

INFX 551 Data Curation  
February 13, 2017  
Jacob Kovacs  
Word count: 928, excluding tables

## Assessment of metadata for Seattle Police Department's 911 Incident Response and Police Report Incidents data

This is an assessment of the metadata for two datasets available through the City of Seattle's open data portal: 911 Incident Response data (SPD, 2017b) and Police Report Incidents data (SPD, 2017d).

### Description of metadata

Seattle's open data portal is powered by Socrata software. Socrata publishes metadata for datasets in an associated "primer" page (SPD, 2017a; SPD, 2017c). Screenshots of primer pages are provided in the Appendix, and Table 1 summarizes the metadata categories featured in the primers.

Metadata can also be viewed by examining the header of the corresponding downloadable JSON field; this second source contains a few more fields, but is less likely to be seen by most members of the public. For this reason, I've restrict my analysis to primer metadata.

**Table 1.** Description of provided metadata for 911 incident response data (SPD, 2017a) and police report incidents data (SPD, 2017c). In addition to these categories, Socrata allows arbitrary files to be attached to a dataset; for the datasets in question, relevant data release policies have been attached.

Name of dataset
Narrative description of dataset
Date that data was last updated
Date that metadata was last updated
Date that dataset was originally uploaded
Interval at which dataset is refreshed (new data added, if available)
List of dataset providers
Name and contact method for dataset owner
Attachments (see note above)
License
Category
Tags
Number of rows and columns
Number of views and downloads
Name, description, datatype, and API field name for each column

### Evaluation of metadata quality

Despite the seeming thoroughness of the categories outlined in Table 1, the metadata for these datasets is badly inadequate. As mentioned in my previous evaluation of this data, many of the user comments associated with the datasets are questions or complaints that stem from lack of metadata (Table 2).

**Table 2.** An illustrative sample of user comments, revealing the inadequacy of provided metadata (qtd. from SPD, 2017b and SPD, 2017d).

Unlike the records within Seattle_Real_Time_Fire_911_Calls, the datetimes here do not have the UTC offset. Should it be assumed it is 0 (UTC), or local?
The data from 9/2016 appears heavily duplicated except for the RMS CDW ID. [...] Does each record reflect a different reporting officer, a different suspect, or are these the result of duplications in the incoming feed?
Could anyone please give these labels some meaning or some codes. Such as what is Zone/Beat, or Census Tract, or the Offense Codes.
[...] Roughly 282K incidents is a fraction of the actual reports on file. Is there a reason why the SPD isn't loading all of the other data? For example, there are very few DUIs here, yet that's one of the most common.
Is there a reason that rape is not included in this data set?
Is there any metadata available for the explanation of variables.
Any chance we can get a codebook to explain what the variables identify (such as Event clearance date, etc.)

In short, the metadata is *not* comprehensive. The contextual and provenance metadata is not sufficient to answer—at the necessary level of granularity—who created the dataset; why and how they did so; and what sort of processing the data underwent prior to publishing. In her primer on open crime data, Lord (n.d.) notes that datasets derived from computer-aided dispatch (CAD) systems are more timely but less reliable than police reports; this is vital contextual information that absolutely should be made explicit in the metadata for these datasets.

In addition, the descriptive metadata is not sufficient to explain the encoding of datetimes, the meaning of administrative codes, or even the basic definition of different fields. A savvy user might be able to infer the contents of several fields based only on their names, but one of the primary reasons for publishing government data is to make it broadly accessible, which includes users who don't know anything about Census geographic units, policing geographic units, offense codes, etc.

### Evaluation of adherence to metadata standards

As a platform, it appears that Socrata neither imposes a metadata standard nor makes a serious effort to recommend compliance with one (for instance in their knowledge base article on "metadata best practices"; Socrata, 2016). The only metadata *required* for publishing a dataset is its title; dataset description, category, tags, licensing, data provider, source link, and row label are merely recommended (Socrata, 2017). My search of the Socrata Knowledge Base (<https://support.socrata.com/>) failed to uncover any suggestion that Socrata's recommended metadata is based on a metadata standard; however, Socrata does allow site administrators to specify custom metadata fields (Socrata, 2015), which permits adoption of broadly-used metadata standards or even import of local metadata standards.

The most attractive candidate for a metadata standard is the Project Open Data (POD) schema 1.1, used by the U.S. federal government for its data portal (Data.gov, n.d.). This standard seems fairly extensive; it is meant to be applied at the level of individual datasets or whole catalogs (POD, 2014, "Metadata schema"), and has crosswalks for conversion to several other metadata standards (POD, 2014, "Metadata resources"). Table 3 summarizes (to the best of my ability, as someone not trained in metadata) POD 1.1 standards for dataset-level metadata.

