# Open Data & Metadata Schema Overview

Philip Ashlock

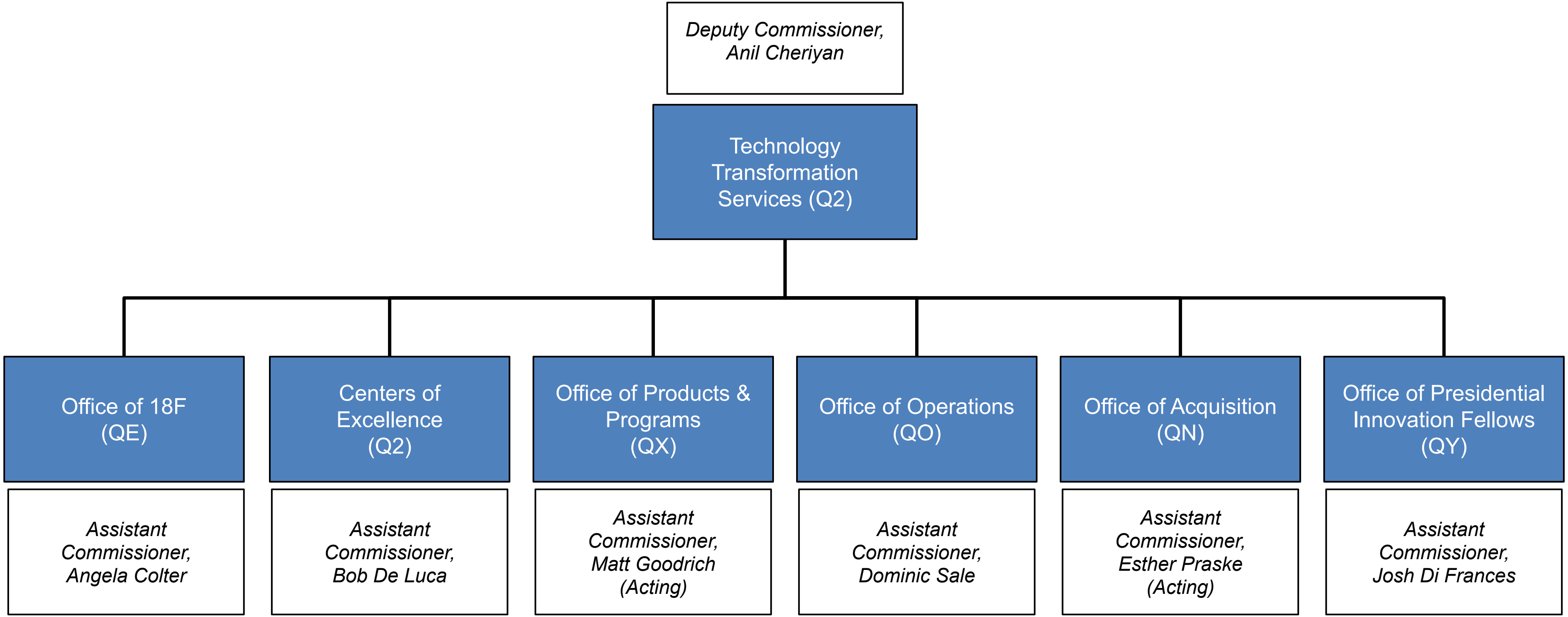
***Committee on Transparency and Reproducibility of Federal Statistics***
*May 2019*

GSA Technology Transformation Services

# My Role & Background

- Computer science, IT, civic tech, data standards, open data, open government, open source
- No formal background in statistics
- Entered the Federal government in 2012, Data.gov in 2013
- Led the Data & Analytics Portfolio at GSA TTS since 2016
- Support OFCIO on data policies including Federal Data Strategy
- Also support Office of the Chief Statistician on the National Reporting Platform for the Sustainable Development Goals (sdg.data.gov)
- Leading new project called US Data Federation to support distributed data management projects and data standards and interoperability efforts across government

# Technology Transformation Services Org Chart



*Deputy Commissioner, Anil Cheriyan*

Technology Transformation Services (Q2)

Office of 18F (QE)

*Assistant Commissioner, Angela Colter*

Centers of Excellence (Q2)

*Assistant Commissioner, Bob De Luca*

Office of Products & Programs (QX)

*Assistant Commissioner, Matt Goodrich (Acting)*

Office of Operations (QO)

*Assistant Commissioner, Dominic Sale*

Office of Acquisition (QN)

*Assistant Commissioner, Esther Praske (Acting)*

Office of Presidential Innovation Fellows (QY)

*Assistant Commissioner, Josh Di Frances*

DATA.GOV

# Overview

Data.gov is the U.S. Government's open data site, designed to unleash the power of government open data to help the public, achieve agency missions, drive innovation, and uphold the ideals of an open and transparent government.

- Unified Federal open data program and national cross-government catalog
- Launched in 2009 with 47 datasets and grown to over 200,000, approximately 9000 APIs. 60 Federal agencies (and over 100 sub-agencies) and 50 non-federal (state, city, county)
- Enterprise inventories of metadata of public and non-public datasets
- Statutory Requirement under OPEN Government Data Act (part of 2019 Foundations for Evidence Based Policymaking Act)
- Expands scope of Data.gov with more than 100 additional agencies, makes GSA's operation of Data.gov required by statute

