inkling of where to start.

Creating an uptick in requests.

There is a concern about the possible increase of public records requests when making other public records or datasets available as open data. The concern is rooted in the idea that such an uptick will only add to the increasing costs related to fulfilling public records requests, and that open data has not solved this issue but instead made it worse. Such a possible consequence of publishing open data is entirely new in our digital age because of the quick and accessible nature of the internet to the public and those concerned with government data.

It is difficult to determine whether an increase in public records requests is caused by the publication of data or with the implementation of a new open data portal. However, average increases in public records request often arise from incidences outside the department, such as cases when there is a police officer involved shooting and a subsequent increase in requests for their body-cam footage.

Insights: Increases in requests have occurred and will continue to occur regardless of open data publication. It is important to focus on the various datasets that can be published and their respective drops in requests, instead of the general overall public records request numbers.

INTERVIEW CONSOLIDATION: A THEMATIC DISTILLATION

Open Data Alliance

Creating a community of sharing around open data.

One hope for the Open Data Alliance, expressed by many, is the creation of a regional community that focuses on open data that isn't directly associated with the government but is made up of government workers. To really solidify the communal aspect of such an alliance, the voicing of opinions should be embraced by all members. Understanding that almost each person in this organization is familiar with the difficulties of publishing open data and utilizing that commonality as a basis for such a community will allow it flourish.

Insights: To be a successful organization, members must be open and accepting of all ideas.

Standardizing data to ease sharing.

Informal sharing and aggregation of data across departments and jurisdictions can be helpful in gaining insights, but the biggest issue is the standardization of this data. Unstructured, unclean data means more time spent preparing the data for analysis and takes up precious collaborative time if not addressed. Finding ways as a group to agree on an informal standardization would help ease sharing among the members of the Open Data Alliance.

Insights: Standardization takes time to create initially but is incredibly useful long-term.

A support system to learn/teach ways to work with data.

Taking advantage of the skills and education of the group would be extremely beneficial to ease the sharing of data and could result in people bringing good ideas back to their parent organization and affecting positive change.

The Open Data Alliance, for many interviewed, seemed an excellent way to support people who are working on a project related to open data or are interested in advocating for its use in their department. Sharing skills through peer teaching could create a support system for people new to the field. The goal of the alliance is to provide opportunities where one can be vulnerable enough to learn and strong enough to teach.

Insights: Utilize the knowledge and skills of all members and be sure to support members who are learning skills for the first time.

Open Data Alliance

Our Mission The Open Data Alliance: data-driven, human-focused.

The five pillars of open data. Open data is...

Single origin It is as close to the source of the supply chain as possible, not composed of aggregated or modified forms.

Minimally refined All public aspects of the data are made available when possible—the default is open.

Structured and annotated It is easily parsable with computer software, the metadata are well-defined and specified, and the features are clearly explained

Free, gratis, and libre It is provided with no monetary compensation or barriers to entry, and that it is portable to a wide variety of formats, non-proprietary, license-free, etc.

User-focused and customer-centric It is initially produced and structured through public dialogue, to address the needs of citizens, and is accessible to the widest array of users for a panoply of purposes.

As a member of this organization and to address the five pillars of open data, you should...

Be open about being open Communicate with internal and external bodies to champion the open data related work your organization is doing. Share your successful processes and innovations, but also failures and obstacles.

Maintain common standards Rather than rely on institutional knowledge, attempt to codify and collect whatever you can in a lasting, sustainable format, such that new entrants into your discipline understand the intricacies of both your data and the processes that produce it.

Test your own products Regularly work with the data you provide to external parties in the form that they receive it, to understand the pleasures and pitfalls of the information you are sharing. Audit your material regularly and ensure that it is regularly updated.

Investigate possible risks Before releasing a dataset or constructing a new system, attempt to think from the perspective of someone who would use the data for nefarious purposes, or the circumstances under which things could go wrong. Act accordingly to prevent abuse and mitigate risk.

Open other things up Remember that the default should be open; as legacy systems are spun down, try to find other avenues to proactively disclose information, using the principles above.

IDENTIFYING VALUABLE DATASETS FOR OPEN DATA

Introduction

This document is for the members of the Open Data Alliance to help with identifying valuable datasets for possible publication as open data. Included are general categories from the Washington State Office of the Chief Information Officer, popular categories identified by a study at the Sunlight Foundation, categories identified by our analysis of public records requests from the cities of Seattle, Port Orchard, and Olympia, as well as a set of recommendations to identifying valuable datasets.