**Table 3.** A summary of the Project Open Data v1.1 metadata standards for dataset fields, omitting POD standards for catalog-level metadata and dataset distribution metadata. Mandatory elements are denoted with an asterisk (\*); definitions for each element are my personal interpretation of those provided by POD *unless* they are enclosed within quotes, in which case they are quoted from POD, 2014, “Metadata schema”.

<b>accessLevel*</b>	Describes whether data is intended to be publicly available; restricted to values `public`, `restricted public`, or `non-public`.
<b>bureauCode*</b>	Denotes federal agency or agencies responsible for dataset (mandatory only for U.S. federal agencies).
<b>fn*</b>	Full name for dataset point-of-contact.
<b>hasEmail*</b>	Email for dataset point-of-contact.
<b>description*</b>	Narrative description of dataset, intended for human audience.
<b>identifier*</b>	“Each identifier must be unique across the agency’s catalog and remain fixed”.
<b>keyword*</b>	List of relevant keywords describing the dataset.
<b>modified*</b>	ISO 8601 formatted date of last update, or ISO 8601 formatted interval for frequently refreshed data.
<b>programCode*</b>	List of codes for related U.S. federal programs (mandatory only for federal agencies).
<b>publisher*</b>	Container for further metadata fields describing the dataset publisher: name*, subOrganizationOf, @type.
<b>title*</b>	Title without acronyms, meant for human audience.
<b>accrualPeriodicity</b>	Frequency of data updates, described with an ISO 8601 code.
<b>conformsTo</b>	URI for data standard, if applicable.
<b>describedBy</b>	URI for human-readable data dictionary defining dataset fields.
<b>dataQuality</b>	Boolean indicating whether the dataset meets federal data quality guidelines.
<b>issued</b>	ISO 8601 formatted date when the dataset was first published.
<b>landingPage</b>	URL for dataset-specific webpage.
<b>language</b>	RFC 5646 code for (human) language of dataset.
<b>license</b>	URL for applicable license.
<b>primaryITInvestmentUII</b>	IT Unique Investment Identifier; only relevant for tracking federal government expenditures.
<b>references</b>	URLs to any documents that supplement the dataset, excluding the data dictionary URL from the describedBy field.
<b>rights</b>	Explanation for status of `restricted public` or `non-public` datasets; instructions for appropriate access.
<b>spatial</b>	Required for spatial datasets; denotes one of bounding coordinates, point coordinates, a Geographic Markup Language- described feature, or a name from the GeoNames database.
<b>systemOfRecords</b>	Relevant only for federal agencies; URL to to relevant entry on Federal Register.
<b>temporal</b>	ISO 8601 formatted start and end dates, expressing the period to which the dataset applies.
<b>theme</b>	List of categories; can be specific to agency or sourced from ISO Topic Categories.
<b>isPartOf</b>	When datasets are related, there should be a parent dataset, and for its children, the value of this field should be the parent’s identifier.

Table 4 provides my assessment of how well metadata for the 911 Incident Response and Police Report Incidents datasets aligns with the POD 1.1 specification; in brief, the correspondence is quite good, as most missing elements are nonessential.

**Table 4.** Evaluating the correspondence of 911 Incident Response metadata (SPD, 2017a) and Police Report Incidents metadata (SPD, 2017c) with POD 1.1 metadata specification (POD, 2014, “Metadata schema”).

POD 1.1 element	Socrata/Seattle Open Data metadata status
accessLevel	<b>Absent</b>
bureauCode	n/a
fn	Present
hasEmail	Present
description	Present
identifier	Present
keyword	Present
modified	Present
programCode	n/a
publisher	Present (as list of dataset providers)
title	Present
accrualPeriodicity	Present
conformsTo	<b>Absent</b>
describedBy	Present
dataQuality	n/a
issued	Present
landingPage	Present (as URL of primer page)
language	<b>Absent</b>
license	Present
primaryITInvestmentUII	n/a
references	Present (as links to attachments)
rights	<b>Absent</b>
spatial	<b>Absent</b>
systemOfRecords	n/a
temporal	<b>Absent</b>
theme	Present
isPartOf	<b>Absent</b>

### Enrichment proposal

Based on this evaluation, I feel comfortable asserting that, for these datasets, (1) reusability is a more serious problem than discoverability, and (2) lack of a metadata standard is not the primary barrier to reuse. The first barrier falls outside the scope of metadata, even: it is the lack of *data standards*. Compared to crime datasets from other cities like Chicago (CPD, 2017) and to the SpotCrime Open Crime Standard (SOCS), Seattle is missing descriptive fields that would add a lot of richness to the dataset. In particular, the wholesale redaction of narrative from police reports data is motivated by commendable privacy concerns, but it is out of line with peer cities and reduces the value of the data.