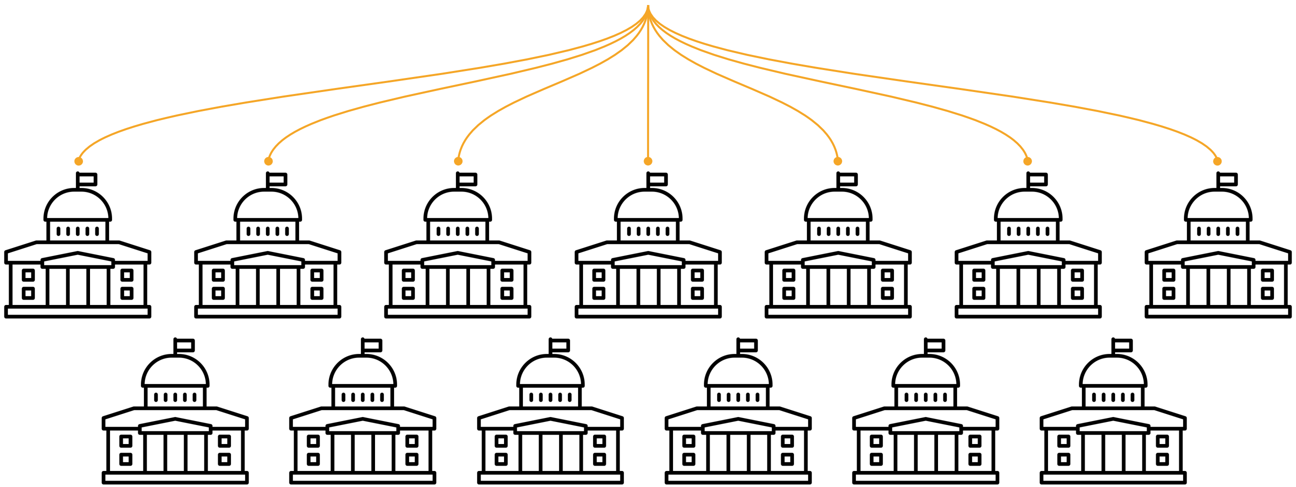
# Data.gov Harvesting Model

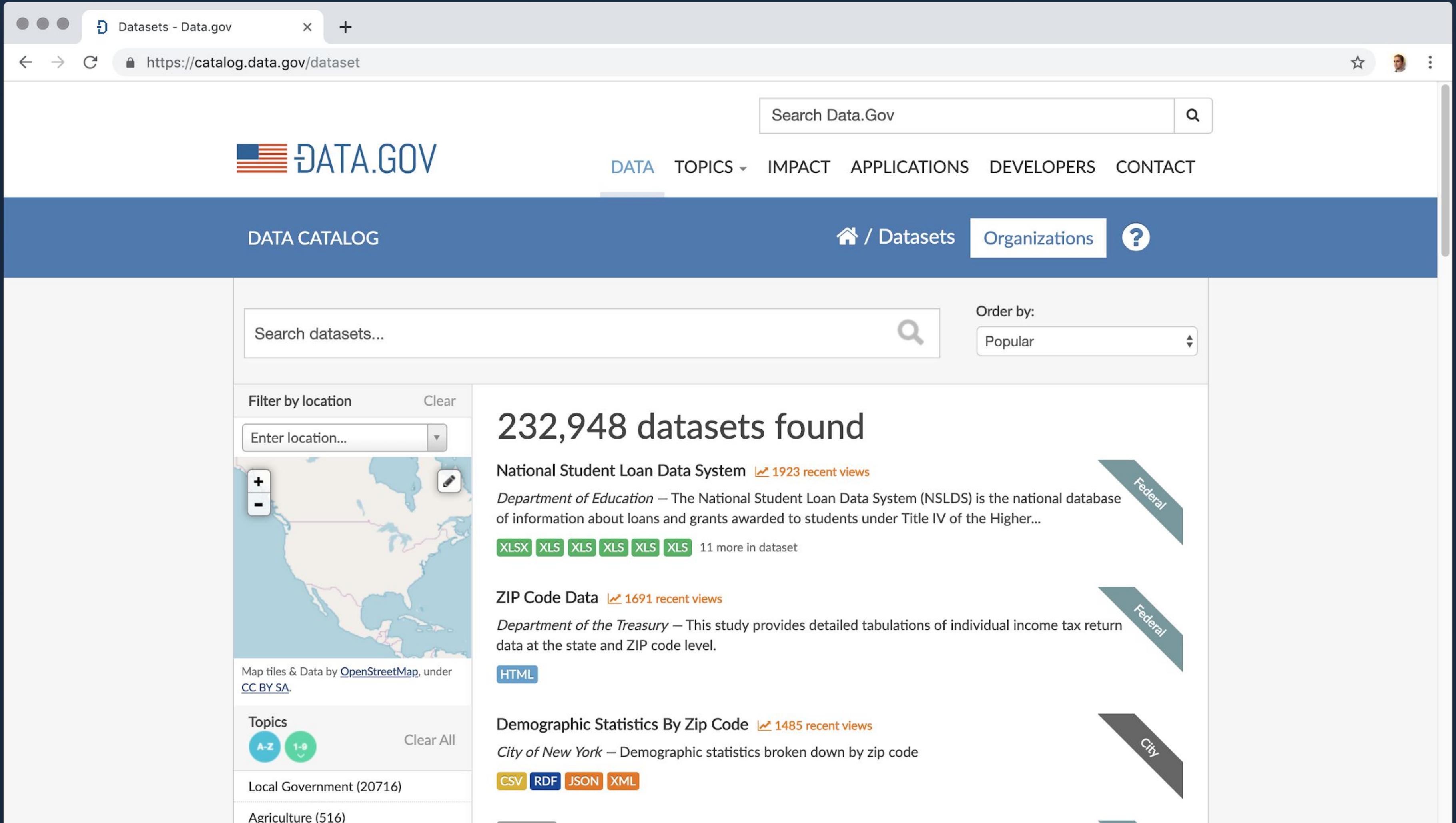
- Data.gov does not host data, it simply aggregates metadata like a library card catalog
- Data.gov pulls metadata from harvest sources and synchronizes them to catalog.data.gov
- Data.gov is not used to make changes to metadata
- All changes to metadata are made at the harvest source rather than at data.gov
- If a metadata record is no longer listed by the harvest source is will be deleted on catalog.data.gov during the next harvest job
- Most harvest sources on catalog.data.gov are synchronized every 24 hour

Google Dataset Search

data.world

DATA.GOV





Local Government (20716)

Agriculture (516)

# 232,948 datasets found

**National Student Loan Data System** 1923 recent views

*Department of Education* — The National Student Loan Data System (NSLDS) is the national database of information about loans and grants awarded to students under Title IV of the Higher...

XLSX XLS XLS XLS XLS XLS 11 more in dataset

Federal

**ZIP Code Data** 1691 recent views

*Department of the Treasury* — This study provides detailed tabulations of individual income tax return data at the state and ZIP code level.

HTML

Federal

**Demographic Statistics By Zip Code** 1485 recent views

*City of New York* — Demographic statistics broken down by zip code

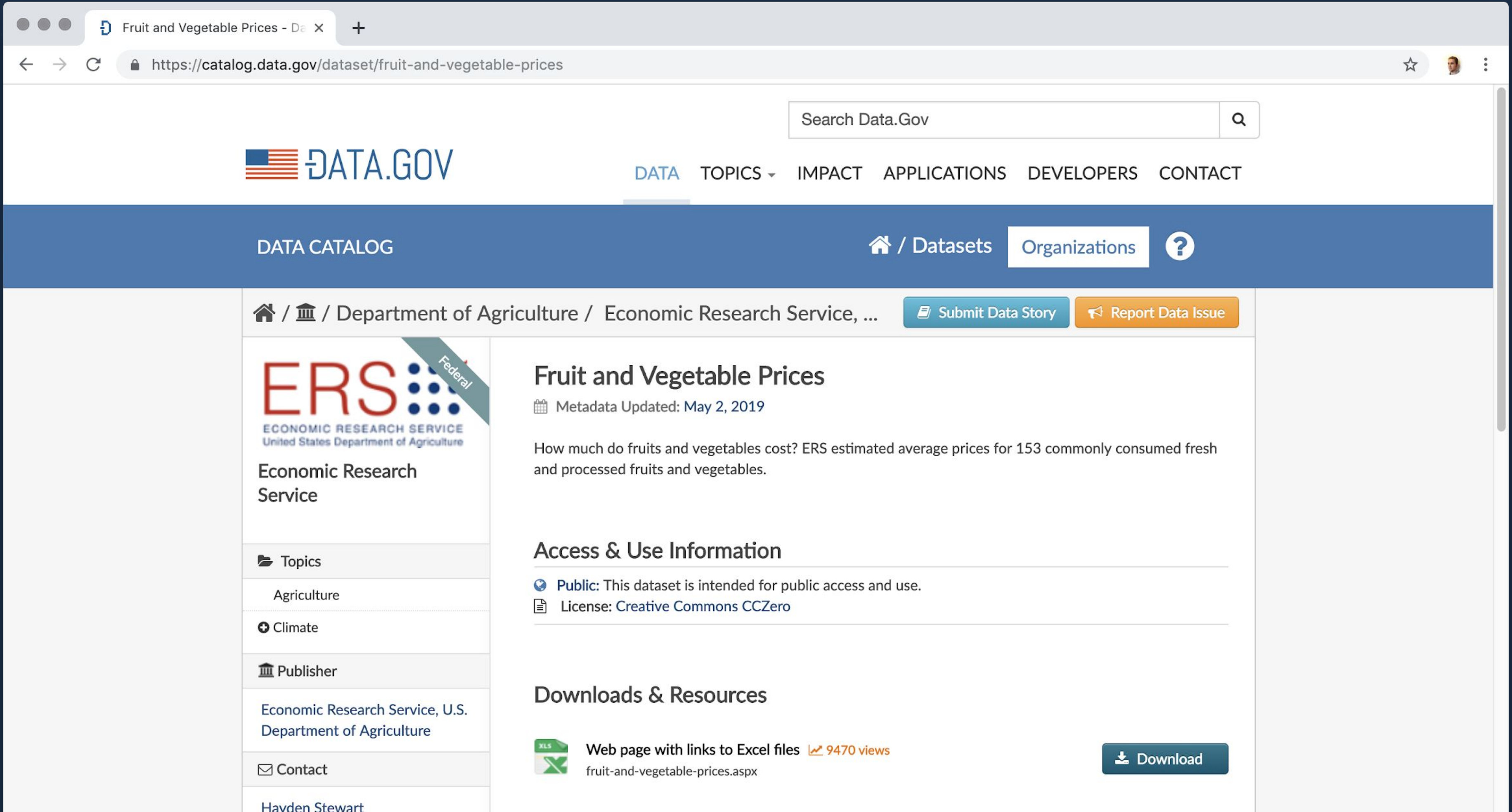
CSV RDF JSON XML

City

# All Data.gov Components

- National Data Catalog - catalog.data.gov
- Blog, Apps, Events - data.gov
- Metadata management SaaS - inventory.data.gov
- Help Desk CRM
- Policy support and interagency bi-weekly meetings
- Project Open Data website - (Under revision for new law)
- Project Open Data Dashboard
- In development: Data Catalog Kit (Multi-tenant CKAN)

