

Data.gov, DCAT, & data.json

**Federal Committee on Statistical Methodology
Metadata Workshop September 14, 2018**



**Rebecca Williams, Digital Services Expert
Office of the Federal Chief Information Officer
Office of Management and Budget**

- 03** The Evolution/Decentralization of Data.gov
- 08** Adopting the WC3's Data Catalog Vocabulary (DCAT) for the United States
- 15** First the agencies, then the vendors, then the world

Contents

The Evolution/Decentralization of Data.gov

ALEXIS MADRIGAL SCIENCE 05.21.09 06:05 PM

DATA.GOV LAUNCHES TO MIXED REVIEWS

SHARE



Data.gov launched today with 47 datasets from across the government.

That's a tiny fraction of the Feds' gargantuan information stores, and the site is clearly in beta, but open-government advocates see the new site as a sign of good things to come for [government transparency](#).

"Data.gov says that our information is your information," said Ellen Miller, executive director of the [Sunlight Foundation](#). "It bothers me less that there are 50 feeds available today because it represents this enormous change in attitude about what public means. It means it's online. It's means it's available. I think it's a dramatic breakthrough in the role of government."

The sheer scale of government data presents many problems, as we've noted in the [Open Up Government Data wiki](#). More than 100



THE DIRECTOR

May 9, 2013

EO 13642

Title 3—The President

Executive Order 13642 of May 9, 2013

Making Open and Machine Readable the New Default for Government Information

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

Section 1. General Principles. Openness in government strengthens our democracy, promotes the delivery of efficient and effective services to the public, and contributes to economic growth. As one vital benefit of open government, making information resources easy to find, accessible, and usable can fuel entrepreneurship, innovation, and scientific discovery that improves Americans' lives and contributes significantly to job creation.

Decades ago, the U.S. Government made both weather data and the Global Positioning System freely available. Since that time, American entrepreneurs and innovators have utilized these resources to create navigation systems, weather newscasts and warning systems, location-based applications, precision farming tools, and much more, improving Americans' lives in countless ways and leading to economic growth and job creation. In recent years, thousands of Government data resources across fields such as health and medicine, education, energy, public safety, global development, and finance have been posted in machine-readable form for free public use on Data.gov. Entrepreneurs and innovators have continued to develop a vast range of useful new products and businesses using these public information resources, creating good jobs in the process.


To promote continued job growth, Government efficiency, and the social good that can be gained from opening Government data to the public, the default state of new and modernized Government information resources shall be open and machine readable. Government information shall be managed as an asset throughout its life cycle to promote interoperability and openness, and, wherever possible and legally permissible, to ensure that data are released to the public in ways that make the data easy to find, accessible, and usable. In making this the new default state, executive departments and agencies (agencies) shall ensure that they safeguard individual privacy, confidentiality, and national security.


Sec. 2. Open Data Policy. (a) The Director of the Office of Management and Budget (OMB), in consultation with the Chief Information Officer (CIO), Chief Technology Officer (CTO), and Administrator of the Office of Information and Regulatory Affairs (OIRA), shall issue an Open Data Policy to advance the management of Government information as an asset, consistent with my memorandum of January 21, 2009 (Transparency and Open Government), OMB Memorandum M-10-06 (Open Government Directive), OMB and National Archives and Records Administration Memorandum M-12-18 (Managing Government Records Directive), the Office of Science and Technology Policy Memorandum of February 22, 2013 (Increasing Access to the Results of Federally Funded Scientific Research), and the CIO's strategy entitled "Digital Government: Building a 21st Century Platform to Better Serve the American People." The Open Data Policy shall be updated as needed.

(b) Agencies shall implement the requirements of the Open Data Policy and shall adhere to the deadlines for specific actions specified therein.

M-13-13

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Sylvia M. Burwell, 
Director

Steven VanRoekel, 
Federal Chief Information Officer

Todd Park, 
U.S. Chief Technology Officer

Dominic J. Mancini, 
Acting Administrator, Office of Information and Regulatory Affairs

SUBJECT: Open Data Policy—Managing Information as an Asset

Information is a valuable national resource and a strategic asset to the Federal Government, its partners, and the public. In order to ensure that the Federal Government is taking full advantage of its information resources, executive departments and agencies (hereafter referred to as "agencies") must manage information as an asset throughout its life cycle to promote openness and interoperability, and properly safeguard systems and information. Managing government information as an asset will increase operational efficiencies, reduce costs, improve services, support mission needs, safeguard personal information, and increase public access to valuable government information.