At the level of metadata, the priority should be to populate the ‘data dictionary’ fields that describe the contents of the dataset, including various codes and encodings. Again, the City of Chicago’s comparable dataset provides this information using the same Socrata platform: important qualifiers are given, such as “this is sometimes a best estimate”, “partially redacted address”, “See the list of IUCR codes at [URL]”, etc. (CPD, 2017).

The second priority should be to write better narrative descriptions that reflect important contextual information—the who, what, when, where, and why of the dataset. Some of

these factors are already present in the metadata; what's missing is a narrative that connects the nouns, giving the audience a better picture of the process underlying the data.

Finally, I tentatively recommend adoption of the POD 1.1 metadata standard, because it is designed for the domain of open government data and because it has been developed and adopted by a hugely significant player (the U.S. federal government). The major benefits of standardized metadata would be increased discoverability and interoperability; however, my recommendation is tentative because of the potential cost of migrating existing metadata. A quick intermediate solution (or a potential alternative solution) would be to develop and publish crosswalks between the local standard and other major standards.

## References

- Chicago Police Department (CPD). (2017, February 13). [dataset]. Retrieved from <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>
- Data.gov. (n.d.). About Data.gov. Retrieved from <https://www.data.gov/about>
- Project Open Data (POD). (2014, November 6). Project Open Data metadata schema v1.1. Retrieved from <https://project-open-data.cio.gov/v1.1/schema/>
- Project Open Data (POD). (2014, November 4). Metadata resources for schema v1.1. Retrieved from <https://project-open-data.cio.gov/v1.1/metadata-resources>
- Socrata. (2015, November 5). Use custom metadata fields and categories. Retrieved from <https://support.socrata.com/hc/en-us/articles/202787576-Use-custom-metadata-fields-and-categories>
- Socrata. (2016, February 8). Metadata: Best practices. Retrieved from <https://support.socrata.com/hc/en-us/articles/202950168-Metadata-best-practices>
- Socrata. (2017, January 23). Define metadata at import. Retrieved from <https://support.socrata.com/hc/en-us/articles/202950148>
- Seattle Police Department (SPD). (2017a, January 21). 911 incident response. Retrieved from <https://data.seattle.gov/Public-Safety/Seattle-Police-Department-911-Incident-Response/3k2p-39jp>
- Seattle Police Department (SPD). (2017b, January 21). 911 incident response [dataset]. Retrieved from <https://data.seattle.gov/Public-Safety/Seattle-Police-Department-911-Incident-Response/3k2p-39jp/data>
- Seattle Police Department (SPD). (2017c, January 21). Police report incident. Retrieved from <https://data.seattle.gov/Public-Safety/Seattle-Police-Department-Police-Report-Incident/7ais-f98f>
- Seattle Police Department (SPD). (2017d, January 21). Police report incident [dataset]. Retrieved from <https://data.seattle.gov/Public-Safety/Seattle-Police-Department-Police-Report-Incident/7ais-f98f/data>
- Seattle Police Department (SPD). (n.d. a). Public rules for CAD data release [Microsoft Word file]. Retrieved from <https://data.seattle.gov/api/assets/98E5052D-5E37-4BB4-B903-50BD73C97CCC?download=true>
- Seattle Police Department (SPD). (n.d. b). Public rules for RMS data release [Microsoft Word file]. Retrieved from <https://data.seattle.gov/api/assets/BA4FF139-F37F-4CCA-871D-C827E8211DB0?download=true>
- Suzan, B. (2014, March 12). The SpotCrime Open Crime Standard (SOCS). Retrieved from <http://blog.spotcrime.com/2014/03/the-spotcrime-open-crime-data-standard.html>

## Appendix

### Seattle Police Department 911 Incident Response Public Safety

View Data → Download API Share ...

This dataset is all the Police responses to 9-1-1 calls within the city. Police response data shows all officers dispatched. To protect the security of a scene, the safety of officers and the public, and sensitive ongoing investigation, these events are added to the data.seattle.gov only after the incident is considered safe to close out. Data is refreshed on a 4 hour interval.