# National Data Catalog - catalog.data.gov


Fruit and Vegetable Prices - Da
https://catalog.data.gov/dataset/fruit-and-vegetable-prices
Search Data.Gov
DATA.GOV
DATA
TOPICS
IMPACT
APPLICATIONS
DEVELOPERS
CONTACT
DATA CATALOG
/ Datasets
Organizations
/ / Department of Agriculture / Economic Research Service, ...
Submit Data Story
Report Data Issue
ERS
ECONOMIC RESEARCH SERVICE
United States Department of Agriculture
Federal
Download

Economic Research Service

Topics

Agriculture

Climate

Publisher

Economic Research Service, U.S. Department of Agriculture

Contact

Hayden Stewart

## Fruit and Vegetable Prices

Metadata Updated: May 2, 2019

How much do fruits and vegetables cost? ERS estimated average prices for 153 commonly consumed fresh and processed fruits and vegetables.

### Access & Use Information

Public: This dataset is intended for public access and use.

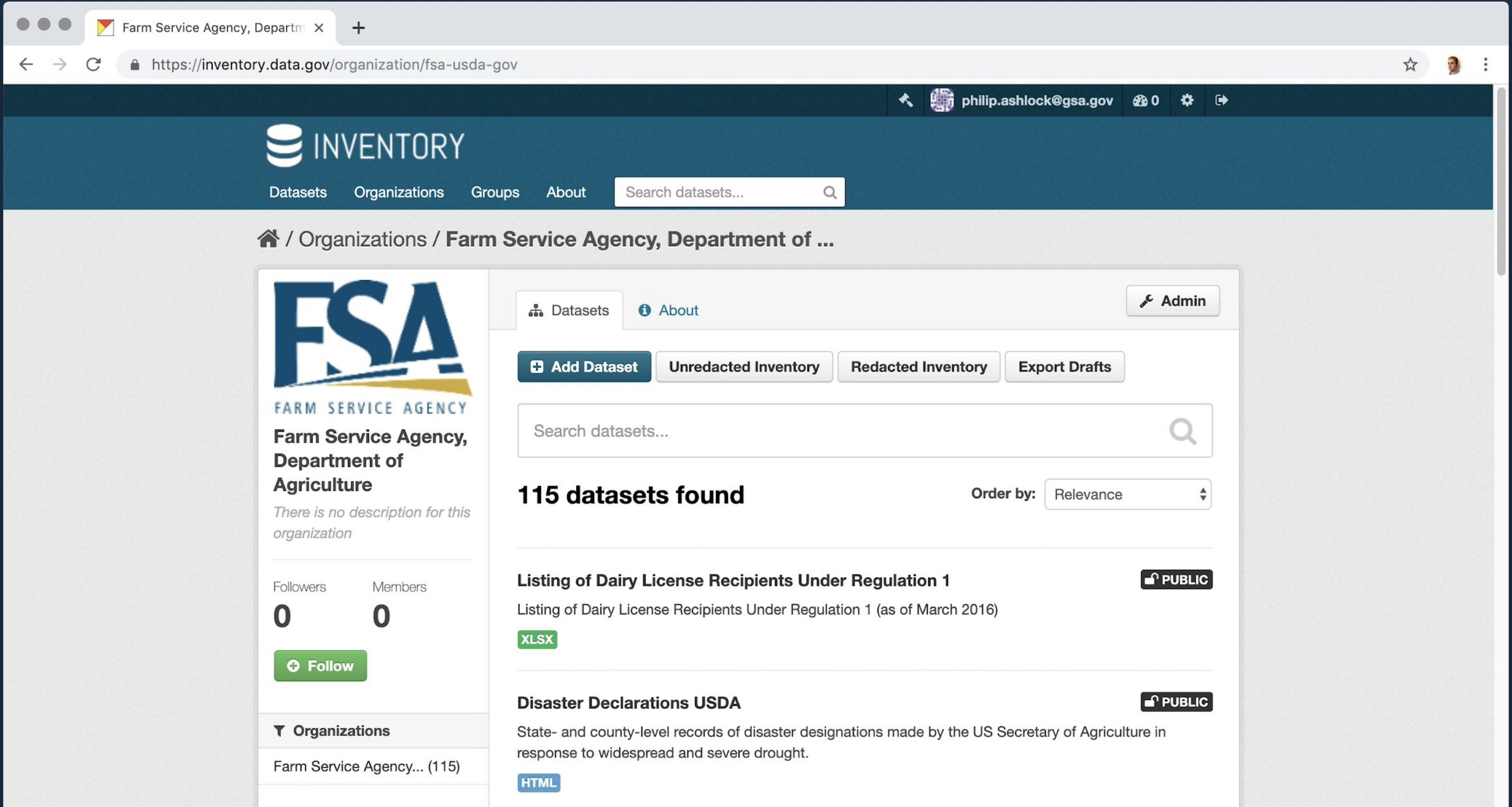
License: Creative Commons CCZero

### Downloads & Resources

Web page with links to Excel files 9470 views

fruit-and-vegetable-prices.aspx

# Metadata management SaaS - inventory.data.gov



/ Organizations / **Farm Service Agency, Department of ...**

FARM SERVICE AGENCY

**Farm Service Agency, Department of Agriculture**

*There is no description for this organization*

| Followers | Members |
|---|---|
| 0 | 0 |

**Organizations**

Farm Service Agency... (115)

**115 datasets found**

**Order by:** Relevance

**Listing of Dairy License Recipients Under Regulation 1** PUBLIC

Listing of Dairy License Recipients Under Regulation 1 (as of March 2016)