Making information resources accessible, discoverable, and usable by the public can help fuel entrepreneurship, innovation, and scientific discovery—all of which improve Americans' lives and contribute significantly to job creation. For example, decades ago, the Federal Government made both weather data and the Global Positioning System (GPS) freely available to anyone. Since then, American entrepreneurs and innovators have used these resources to create navigation systems, weather newscasts and warning systems, location-based applications, precision farming tools, and much more.

Pursuant to Executive Order of May 9, 2013, *Making Open and Machine Readable the New Default for Government Information*, this Memorandum establishes a framework to help institutionalize the principles of effective information management at each stage of the information's life cycle to promote interoperability and openness. Whether or not particular information can be made public, agencies can apply this framework to all information resources to promote efficiency and produce value.

Specifically, this Memorandum requires agencies to collect or create information in a way that supports downstream information processing and dissemination activities. This includes using machine-readable and open formats, data standards, and common core and extensible metadata for all new

PROJECT OPEN DATA

Open Data Policy – Managing Information as an Asset

1. Background

Data is a valuable national resource and a strategic asset to the U.S. Government, its partners, and the public. Managing this data as an asset and making it available, discoverable, and usable – [in a word, open](#) – not only strengthens our democracy and promotes efficiency and effectiveness in government, but also has the potential to create economic opportunity and improve citizens' quality of life.

For example, when the U.S. Government released weather and GPS data to the public, it fueled an industry that today is valued at tens of billions of dollars per year. Now, weather and mapping tools are ubiquitous and help everyday Americans [navigate their lives](#).

The ultimate value of data can often not be predicted. That's why the U.S. Government released a [policy](#) that instructs agencies to manage their data, and information more generally, as an asset from the start and, wherever possible, release it to the public in a way that makes it open, discoverable, and usable.

The White House developed Project Open Data – this collection of code, tools, and case studies – to help agencies adopt the Open Data Policy and unlock the potential of government data. Project Open Data will evolve over time as a community resource to facilitate broader adoption of open data practices in government. Anyone – government employees, contractors, developers, the general public – can view and contribute. Learn more about [Project Open Data Governance](#) and dive right in and help to build a better world through the power of open data.

2. Definitions

This section is a list of definitions and principles used to guide the project.

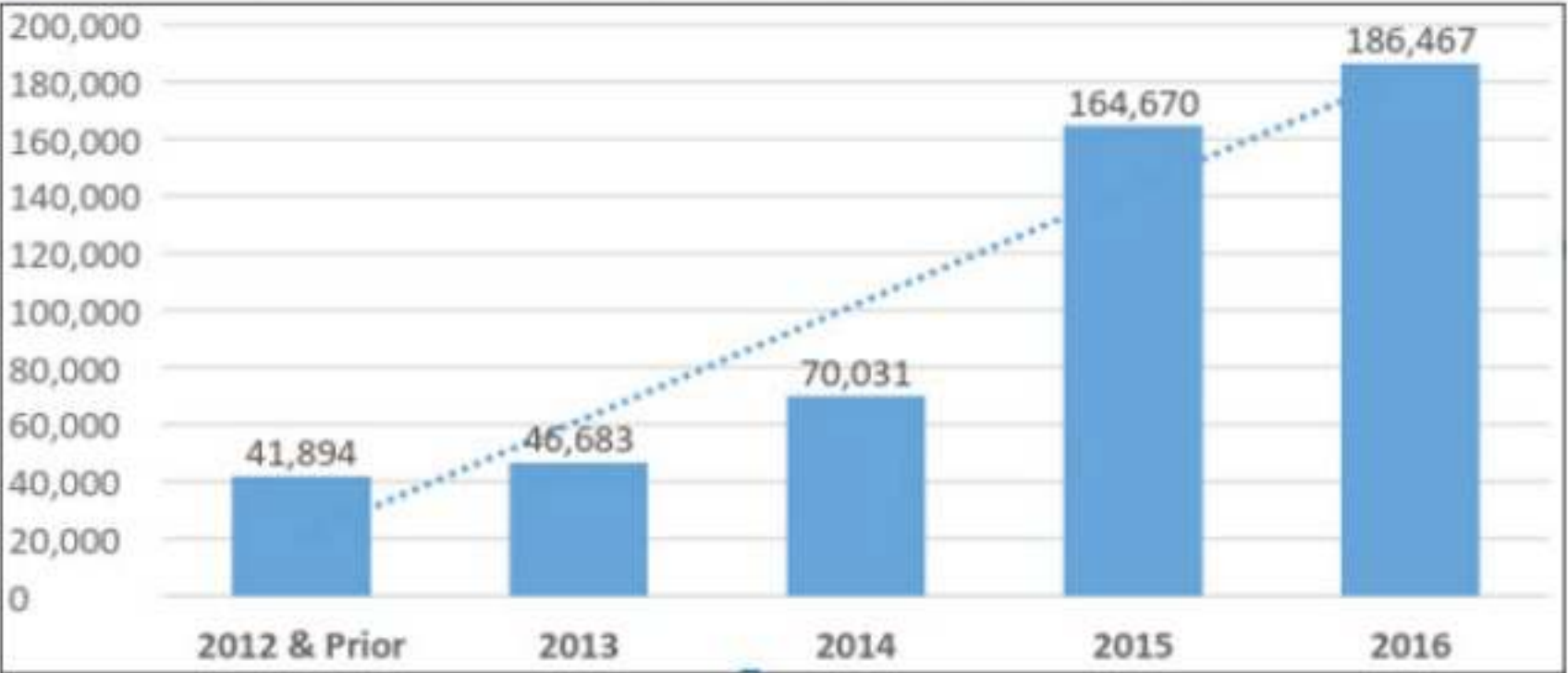
2-1 [Open Data Principles](#) - The set of open data principles.

