OPEN DATA GUIDEBOOKS

GETTING META WITH METADATA

A City's guide to high quality, discoverable, and understandable open data.

What Works Cities | Bloomberg Philanthropies

Table of Contents

Open Data Metadata Guide

Metadata - descriptive information about data - is critical to helping visitors find and use published data effectively. Good metadata reduces the need for visitors to seek personal assistance, helps prevent misinterpretation of data, and encourages higher data quality.

Metadata is generally divided into two types:

- Metadata that provides an overview of the data. This kind of metadata helps people find the data through internet searches, while navigating your portal, or even while navigating other data portals which might include your catalog.
- Metadata that provides details about specific parts of your data. This kind of metadata enables people to use your data effectively, by helping them understand the various elements it includes and potential limitations.

This guide avoids the details of specific technologies; however, it takes into account existing (and emerging) national and international standards for metadata. Refer to the Additional Resources section for more information.

To assist cities in advancing open data programs in their own communities, the Center for Government Excellence at Johns Hopkins University, a partner in the What Works Cities initiative, has created this metadata guide. By learning from the experiences of other cities and following Center-developed best practices, cities will have a greater understanding of metadata and be well on their way to national leadership in open government data.

Categories

As an open data portal grows beyond 25-30 datasets, it is more helpful to visitors if they can browse for data by a subject matter or theme. These categories are short, at most a word or two, and allow related data to be grouped together. Categories also empower visitors to explore available data for inspiration, rather than requiring them to use a search tool to find something specific.

Creating categories is often a fundamental step when implementing an open data portal. Categories do not need to be permanent; it makes sense to have three to four categories for a small number of datasets, and re-evaluate them on an annual basis as more data is published. Most mature open data portals have 8 to 12 categories. Having too many might mean that the categories are not broad enough. Having too few, especially when combined with a large number of datasets, might mean that the categories are too broad and less helpful for visitors.

Although there is no consistent set of categories between open data portals, the following are quite common and might serve as a starting point: *Business*, *Education*, *Environment*, *Finance*, *Health*, *Human (or Social) Services*, *Property*, *Public Safety*, *Recreation*, and *Transportation*. A librarian or information architect can provide insight and assistance when creating or revising the list of categories.

Dataset Metadata

Without dataset metadata, a catalog of published data could not exist. Many open data portals include the necessary tools to create dataset metadata when publishing new data. Some open data portals automatically update the metadata when editing datasets. Each dataset you publish will include many of the following metadata elements.

Basic Elements

Basic metadata elements provide the most important pieces of information to help visitors find data and determine if it is what they need. Many of these items will appear directly in catalog navigation pages or search results.

- **Title (or Name)**: Human-readable name for the data. It should be in plain English and include sufficient detail to facilitate search and discovery. Acronyms should be avoided.
- **Description**: Human-readable description (e.g., an abstract) with sufficient detail to enable a user to quickly understand whether the asset is of interest.
- **Category (or Theme)**: Main thematic category of the dataset, usually chosen from a predefined list. Refer to the Categories section of this guide for more information. Some open data portals limit a dataset to one category; others allow multiple.
- **Keywords (or Tags)**: Tags (or keywords) are generally single words which help visitors discover the data; please include terms that would be used by technical and non-technical users. Keywords can also be used by recommendation engines to help visitors discover similar datasets.
- **Modification Date**: The most recent date on which the dataset was changed, updated, or modified.
- **Contact Information**: The name and email address of the publisher of a dataset.
- **License**: Often datasets on open data portals are available in the public domain with no restritions on reuse (usually this is noted in the site's Terms of Service or Data Policy), however there may be circumstances where a specific dataset is offered using a different license.

Advanced Elements

Advanced metadata elements provide helpful information that allows third-party software to consume both data catalogs and datasets. These items might not appear in catalog navigation pages or search results, but allow for sharing with other open data portals and

search engines.

- **Frequency**: The frequency with which dataset is updated, in plain English. For example, “Never,” “Hourly,” “Daily,” “Weekdays,” “Weekly,” “Semi-monthly,” “Monthly,” “Quarterly,” “Semi-annually,” “Annually,” etc. This helps visitors know how often they should check for new data, and is particularly valuable for software programmers who may set up automatic downloads.
- **Temporal Coverage**: The range of time included in this dataset. This may reflect a general range for all the records, or may reflect the earliest and latest dates from records in the data.
- **Spatial Coverage**: The geographic area for which this dataset is relevant. A place name - particularly one associated with clear boundaries - is most commonly used. If the dataset includes geospatial information, spatial coverage can represent a bounding rectangle or polygon of all the geography contained within it, though this is uncommon.

Refer to Appendix A for sample dataset metadata.

Column Metadata

Although column metadata is often limited or left out entirely, it is very helpful to data consumers who frequently work with, write software for, or analyze datasets. Column metadata attributes provide important details about the data which the column contains. Many open data portals include the necessary tools to create column metadata when publishing new data.

- **Name**: Human-readable name of the column. It should be in plain English and usually a word, or a few words at the most.
- **Description**: Human-readable description of the column's contents. This description should include how values in this column are created or updated; address any data quality concerns, such as unexpected or unusual values;, and explain any meanings which might be stored as codes,often used for record classification, and more frequent in source data systems designed for limited storage space.
- **Data Type**: Specifying a data type helps improve the consistency and quality of data. Common data types are text, numbers, dates/times, booleans (yes/no or true/false), and geometry (points, lines, polygons). Some open data portals will prevent records from being added or updated if the type of a value is incorrect.
- **Required**: Specifying whether a value is required in the column for every row in the table helps improve the quality of data. Some open data portals will prevent records from being added or updated if the column is marked as required but the datum was not included.
- **Machine Name**: Machine-readable version of the column's Name. This is often a copy of the Name, with changes that make it suitable for computer software to use. These changes may include replacing spaces with underscores (or removing them entirely), applying camel-case, and/or ensuring it is unique from other column names.