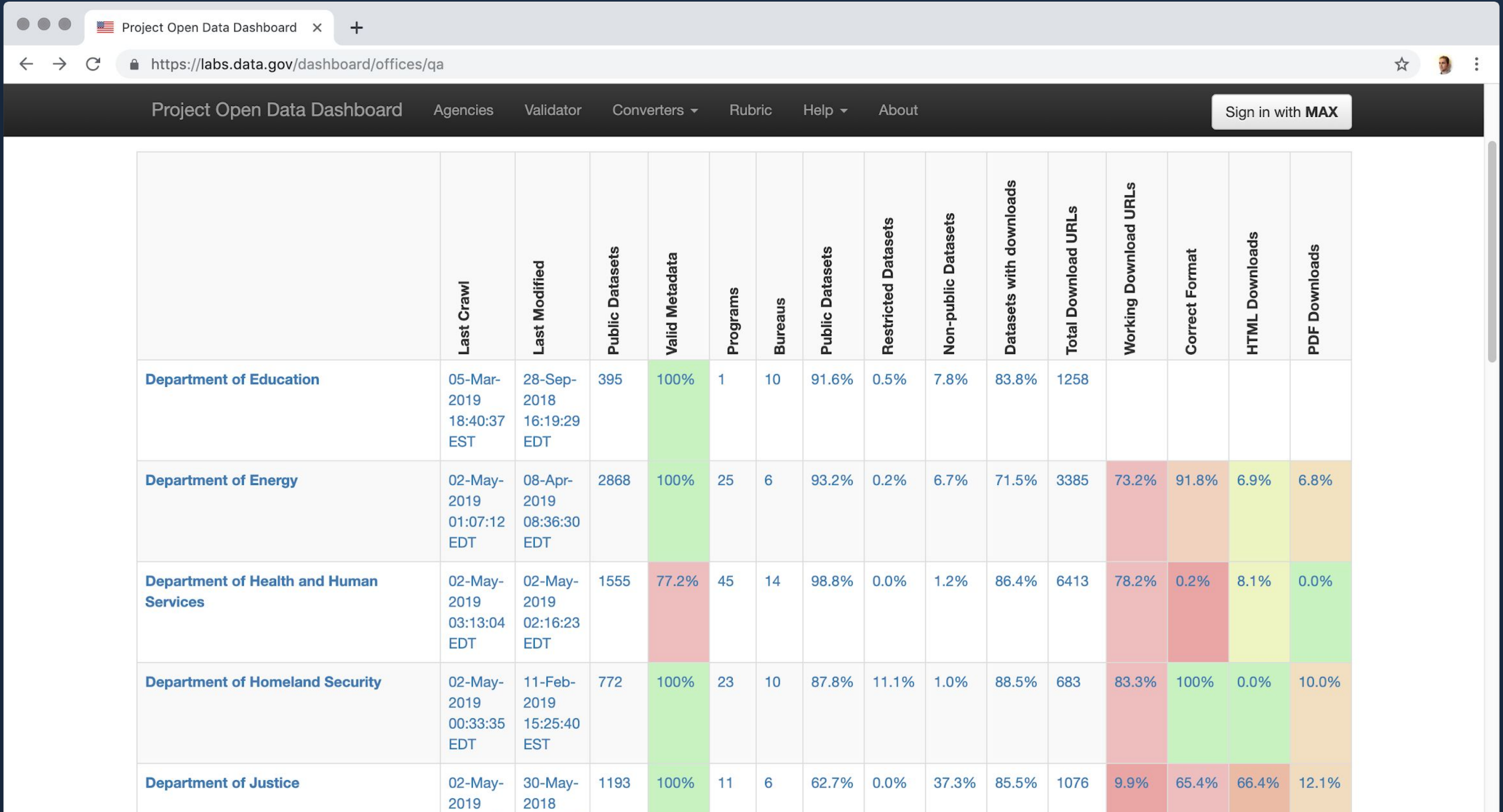
XLSX

**Disaster Declarations USDA** PUBLIC

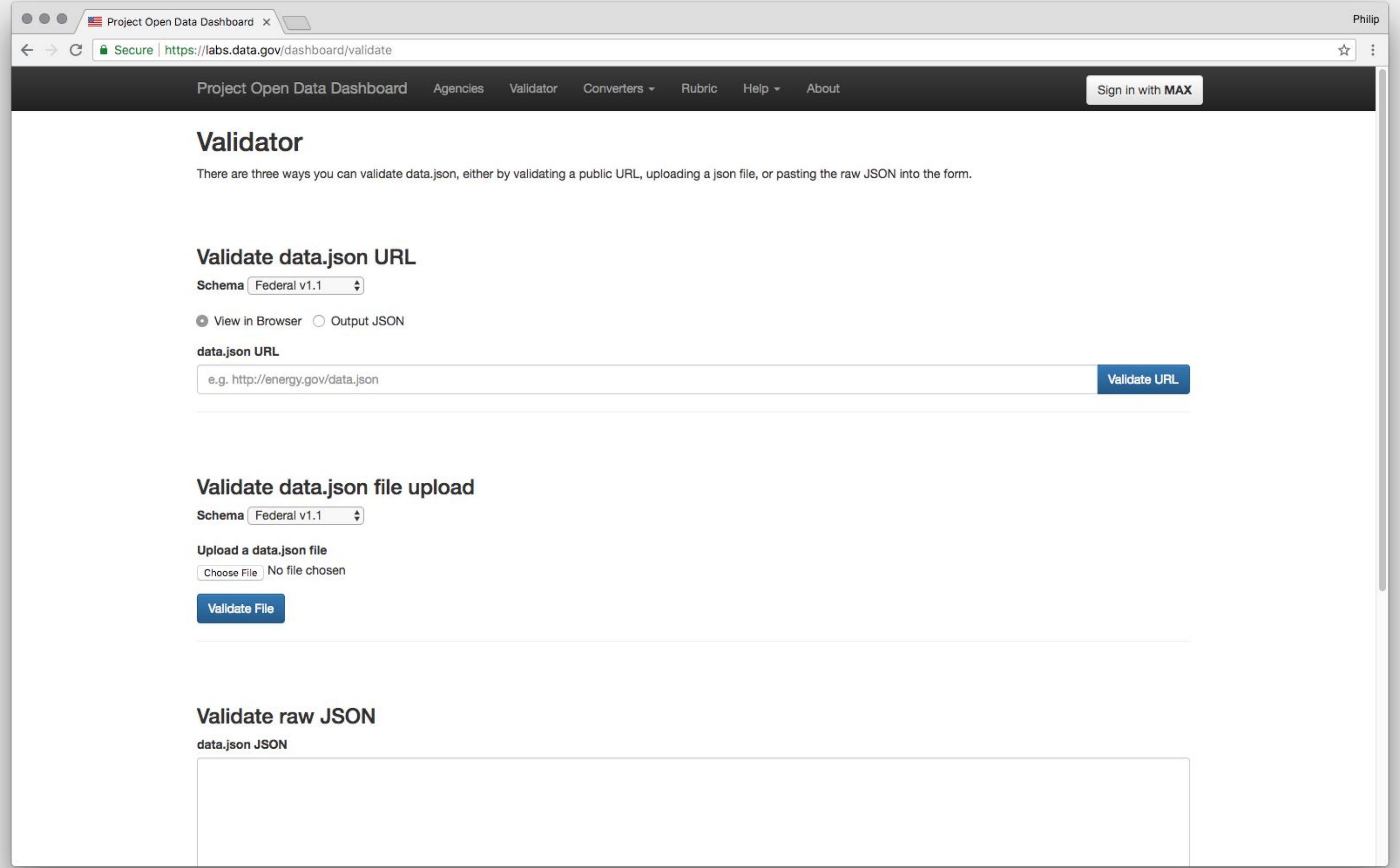
State- and county-level records of disaster designations made by the US Secretary of Agriculture in response to widespread and severe drought.

HTML

# Project Open Data Dashboard


Project Open Data Dashboard
https://labs.data.gov/dashboard/offices/qa
Project Open Data Dashboard
Agencies
Validator
Converters
Rubric
Help
About
Sign in with MAX

| | Last Crawl | Last Modified | Public Datasets | Valid Metadata | Programs | Bureaus | Public Datasets | Restricted Datasets | Non-public Datasets | Datasets with downloads | Total Download URLs | Working Download URLs | Correct Format | HTML Downloads | PDF Downloads |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Department of Education** | 05-Mar-2019 18:40:37 EST | 28-Sep-2018 16:19:29 EDT | 395 | 100% | 1 | 10 | 91.6% | 0.5% | 7.8% | 83.8% | 1258 | | | | |
| **Department of Energy** | 02-May-2019 01:07:12 EDT | 08-Apr-2019 08:36:30 EDT | 2868 | 100% | 25 | 6 | 93.2% | 0.2% | 6.7% | 71.5% | 3385 | 73.2% | 91.8% | 6.9% | 6.8% |
| **Department of Health and Human Services** | 02-May-2019 03:13:04 EDT | 02-May-2019 02:16:23 EDT | 1555 | 77.2% | 45 | 14 | 98.8% | 0.0% | 1.2% | 86.4% | 6413 | 78.2% | 0.2% | 8.1% | 0.0% |
| **Department of Homeland Security** | 02-May-2019 00:33:35 EDT | 11-Feb-2019 15:25:40 EST | 772 | 100% | 23 | 10 | 87.8% | 11.1% | 1.0% | 88.5% | 683 | 83.3% | 100% | 0.0% | 10.0% |
| **Department of Justice** | 02-May-2019 | 30-May-2018 | 1193 | 100% | 11 | 6 | 62.7% | 0.0% | 37.3% | 85.5% | 1076 | 9.9% | 65.4% | 66.4% | 12.1% |



Project Open Data Dashboard
Philip
Secure | https://labs.data.gov/dashboard/validate
Project Open Data Dashboard
Agencies
Validator
Converters
Rubric
Help
About
Sign in with MAX

# Validator

There are three ways you can validate data.json, either by validating a public URL, uploading a json file, or pasting the raw JSON into the form.