2-2 [Standards, Specifications, and Formats](#) - Standards, specifications, and formats supporting open data objectives.

2-3 [Open Data Glossary](#) - The glossary of open data terms.

2-4 [Project Open Data Metadata Schema](#) - The schema used to describe datasets, APIs, and published data at [agency.gov/data](#).

Figure C2: Data Sets Available on Data.gov increased by over 400% from 2012 - 2016¹⁷



STATE OF FEDERAL IT REPORT / PUBLIC RELEASE VERSION 1.0

Adopting the WC3's Data Catalog Vocabulary (DCAT) for the United States



Data Catalog Vocabulary (DCAT)

W3C Recommendation 16 January 2014

This version:

<http://www.w3.org/TR/2014/REC-vocab-dcat-20140116/>

Latest published version:

<http://www.w3.org/TR/vocab-dcat/>

Implementation report:

http://www.w3.org/2011/gld/wiki/DCAT_Implementations

Previous version:

<http://www.w3.org/TR/2013/PR-vocab-dcat-20131217/>

Editors:

[Fadi Maali](#), [DERI](#), [NUI Galway](#)

[John Erickson](#), [Tetherless World Constellation \(RPI\)](#)

Contributors:

[Phil Archer](#), [W3C/ERCIM](#)

Please refer to the [errata](#), a list of issues with this document discovered after publication.

This document is also available in this non-normative format: [diff to previous version](#)

The English version of this specification is the only normative version. Non-normative [translations](#) may also be available.

Copyright © 2012-2014 W3C® ([MIT](#), [ERCIM](#), [Keio](#), [Beihang](#)), All Rights Reserved. W3C [liability](#), [trademark](#) and [document use](#) rules apply.

Abstract

Feedback on common core schema #49

New issue

Closed

jpmckinney opened this issue on May 16, 2013 · 4 comments



jpmckinney commented on May 16, 2013

Contributor

...

In general, great work! A few questions/comments:

- `contactPoint` is being added to `DCAT`. To maintain conformance, consider renaming the `person` property to `contactPoint`.
 - see #57
- I prefer `schema:email` to `foaf:mbox`, but either is fine. Just letting you know that you can call that property `email` as the two properties are `owl:sameAs`, I believe.
- ~~Did `systemRecords` come from an existing vocabulary, or did you need to mint it for your use case? If it is new, it would be good practice to mention it, as you've done for `accessLevel`.~~
 - fixed #60
- ~~`dc:byteSize` has been renamed to `dc:byteSize`, which is a clearer term. As mentioned in #32, byte size is much less prone to error than various abbreviations for byte size, like GB, GiB, etc.~~
 - `size` has been removed #114
- Dublin Core has a different, more flexible way of expressing `dc:temporal`. You may want to consider making it an object with "start" and "end" fields. This would also solve #43
- ~~`CV`, granularity was dropped from DCAT because it was underspecified and publishers simply didn't know how/when/why to use it. You may want to drop it as well unless you have a very clear use case in mind (that others share).~~
 - fixed #115
- `dataDictionary` was also dropped due to inconsistent usage. `dataQuality` was dropped from DCAT for similar reasons, but I see you have a clear definition of what it means, so it makes sense to keep that one. Just flagging these to let you know that you can very likely get rid of them if you were already on the fence about them.
 - see #57
- `WebService` and `Feed` are classes in DCAT (subclasses of `Distribution` to be specific), not properties. To be consistent, you can have a `@type` field (see #23 for why this term should be used) on each `Distribution` which would be used to say whether the distribution is a `WebService`, `Feed` or `Download`. This also solves the issue of having multiple APIs and feeds for the same dataset, which is not handled by the current spec.
 - see #57
- As mentioned in #39, please use arrays and not comma-separated values.
- It would be cool to add the three properties mentioned in #23 (`@context`, `@id` and `@type`)