Refer to Appendix B for sample column metadata.

Additional Resources

This metadata guide is based upon current best practices in the US open government data sector. The following are additional resources which may be helpful for greater detail and guidance:

- Project Open Data Metadata Schema
- Dublin Core
- Federal Geographic Data Committee - Metadata
- The Open Metadata Handbook

Appendix A: Sample Dataset Metadata

Standard Dataset Fields

The U.S. federal government has created the Project Open Data metadata schema standard to implement the federal open data policy. The Project Open Data schema is based on the international DCAT metadata schema used by open data programs around the world and has been mapped to many standards. The Project Open Data schema must be preseneted as a JSON file to be ingested by Data.gov. This schema is natively available with many open data portal providers including: *Azavea*, *Esri Open Data*, *NuCivic's DKAN*, *OpenGov*, and *Socrata*, and is easily added to *CKAN* sites with an extension or can be generated on an ad hoc basis with these tools.

Field	Label	Definition	Required
title	Title	Human-readable name of the asset. Should be in plain English and include sufficient detail to facilitate search and discovery.	Always
description	Description	Human-readable description (e.g., an abstract) with sufficient detail to enable a user to quickly understand whether the asset is of interest.	Always
keyword	Tags	Tags (or keywords) help users discover your dataset; please include terms that would be used by technical and non-technical users.	Always
modified	Last Update	Most recent date on which the dataset was changed, updated or modified.	Always
publisher	Publisher	The publishing entity and optionally their parent organization(s).	Always
contactPoint	Contact Name and Email	Contact person's name and email for the asset.	Always
identifier	Unique Identifier	A unique identifier for the dataset or API as maintained within an Agency catalog or database.	Always
		The degree to which this dataset **could** be made publicly-available, *regardless of whether it has been*	

accessLevel	Public Access Level	*made available*. Choices: public (Data asset is or could be made publicly available to all without restrictions), restricted public (Data asset is available under certain use restrictions), or non-public (Data asset is not available to members of the public).	Always
license	License	The license or non-license (i.e. Public Domain) status with which the dataset or API has been published. See Open Licenses for more information.	If-Applicable
rights	Rights	This may include information regarding access or restrictions based on privacy, security, or other policies. This should also serve as an explanation for the selected "accessLevel" including instructions for how to access a restricted file, if applicable, or explanation for why a "non-public" or "restricted public" data asset is not "public," if applicable. Text, 255 characters.	If-Applicable
spatial	Spatial	The range of spatial applicability of a dataset. Could include a spatial region like a bounding box or a named place.	If-Applicable
temporal	Temporal	The range of temporal applicability of a dataset (i.e., a start and end date of applicability for the data).	If-Applicable
distribution	Distribution	A container for the array of Distribution objects. See Dataset Distribution Fields below for details.	If-Applicable
@type	Metadata Type	IRI for the JSON-LD data type. This should be `dcat:Dataset` for each Dataset.	No
accrualPeriodicity	Frequency	The frequency with which dataset is published.	No
conformsTo	Data Standard	URI used to identify a standardized specification the dataset conforms to.	No
describedBy	Data Dictionary	URL to the data dictionary for the dataset. Note that documentation other than a data dictionary can be referenced using Related Documents (`references`).	No

describedByType	Data Dictionary Type	The machine-readable file format (IANA Media Type also known as MIME Type) of the dataset's Data Dictionary (`describedBy`).	No
isPartOf	Collection	The collection of which the dataset is a subset.	No
issued	Release Date	Date of formal issuance.	No
language	Language	The language of the dataset.	No
landingPage	Homepage URL	This field is not intended for an agency's homepage (e.g. www.agency.gov), but rather if a dataset has a human-friendly hub or landing page that users can be directed to for all resources tied to the dataset.	No
references	Related Documents	Related documents such as technical information about a dataset, developer documentation, etc.	No
theme	Category	Main thematic category of the dataset.	No

Federal Dataset Fields

The U.S. federal requirement also requires the following metadata fields. You should consider requiring local department codes, systems of record, and associated IT spending if helpful for your open data catalog. If you do not have unique governmentwide codes related to these areas, you might consider creating those.

Field	Label	Definition	Required
bureauCode[USG]	Bureau Code	Federal agencies, combined agency and bureau code from OMB Circular A-11, Appendix C (PDF, CSV) in the format of `015:11` .	Always
programCode[USG]	Program Code	Federal agencies, list the primary program related to this data asset, from the Federal Program Inventory. Use the format of `015:001` .	Always
dataQuality[USG]	Data Quality	Whether the dataset meets the agency's Information Quality Guidelines (true/false).	No
primaryITInvestmentUII[USG]	Primary IT Investment UII	For linking a dataset with an IT Unique Investment Identifier (UII).	No
systemOfRecords[USG]	System of Records	If the system is designated as a system of records under the Privacy Act of 1974, provide the URL to the System of Records Notice related to this dataset.	No

Appendix B: Sample Column Metadata