## Validate data.json URL

**Schema** Federal v1.1

● View in Browser ○ Output JSON

**data.json URL**

e.g. http://energy.gov/data.json

## Validate data.json file upload

**Schema** Federal v1.1

**Upload a data.json file**

Choose File No file chosen

## Validate raw JSON

**data.json JSON**

# Open Data & Data Management Policies

- Foundations for Evidence-Based Policymaking Act / OPEN Government Data Act (2019)
- Open Data Policy (M-13-13 / 2013)
- Geospatial Data Act / Coordination of Geographic Information and Related Spatial Data Activities (OMB A-16 / 2018)
- Executive Order on Maintaining American Leadership in Artificial Intelligence (2019)
- Open Data CAP Goal, GPRA / Federal Data Strategy (2013-2017, 2018-2020)

# Data.gov Responsibilities in OPEN Gov Data Act

## Federal Data Catalog

IN GENERAL.—The Administrator of General Services shall maintain a single public interface online as a point of entry dedicated to sharing agency data assets with the public, which shall be known as the 'Federal data catalogue'. The Administrator and the Director shall ensure that agencies can submit public data assets, or links to public data assets, for publication and public availability on the interface.

## Repository

The Director shall collaborate with the Office of Government Information Services and the Administrator of General Services to develop and maintain an **online repository of tools, best practices, and schema standards** to facilitate the adoption of open data practices across the Federal Government, which shall

"(A) include any definitions, regulations, policies, checklists, and case studies related to open data policy;

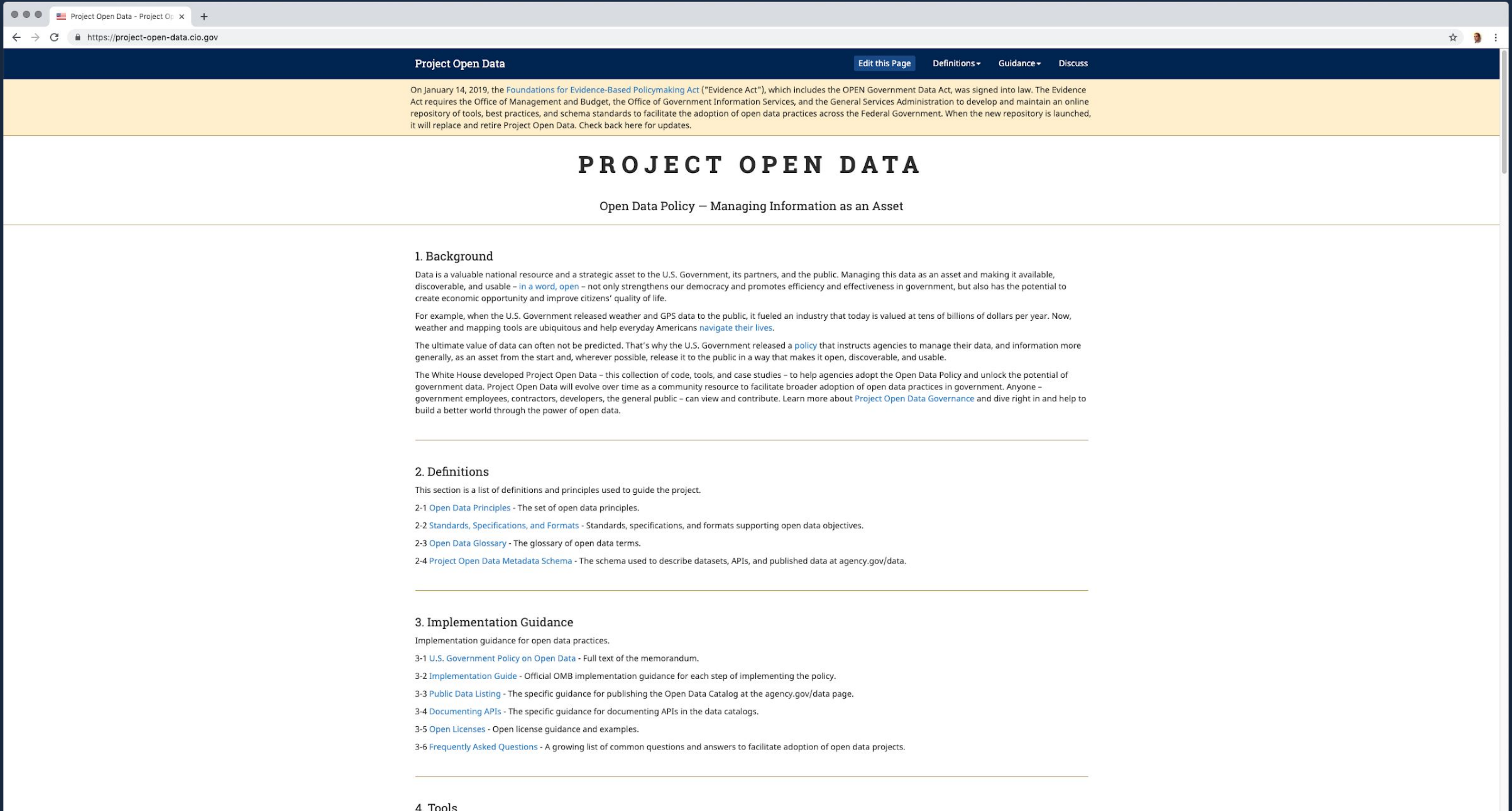
"(B) facilitate collaboration and the adoption of best practices across the Federal Government relating to the adoption of open data practices; and

"(C) be made available on the Federal data catalogue maintained under paragraph (1).

# Project Open Data Metadata Schema

“DCAT-US”

# Project Open Data - project-open-data.cio.gov


Project Open Data - Project Op
https://project-open-data.cio.gov
Project Open Data
Edit this Page
Definitions
Guidance
Discuss

On January 14, 2019, the Foundations for Evidence-Based Policymaking Act ("Evidence Act"), which includes the OPEN Government Data Act, was signed into law. The Evidence Act requires the Office of Management and Budget, the Office of Government Information Services, and the General Services Administration to develop and maintain an online repository of tools, best practices, and schema standards to facilitate the adoption of open data practices across the Federal Government. When the new repository is launched, it will replace and retire Project Open Data. Check back here for updates.

## 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.

### 3. Implementation Guidance

Implementation guidance for open data practices.

3-1 U.S. Government Policy on Open Data - Full text of the memorandum.

3-2 Implementation Guide - Official OMB implementation guidance for each step of implementing the policy.

3-3 Public Data Listing - The specific guidance for publishing the Open Data Catalog at the agency.gov/data page.