Ref.: the latest version of DCAT

Assignees

No one assigned

Labels

question

schema v1.1

Projects

None yet

Milestone

No milestone

3 participants



Version 1.1 Update

In the year since the release of the Open Data Policy, agencies and the public have suggested several updates to the metadata schema. In the interest of stability, these updates have been tied together into a methodical update to a version 1.1 of the metadata schema. Each issue has been rigorously discussed in its own issue thread and at the [July government-wide offsite session](#) dedicated to this update.

Changes

- [Changelog for the version 1.1 schema](#).
- These updates have been managed through [issues in a single milestone](#). Each issue within the milestone contains the related discussion and a link to the proposed edits.
- The proposed edits can also be found bundled in [this combined pull request](#).

Resources

- [New updated metadata schema page](#).
- [Field mapping from v1.0 to v1.1 schema](#).
- [Updated Metadata Schema v 1.1 Diagram](#).
- [Sample data.json files with the version 1.1 schema](#).

Guidance

- [“Project Open Data Metadata Updates v 1.1” presentation slides and video](#) from the [October 15th webinar](#).

Supplemental Information

Agencies are encouraged to supplement these requirements with other information, outreach and tools (e.g. blog posts, GitHub tools, customer engagement tools, etc.). Components, bureaus, and programs are also encouraged to highlight their work implementing the Open Data Policy through their own channels.

Catalog Fields

Label	POD v1.1	POD v1.0	CKAN API	DCAT	Schema.org
Metadata Context	@context	n/a	n/a	n/a	n/a
Metadata Catalog ID	@id	n/a	n/a	n/a	n/a
Metadata Type	@type	n/a	n/a	n/a	itemtype attribute
Schema Version	conformsTo	n/a	n/a	n/a	n/a
Schema URL	describedBy	n/a	n/a	n/a	n/a
Dataset	dataset	n/a	results	dct:dataset	dataset

Dataset Fields

Note the mapping for **license** and **rights** from Project Open Data to DCAT applies the fields from the Dataset object in Project Open Data to each of the Distribution objects in DCAT.

Label	POD v1.1	POD v1.0	CKAN API	DCAT	Schema.org
Metadata Type	@type	n/a	n/a	n/a	itemtype attribute
Title	title	title	title	dct:title	name
Description	description	description	notes	dct:description	description
Tags	keyword	keyword	tags	dcat:keyword	keywords
Last Update	modified	modified	n/a	dct:modified	dateModified
Publisher	<i>publisher</i> → name	<i>publisher</i>	<i>organization</i> → title	dct:publisher → foaf:name	<i>publisher</i> → Organization:name
Publisher Parent Organization	<i>publisher</i> → subOrganizationOf	n/a	n/a	dct:publisher → org:subOrganizationOf	<i>publisher</i> → Organization:memberOf
Contact Name	<i>contactPoint</i> → fn	<i>contactPoint</i>	<i>maintainer</i>	dcat:contactPoint → vcard:fn	<i>provider</i> → Person:name

Important aspects to keep in mind

- The DCAT Application profile for data portals in Europe (DCAT-AP) is a specification based on W3C's Data Catalogue vocabulary (DCAT) for describing metadata of public sector datasets in Europe.
- It is the standard used by the European Data Portal as well.
- Version 1.1 is released end of 2015 and available here:
https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-ap-v11
- Information on the national extensions of DCAT-AP can be found here:
<https://joinup.ec.europa.eu/document/national-extensions-analysis-dcat-ap>
- Updates and news can be found here:
<https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe>
- A DCAT-AP validation tool is available here:
https://joinup.ec.europa.eu/software/dcat-ap_validator/asset_release/dcat-ap-validation-service-110
- The next slide provides the mandatory and recommended classes.