Steps Needed for Identifying Valuable Datasets

Know Your Users – *Connect with the public*

Although connecting with the public may seem obvious and you might be doing so already, there may be other tactics to explore. However, one can only derive so much information on trends from public records requests. It is vital that as open data projects begin or continue, a connection and deep understanding of the needs of the public is clear and present when deciding datasets to choose for publication. Many citizens may not realize that their public records requests are information requests in truth, and it is these opportunities that provide space for possible valuable dataset identification. With such a community engagement tactic, as supported by resources in the Growth & Outreach document, any group or person will begin to identify the immediate needs of the public that could be addressed by specific datasets. As a side effect, there could also be the identification of a portion of the community that is not always addressed or viewed as being users of data.

Types of engagement: Focus Groups; Interviews; Surveys; Community Meetings; Public Forums

Know Other Users – *Look outside your area*

While knowing how to serve your community is the priority, there is no reason to "reinvent the wheel." Communities may have issues specific to their region and population, but often, the data most useful to them are similar across the board. Identifying open data projects or datasets that are successful or similar can provide opportunities for evidence that your project is worthwhile or even some inspiration for datasets. While in the search for other examples to help bolster the cause, it could also be an avenue for collaboration and resource sharing on a level that benefits both parties involved.

Areas to focus: Cities of similar sizes; Neighboring towns; Cities that lead Open Data Projects; Open Data Organizations

Know Your Data – *Understand the history*

There may be datasets already on the radar for possible publication that seem to be of no significance. In these situations, taking the mindset that each although it may not seem significant, someone out there (and there's always someone) may disagree and find a use for it. Such a situation is why the use of necessary metadata to provide context to the data is incredibly important for the continued use of open data portals and datasets. Giving metadata and a descriptive story about the datasets published will allow for a profound public interaction element.

Don't forget: Asking for complete metadata; Only publish if you can explain it; Make sure to always update

IDENTIFYING VALUABLE DATASETS FOR OPEN DATA

General Categories from WA OCIO[1]

- **Geographic reference data**: parcels, addresses (except category 3 and 4 data), where to obtain state services.
- **Public safety data**: traffic, moving violations, aggregate crime statistics, environmental hazards.
- **Fiscal data**: state salary data, expenditure data, budget data, purchasing data.
- **Health data**: quality and purity of water, food, and air, data that increases healthcare accountability, data that facilitates patient choice, aggregated incidence of diseases and medical conditions in communities (review for HIPAA compliance first).
- **Education data**: capabilities of state funded schools, achievements by the state's student population, results of state support of education.
- **Census and demographic data**: the populations of the communities of the state, trends in migration, diversity, and housing
- **Business and economic data**: aggregate changes in license counts, labor market and employment.
- **Government directory data**: who to contact for help with State services, catalogs of available state services.

The Sunlight Foundation[2]

Topics of the most popular open datasets

1		**Police and crime:** Police incidents; jail bookings; police station locations; crime statistics
2		**Transportation:** Taxi licenses; transit data; traffic counts; road infrastructure data; parking data
3		**Emergency calls:** Police, fire, and EMS responses; 911 calls; response times; incident reports
4		**Development:** Commercial developments; housing developments; property data; housing affordability
5		**Building safety:** Building permits; safety permits; certificates of occupancy
6		**Finance:** Revenue; spending; employee salaries; capital budgets; payments
7		**Elections:** Election results; polling locations; campaign finance reports
8		**Businesses and licenses:** Business licenses; liquor licenses; vendor, contract, and procurement data
9		**Inspections and service requests:** Restaurant health inspections; 311 requests; code violations
10		**Education:** Schools information; student health data; after-school programs; library locations

ODL Categories from Public Records Request Logs

Building and Construction

- Code Violations
  - Fire Code Violations
  - Environmental Assessments
- Property Records

Permitting

- Building Permits; City Project Permits

Police Department

- Police Officer Recordings (Audio/Video)
- Police Incident Reports
- Arrest Records

Transportation

- Automotive Accident Information

[1] Washington State Office of the Chief Information Officer. https://ocio.wa.gov/programs/open-data/guidance-what-data-publish
[2] The Sunlight Foundation. "Who's at the popular table? Our analysis found which open data the public likes." https://sunlightfoundation.com/2017/09/11/whos-at-the-popular-table-our-analysis-found-which-open-data-the-public-likes/