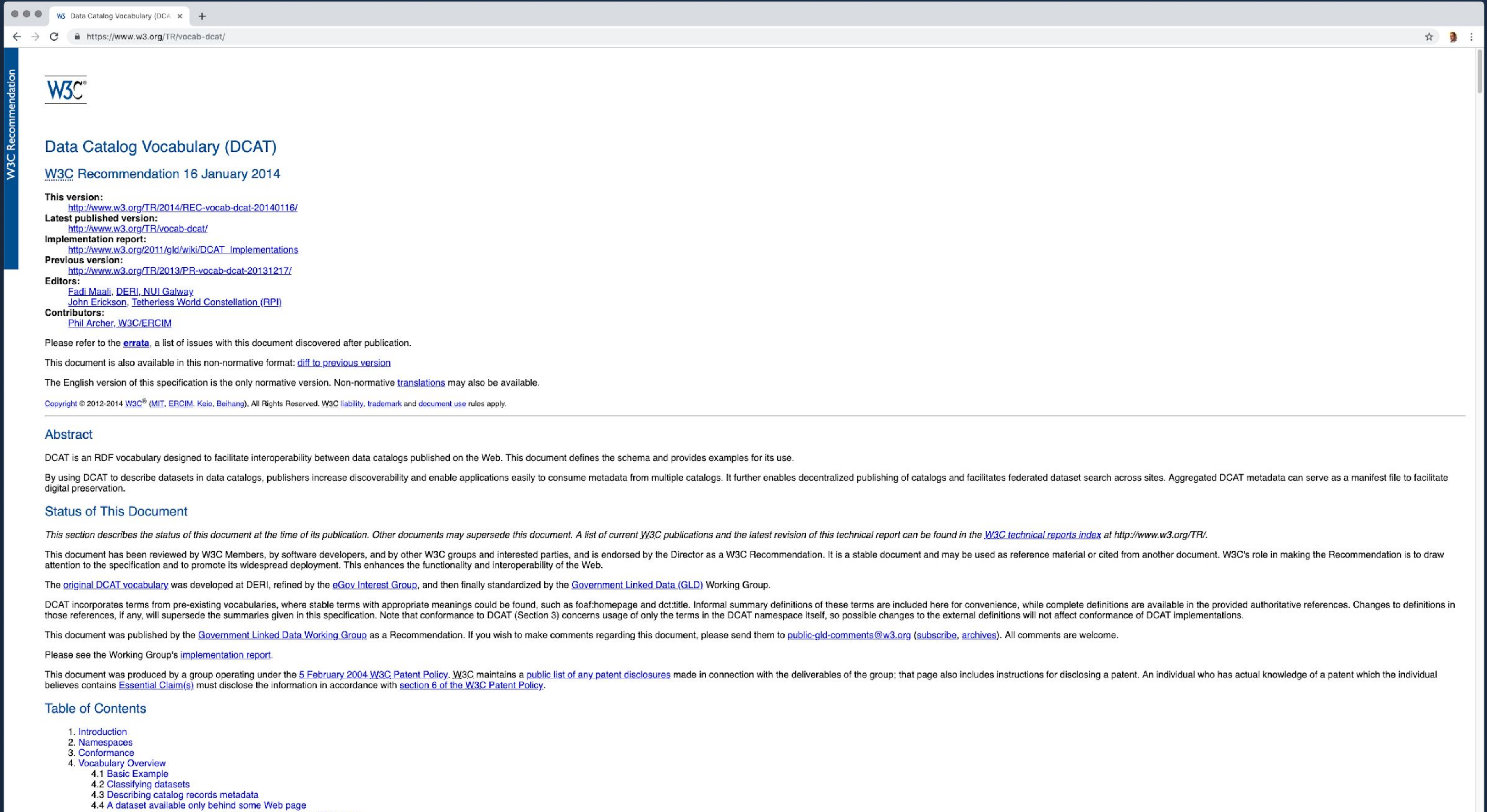
3-4 Documenting APIs - The specific guidance for documenting APIs in the data catalogs.

3-5 Open Licenses - Open license guidance and examples.

3-6 Frequently Asked Questions - A growing list of common questions and answers to facilitate adoption of open data projects.

### 4. Tools

# DCAT - w3.org/TR/vocab-dcat/


W3 Data Catalog Vocabulary (DCA
https://www.w3.org/TR/vocab-dcat/
W3C Recommendation

## 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.

### Abstract

DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. This document defines the schema and provides examples for its use.

By using DCAT to describe datasets in data catalogs, publishers increase discoverability and enable applications easily to consume metadata from multiple catalogs. It further enables decentralized publishing of catalogs and facilitates federated dataset search across sites. Aggregated DCAT metadata can serve as a manifest file to facilitate digital preservation.

### Status of This Document

*This section describes the status of this document at the time of its publication. Other documents may supersede this document. A list of current W3C publications and the latest revision of this technical report can be found in the W3C technical reports index at http://www.w3.org/TR/.*

This document has been reviewed by W3C Members, by software developers, and by other W3C groups and interested parties, and is endorsed by the Director as a W3C Recommendation. It is a stable document and may be used as reference material or cited from another document. W3C's role in making the Recommendation is to draw attention to the specification and to promote its widespread deployment. This enhances the functionality and interoperability of the Web.

The original DCAT vocabulary was developed at DERI, refined by the eGov Interest Group, and then finally standardized by the Government Linked Data (GLD) Working Group.

DCAT incorporates terms from pre-existing vocabularies, where stable terms with appropriate meanings could be found, such as foaf:homepage and dct:title. Informal summary definitions of these terms are included here for convenience, while complete definitions are available in the provided authoritative references. Changes to definitions in those references, if any, will supersede the summaries given in this specification. Note that conformance to DCAT (Section 3) concerns usage of only the terms in the DCAT namespace itself, so possible changes to the external definitions will not affect conformance of DCAT implementations.

This document was published by the Government Linked Data Working Group as a Recommendation. If you wish to make comments regarding this document, please send them to public-gld-comments@w3.org (subscribe, archives). All comments are welcome.

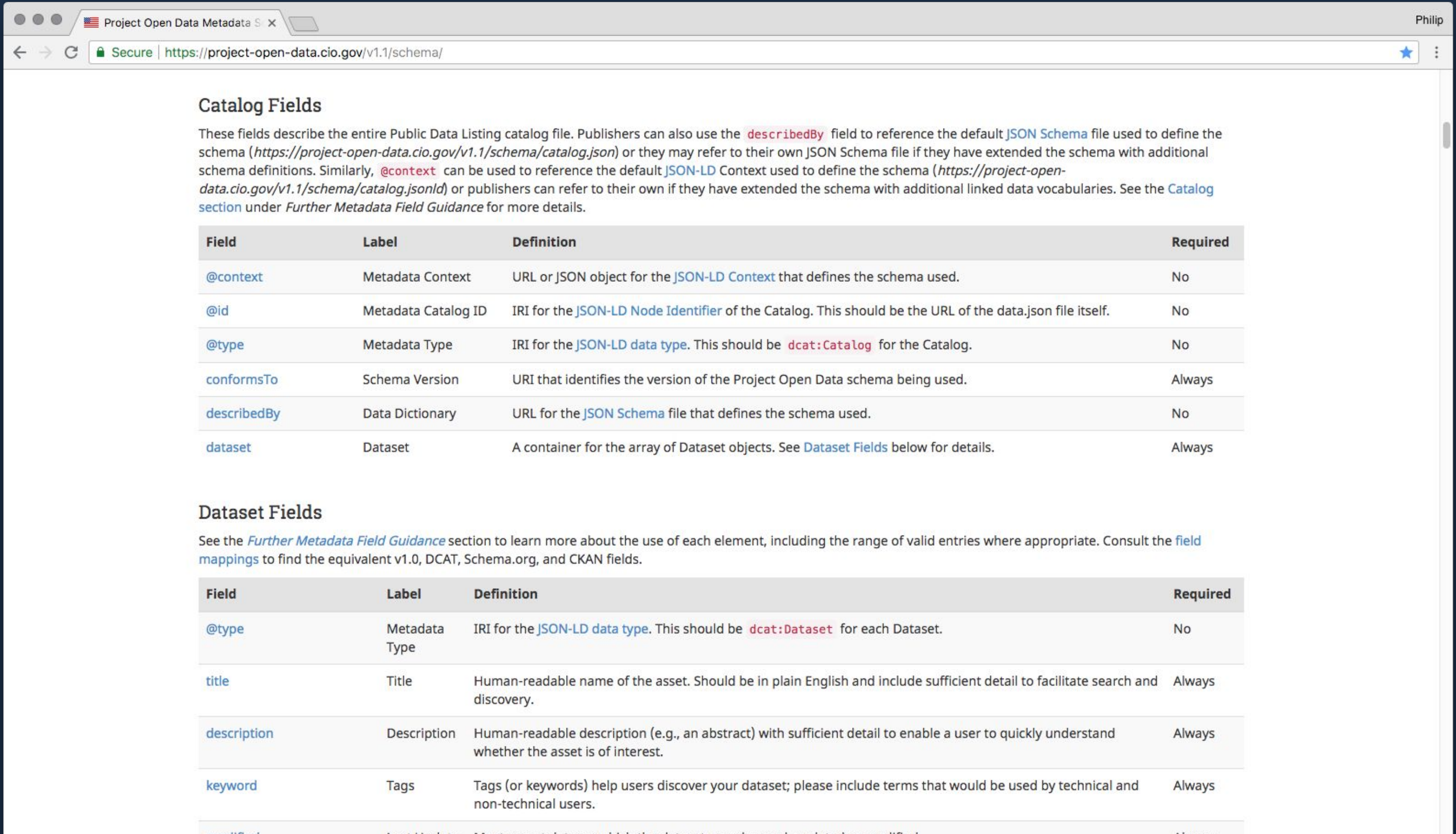
Please see the Working Group's implementation report.