Benefits of DCAT-AP:

By using a common metadata schema to describe datasets:

1. Data publishers increase the discoverability of the data and thus re-use
2. Data re-users can search across platforms without facing difficulties caused by the use of separate models or language differences

**First the agencies,
then the vendors,
then the world**

Project Open Data Dashboard															
https://labs.data.gov/dashboard/offices/2014-11-30/qa															
Project Open Data Dashboard Agencies Validator Converters ▾ Rubric Help ▾ About Sign in with MAX															
	EST	EST													
Department of Health and Human Services	29-Nov-2014 23:01:18 EST	29-Nov-2014 21:09:09 EST	1064	97.3%	12	11	98.0%	0.0%	2.0%	0.0%	0				
Department of Homeland Security	29-Nov-2014 23:01:19 EST	22-Aug-2014 10:15:10 EDT	322	100%	30	1	100%	0.0%	0.0%	0.0%	0				
Department of Housing and Urban Development	26-Nov-2014 23:01:10 EST		170	100%	29	14	87.6%	0.0%	12.4%	0.0%	0				
Department of Justice	29-Nov-2014 23:06:11 EST	22-Sep-2014 09:46:56 EDT	969	100%	10	4	61.1%	0.0%	38.9%	0.0%	0				
Department of Labor	29-Nov-2014 23:01:44 EST	29-Sep-2014 16:02:51 EDT	368	85.6%	24	11	100%	0.0%	0.0%	0.0%	0				
Department of State	29-Nov-2014 23:01:45 EST	29-Nov-2014 16:00:04 EST	113	98.2%	4	1	86.7%	10.6%	2.7%	0.0%	0				
Department of the Interior															

Data Catalog vendors who use Project Open Data v1.1 today



https://data.cityofnewyork.us/ x +

← → ↻ https://data.cityofnewyork.us/data.json ☆ ⓘ |

```
-
{ ...

  • @context: "https://project-open-data.cio.gov/v1.1/schema/catalog.jsonld",
  • @id: "https://data.cityofnewyork.us/data.json",
  • @type: "dcat:Catalog",
  • conformsTo: "https://project-open-data.cio.gov/v1.1/schema",
  • describedBy: "https://project-open-data.cio.gov/v1.1/schema/catalog.json",
  • dataset:
    -
    [ ...
      { ...
        • accessLevel: "public",
        • landingPage: "https://data.cityofnewyork.us/d/h9gi-nx95",
        • issued: "2018-09-14",
        • @type: "dcat:Dataset",
        • modified: "2018-09-14",
        • keyword:
          -
          [ ...
            • "nypd",
            • "collisions",
            • "bigapps",
            • "big apps",
            • "visionzero",
            • "vision",
            • "zero",
            • "nycopendata",
            • "traffic data"
          ],
        • contactPoint:
          -
          { ...
            • @type: "vcard:Contact",
            • fn: "NYC OpenData"
          },
        • publisher:
          -
          { ...
            • @type: "org:Organization",
            • name: "data.cityofnewyork.us"
          },
        ,
      }
    ]
  }
}
```


Harvest sources - Data.gov

+

← → ↺

https://catalog.data.gov/harvest

🔍 ☆ ⓘ 👤 ⋮



DATA

TOPICS -

IMPACT

APPLICATIONS

DEVELOPERS

CONTACT

DATA CATALOG

/ Datasets

Organizations

?

🏠 / Harvest Sources

Organization Types

A-Z

1-9

Clear All

Federal Government (723)

Non-Profit (145)

State Government (33)

University (31)

City Government (28)

Show More Organization Types

Frequency

A-Z

1-9

Clear All

MANUAL (973)

WEEKLY (5)

MONTHLY (2)

DAILY (1)

Type

A-Z

1-9

Clear All

waf (422)

waf-collection (354)

datajson (140)

Search harvest sources...

Order by:

Popular

981 harvest sources found

WV GIS Technical Center

There is no description for this harvest source

— Organization: WV GIS Technical Center

CFPB JSON

There is no description for this harvest source

— Organization: Consumer Financial Protection Bureau

USGS National Structures Dataset (NSD) Downloadable Data Collection

The USGS structures downloadable data from The National Map consists of data to include the name, function, location, and other core information and characteristics of selected manmade facilities. The types of structures collected are largely determined by the needs of disaster planning and emergency response, and homeland security organizations. Structures data are designed to be used in general mapping and in the analysis of structure related activities using geographic information system technology. The National Map structures data is commonly combined with other data themes, such as elevation, boundaries, and transportation, to produce general reference base maps. For additional information, go to <http://nationalmap.gov/structures.html>.