Updated  
February 9, 2017

Data Provided by  
City of Seattle, Department of Information  
Technology, Seattle Police Department

#### About this Dataset

Updated

**February 9, 2017**

Data Last Updated February 9, 2017  
Metadata Last Updated October 25, 2016

Date Created  
October 8, 2010

Views Downloads  
**111K** **194K**

Data Provided by City of Seattle, Department of Information Technology, Seattle Police Department  
Dataset Owner Seattle IT

Contact Dataset Owner

#### Data Owner

Owner Department of Information Technology

#### Refresh Frequency

Frequency hourly

#### Attachments

[CadDataReleaserules.docx](#)

#### Topics

Category Public Safety

Tags 911, police, crime, incident response, census911 incidents

#### Licensing and Attribution

License



#### What's in this Dataset?

Rows Columns  
**1.37M** **19**

### Columns in this Dataset

Column Name	Description	Type
<b>CAD CDW ID</b>	CAD CDWID	Plain Text T
<b>CAD Event Number</b>	CAD Event Number	Plain Text T
<b>General Offense Number</b>	General Offense Number	Plain Text T
<b>Event Clearance Code</b>	Event Clearance Code	Plain Text T
<b>Event Clearance Description</b>	Event Clearance Description	Plain Text T
<b>Event Clearance SubGroup</b>	Event Clearance SubGroup	Plain Text T
<b>Event Clearance Group</b>	Event Clearance Group	Plain Text T
<b>Event Clearance Date</b>	Event Clearance Date	Date & Time
<b>Hundred Block Location</b>	Hundred Block Location	Plain Text T
<b>District/Sector</b>	Sector	Plain Text T
<b>Zone/Beat</b>	Beat	Plain Text T
<b>Census Tract</b>	Census_Tract	Plain Text T
<b>Longitude</b>	Longitude	Number #
<b>Latitude</b>	Latitude	Number #
<b>Incident Location</b>		Location
<b>Initial Type Description</b>		Plain Text T
<b>Initial Type Subgroup</b>		Plain Text T
<b>Initial Type Group</b>		Plain Text T
<b>At Scene Time</b>		Date & Time


# Seattle Police Department Police Report Incident COMMUNITY Public Safety

[View Data](#) [Download](#) [API](#) [Share](#) [...](#)

These incidents are based on initial police reports taken by officers when responding to incidents around the city. The information enters our Records Management System (RMS) and is then transmitted out to data.seattle.gov. This information is published within 6 to 12 hours after the report is filed into the system.

Updated  
February 9, 2017  
Data Provided by  
City of Seattle, Department of Information Technology, Seattle Police Department

## About this Dataset

<p>Updated <b>February 9, 2017</b></p> <p>Data Last Updated: February 9, 2017 Metadata Last Updated: September 12, 2016</p> <p>Date Created July 28, 2010</p> <hr/> <p>Views: <b>67K</b> Downloads: <b>17K</b></p> <hr/> <p>Data Provided by: City of Seattle, Department of Information Technology, Seattle Police Department Dataset Owner: spd2internetData</p> <p><a href="#">Contact Dataset Owner</a></p>	<p>Data Owner Owner: Department of Information Technology</p> <p>Refresh Frequency Frequency: daily</p> <p>Attachments <a href="#">RMSDataReleaserules.docx</a></p> <p>Topics Category: Public Safety Tags: crime, police, police report, census911incidents</p> <p>Licensing and Attribution License: </p>
---	--

## What's in this Dataset?

Rows: <b>96.7K</b>	Columns: <b>19</b>
--------------------	--------------------



## Columns in this Dataset

Column Name	Description	Type
<b>RMS CDW ID</b>	RMS CDWID	Plain Text T
<b>General Offense Number</b>	General Offense Number	Plain Text T
<b>Offense Code</b>	Offense_Code	Plain Text T
<b>Offense Code Extension</b>	Offense_Code_Extension	Plain Text T
<b>Offense Type</b>	Offense Type	Plain Text T
<b>Summary Offense Code</b>	Summary Offense Code	Plain Text T
<b>Summarized Offense Description</b>	Summarized_Offense_Description	Plain Text T
<b>Date Reported</b>	Date Reported	Date & Time 📅
<b>Occurred Date or Date Range Start</b>	Occurred Date or Date Range Start	Date & Time 📅
<b>Occurred Date Range End</b>	Occurred Date Range End	Date & Time 📅
<b>Hundred Block Location</b>	Hundred Block Location	Plain Text T
<b>District/Sector</b>	District/Sector	Plain Text T
<b>Zone/Beat</b>	Zone/Beat	Plain Text T
<b>Census Tract 2000</b>	Census Tract 2000	Plain Text T
<b>Longitude</b>	Longitude	Number #
<b>Latitude</b>	Latitude	Number #
<b>Location</b>		Location 📍
<b>Month</b>	Month crime occurred, specifically added for grouping a...	Number #
<b>Year</b>	Year crime occurred, specifically added for grouping and ...	Number #