This document was produced by a group operating under the 5 February 2004 W3C Patent Policy. W3C maintains a public list of any patent disclosures made in connection with the deliverables of the group; that page also includes instructions for disclosing a patent. An individual who has actual knowledge of a patent which the individual believes contains Essential Claim(s) must disclose the information in accordance with section 6 of the W3C Patent Policy.

### Table of Contents

1. Introduction
2. Namespaces
3. Conformance
4. Vocabulary Overview
    4.1 Basic Example
    4.2 Classifying datasets
    4.3 Describing catalog records metadata
    4.4 A dataset available only behind some Web page

# Project Open Data Metadata Schema v1.1



Project Open Data Metadata S ×
Philip
Secure | https://project-open-data.cio.gov/v1.1/schema/

## Catalog Fields

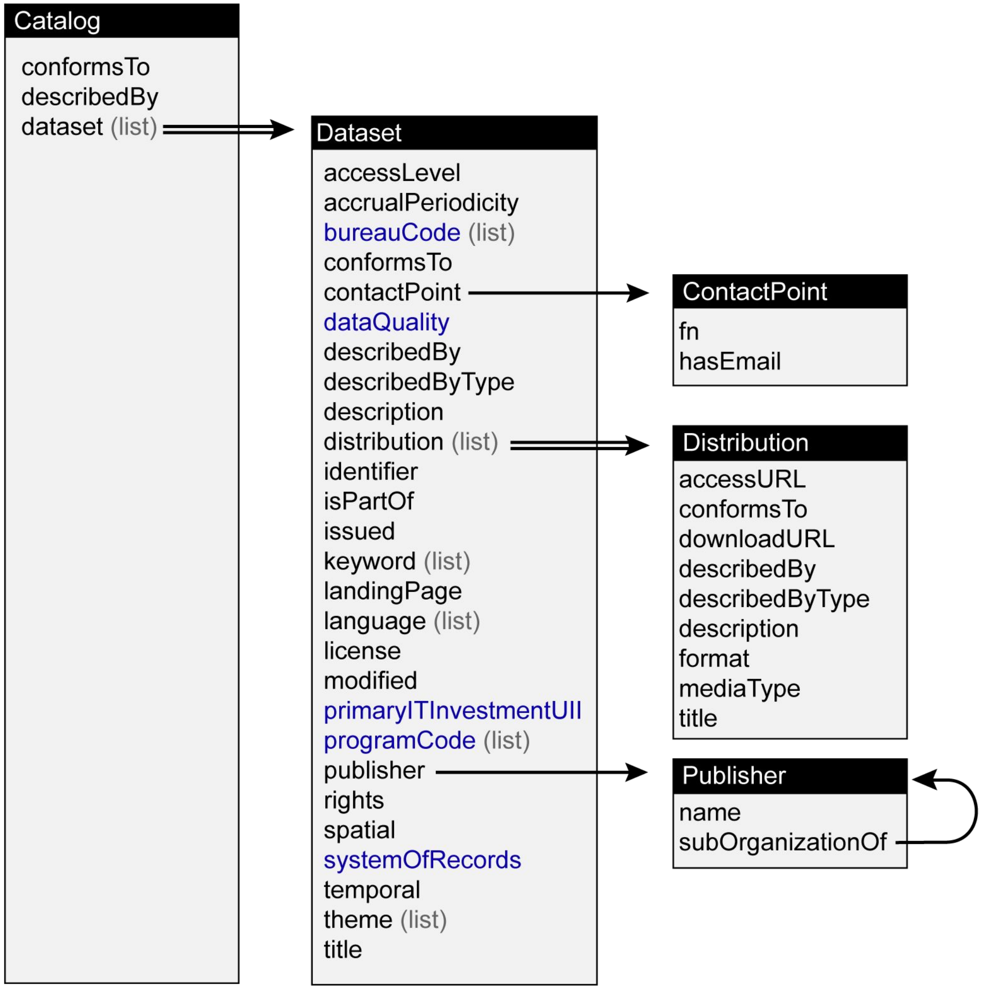
These fields describe the entire Public Data Listing catalog file. Publishers can also use the `describedBy` field to reference the default JSON Schema file used to define the schema (*https://project-open-data.cio.gov/v1.1/schema/catalog.json*) or they may refer to their own JSON Schema file if they have extended the schema with additional schema definitions. Similarly, `@context` can be used to reference the default JSON-LD Context used to define the schema (*https://project-open-data.cio.gov/v1.1/schema/catalog.jsonld*) or publishers can refer to their own if they have extended the schema with additional linked data vocabularies. See the Catalog section under *Further Metadata Field Guidance* for more details.

| Field | Label | Definition | Required |
|---|---|---|---|
| @context | Metadata Context | URL or JSON object for the JSON-LD Context that defines the schema used. | No |
| @id | Metadata Catalog ID | IRI for the JSON-LD Node Identifier of the Catalog. This should be the URL of the data.json file itself. | No |
| @type | Metadata Type | IRI for the JSON-LD data type. This should be `dcat:Catalog` for the Catalog. | No |
| conformsTo | Schema Version | URI that identifies the version of the Project Open Data schema being used. | Always |
| describedBy | Data Dictionary | URL for the JSON Schema file that defines the schema used. | No |
| dataset | Dataset | A container for the array of Dataset objects. See Dataset Fields below for details. | Always |

## Dataset Fields

See the *Further Metadata Field Guidance* section to learn more about the use of each element, including the range of valid entries where appropriate. Consult the field mappings to find the equivalent v1.0, DCAT, Schema.org, and CKAN fields.

| Field | Label | Definition | Required |
|---|---|---|---|
| @type | Metadata Type | IRI for the JSON-LD data type. This should be `dcat:Dataset` for each Dataset. | No |
| 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 |



Catalog

conformsTo
describedBy
dataset (list)

Dataset

accessLevel
accrualPeriodicity
bureauCode (list)
conformsTo
contactPoint
dataQuality
describedBy
describedByType
description
distribution (list)
identifier
isPartOf
issued
keyword (list)
landingPage
language (list)
license
modified
primaryITInvestmentUII
programCode (list)
publisher
rights
spatial
systemOfRecords
temporal
theme (list)
title