— Organization: U.S. Geological Survey, Department of the Interior

The schema.org approach for describing datasets is based on an effort recently standardized at W3C (the [Data Catalog Vocabulary](#)), which we expect will be a foundation for future elaborations and improvements to dataset description. While these industry [discussions](#) are evolving, we are confident that the standards that already exist today provide a solid basis for building a data ecosystem.

Technical Challenges

While we have released the guidelines on publishing the metadata, many technical challenges remain before search for data becomes as seamless as we feel it should be. These challenges include:

- **Defining more consistently what constitutes a dataset:** For example, is a single table a dataset? What about a collection of related tables? What about a protein sequence? A set of images? An API that provides access to data? We hope that a better understanding of what a dataset is will emerge as we gain more experience with how data providers define, describe, and use data.
- **Identifying datasets:** Ideally, datasets should have permanent identifiers conforming to some well known scheme that enables us to identify them uniquely, but often they don't. Is a URL for the metadata page a good identifier? Can there be multiple identifiers? Is there a primary one?
- **Relating datasets to each other:** When are two records describing a dataset "the same" (for instance, if one repository copies metadata from another)? What if an aggregator provides more metadata about the same dataset or cleans the data in some useful way? We are working on clarifying and defining these relationships, but it is likely that consumers of metadata will have to assume that many data providers are using these predicates imprecisely and need to be tolerant of that.
- **Propagating metadata between related datasets:** How much of the metadata can we propagate among related datasets? For instance, we can probably propagate provenance information from a composite dataset to the datasets that it contains. But how much does the metadata "degrade" with such propagation? We expect the answer to be different depending on the application: metadata for search applications may be less precise than, say, for data integration.

Google Dataset Search Beta

Try [boston education data](#) or [weather site:noaa.gov](#)



U.S. Hourly Precipitation Data

catalog.data.gov
data.globalchange.gov
+3more

Updated Feb 8, 2018



U.S. Hourly Precipitation Data Publication

data.nodc.noaa.gov
cmr.earthdata.nasa.gov
+2more

Updated May 2, 2013



U.S. Hourly Climate



NATIONAL CENTERS FOR
ENVIRONMENTAL INFORMATION

U.S. Hourly Precipitation Data



catalog.data.gov



data.globalchange.gov



cmr.earthdata.nasa.gov



ec2-52-38-26-42.us-west-2.compute.amazonaws.com



datamirror.org

27 scholarly articles cite this dataset ([View in Google Scholar](#))

Dataset updated Feb 8, 2018

Dataset provided by

[National Centers for Environmental Information](#)

Content Types

- Article
- Book
- Course
- Dataset**
- Employer Aggregate Rating
- Event
- Fact Check
- Job Posting
- Local Business
- Media
- Occupation
- Paywalled content
- Podcast
- Product
- Recipe
- Review
- Software App
- Speakable
- Top Places List
- Video
- Structured data codelab

Our approach to dataset discovery

We can understand structured data in Web pages about datasets, using either [schema.org Dataset markup](#), or equivalent structures represented in W3C's [Data Catalog Vocabulary \(DCAT\) format](#). We also exploring experimental support for structured data based on [W3C CSVW](#), and expect to evolve and adapt our approach as best practices for dataset description emerge. For more information about our approach to dataset discovery, see [Facilitating the discovery of public datasets](#).

Example

Here's an example for datasets using JSON-LD code and the Schema.org vocabulary in the Structured Data Testing Tool. The following example is based on a [real-world dataset description](#).

[SEE MARKUP](#)

The same vocabulary can be used in JSON-LD (preferred), RDFa 1.1, or Microdata syntax.

It is also possible to use W3C DCAT vocabulary. Here is a simple example using RDFa:

[SEE MARKUP](#)

Guidelines

Contents

- [Our approach to dataset discovery](#)
- Example
- Guidelines
 - Sitemap best practices
 - Source and provenance best practices
- Known Errors and Warnings
- Structured data type definitions
 - Dataset
 - DataCatalog
 - DataDownload
 - Provenance and license
 - Tabular datasets
- Help and tools

Thanks!