DATASETS OF INTEREST

During this project, we used statistical techniques to analyze public records requests. These processes mapped the frequency and relationship between words and helped us identify the popularity of particular themes over time.[1]

Broadly, the topics we discovered fell into the following categories:

Transportation

- Insurance Cases re: Car Accidents
- Parking Infractions
- Stolen Vehicles

Law Enforcement

- Police Officer Recordings (Audio & Video)
- Arrest Records
- Police Incidents

Architecture & Buildings

- Code Violations
- Environmental Assessments
- Construction Permits

Since you are concerned with open data within your department/jurisdiction, you have undoubtedly considered some (or all) of these categories as possible data sources to open up. Try evaluating the gains that might come from releasing data; barriers to doing so; and how you might mitigate risks related to disclosure.

[1] Visit https://github.com/OpenDataLiteracy/OpenDataAlliance-KM to examine the theoretical underpinnings of these processes and take a closer look at what we discovered.

GROWTH & OUTREACH RESOURCES

Introduction

This document is for the Open Data Alliance to guide growth within the alliance and outreach from the alliance to the public. The growth section will focus on resources for alliance members grouped by open data and by technical skills. The outreach section will focus on resources to help aid in community engagement and outreach.

Growth

Open Data

WA OCIO – Guidance: Agency Open Data Plans

- *Intended to guide the creation of an open data plan.*

WA OCIO – Guidance: What Data to Publish

- *Intended to guide the selection of data for publication as open data.*

City of Seattle Open Data Playbook

- *Originally used for Seattle's Open Data Champions – useful tips for selecting datasets.*

The Sunlight Foundation – Guidelines for Open Data Policies

- *Includes direction on what data to choose, how to make it public, and policy creation.*

The Sunlight Foundation – Public Policy for Public Data

- *Intended for initial policy creation regarding open data.*

U.S. City Open Data Census

- *Helpful for comparison to other cities and to understand what is published elsewhere.*

Open Government Data by Joshua Tauberer

- *A book about the principles, practices, and history of open government data.*

Open Data Institute – Guides

- *A curated set of guides that deal directly with all aspects of open data.*

Open Knowledge International – Open Data Handbook

- *A website dedicated to providing guides, value stories, and resources about open data.*

IRM: Aligning Supply and Demand for Better Data Governance

- *A paper with a focus on government and open data utilization.*

GROWTH & OUTREACH RESOURCES

Technical Skills

Resources for learning Python

LearnPython.org

- *Provides interactive Python tutorials for data science.*

Dive Into Python

- *A step-by-step guide for learning python.*

Automate the Boring Stuff

- *A guidebook for beginners learning how to code with python.*

Think Stats 2e

- *An introduction to probably and statistics for python users. Focuses on theory and principles.*

Google's Python Class

- *A class for people with a little programming experience who want to learn python.*

Software Carpentry

- *Tutorials designed for teaching R and Python.*

Conda

- *A scientific computing environment for Python or R.*

Resources for learning R

Data Camp – Introduction to R

- *An introductory course for learning R.*

Swirl

- *Provides interactive R tutorials for data science.*

Data Science Journal by Cecilia Lee

- *A blog dedicated to teaching people how to use R for programming.*

Tutorials Point – Learning R

- *A detailed course/tutorial for learning R.*

GROWTH & OUTREACH RESOURCES

Other Resources

NNIP – Beyond PDFs: Visualizing Your Data

- *Video presentation about visualizations and recommendations.*

Open Data in a Day

- *An introductory course focused on open data properties such as cleaning.*

Sunlight Academy

- *A plethora of interactive tutorials for working with government data.*

Outreach

The Sunlight Foundation & What Works Cities: A Guide to Tactical Data Engagement

- *A guide for government officials to work with the community regarding open data.*

The Sunlight Foundation: Tactical Data Engagement Playbook: Data User Groups

- *Similar to the TDE above but with a focus on user groups.*

National Neighborhood Indicators Partnership's Guide to Starting a Local Data Intermediary

- *For possible future use when funding becomes a factor. Focus on local usage.*

The Sunlight Foundation: Does your Open Data project match community needs?

- *A quiz for identifying the compatibility of an OD project and community needs.*

Western Pennsylvania Regional Data Center: Create Your Own Data User Guides

- *A template creating data user guides for your public users.*

Open Seattle

- *A Code for America Brigade that works on projects in the state of Washington.*

Democracy Lab

- *Provides contacts for tech-for-good projects in Seattle.*