ContactPoint

fn
hasEmail

Distribution

accessURL
conformsTo
downloadURL
describedBy
describedByType
description
format
mediaType
title

Publisher

name
subOrganizationOf

# US Government-Specific Metadata Elements

## dataQuality

- Whether the dataset meets the agency's Information Quality Guidelines. Required if Applicable

## systemOfRecords

- 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. Not Required

## primaryITInvestmentUII

- For linking a dataset with an IT Unique Investment Identifier (UII). Not Required

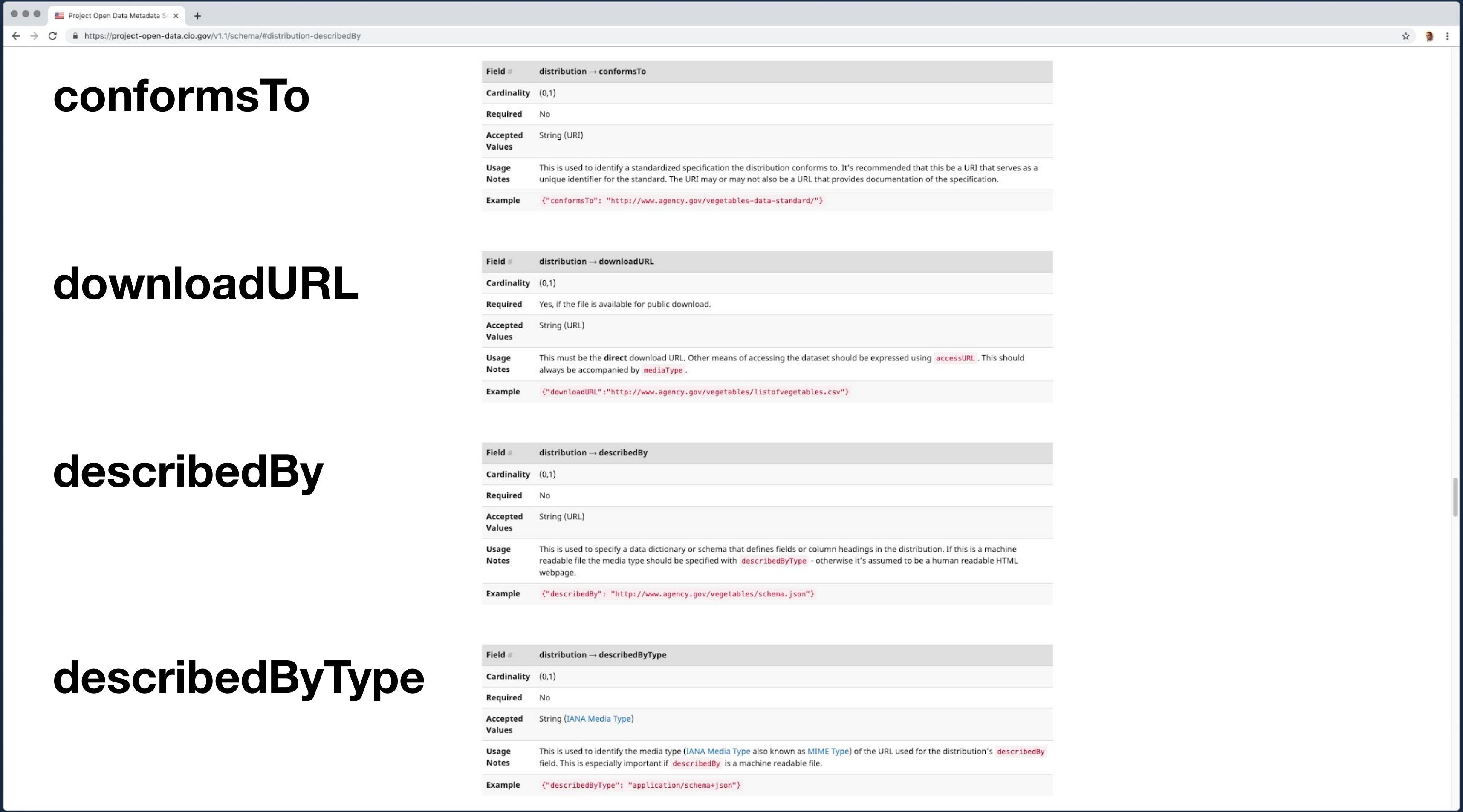
## bureauCode

- Federal agencies, combined agency and bureau code from OMB Circular A-11, Appendix C (PDF, CSV) in the format of 015:11. Required

## programCode

- Federal agencies, list the primary program related to this data asset, from the Federal Program Inventory. Use the format of 015:001. Required

# Extending the Schema to Lower Level Metadata


Project Open Data Metadata S
https://project-open-data.cio.gov/v1.1/schema/#distribution-describedBy

## conformsTo

| Field | distribution → conformsTo |
|---|---|
| Cardinality | (0,1) |
| Required | No |
| Accepted Values | String (URI) |
| Usage Notes | This is used to identify a standardized specification the distribution conforms to. It's recommended that this be a URI that serves as a unique identifier for the standard. The URI may or may not also be a URL that provides documentation of the specification. |
| Example | `{"conformsTo": "http://www.agency.gov/vegetables-data-standard/"}` |

## downloadURL

| Field | distribution → downloadURL |
|---|---|
| Cardinality | (0,1) |
| Required | Yes, if the file is available for public download. |
| Accepted Values | String (URL) |
| Usage Notes | This must be the **direct** download URL. Other means of accessing the dataset should be expressed using `accessURL` . This should always be accompanied by `mediaType` . |
| Example | `{"downloadURL":"http://www.agency.gov/vegetables/listofvegetables.csv"}` |

## describedBy

| Field | distribution → describedBy |
|---|---|
| Cardinality | (0,1) |
| Required | No |
| Accepted Values | String (URL) |
| Usage Notes | This is used to specify a data dictionary or schema that defines fields or column headings in the distribution. If this is a machine readable file the media type should be specified with `describedByType` - otherwise it's assumed to be a human readable HTML webpage. |
| Example | `{"describedBy": "http://www.agency.gov/vegetables/schema.json"}` |

## describedByType

| Field | distribution → describedByType |
|---|---|
| Cardinality | (0,1) |
| Required | No |
| Accepted Values | String (IANA Media Type) |
| Usage Notes | This is used to identify the media type (IANA Media Type also known as MIME Type) of the URL used for the distribution's `describedBy` field. This is especially important if `describedBy` is a machine readable file. |
| Example | `{"describedByType": "application/schema+json"}` |

# Future Evolution & Expansion of Metadata Support

- Updates to DCAT, W3C Working Group thru June 2019
- Updates to Project Open Data Metadata Schema v1.1 (“DCAT-US”)
  - GitHub issue backlog on Project Open Data & Data.gov repositories
  - Potential for improved support for statistical metadata
- Further integration and harmonization across metadata standards: ISO 19115, Schema.org, SDMX, Tabular Data Packages, CSVW
- Better integration of lower level metadata in Data Catalogs, searchability, and cross linking
- Use of DOI’s and better citability
- Integration with other technologies, e.g. DAT, IPFS, Jupyter Notebooks
- Better support for domain specific data standards including automated aggregation and validation

# Some Relevant Communities

- Federal Data Strategy - https://strategy.data.gov
- U.S. Data Federation  - https://federation.data.gov
- Interagency Open Data Working Group
- Interagency Data Exchange Community of Practice
- Interagency Committee on Standards Policy (ICSP)
- W3C Data Exchange Working Group - https://www.w3.org/2017/dxwg/charter
- Open SDG - https://open-sdg.readthedocs.io
- Libraries + Network - https://libraries.network
- The Maintainers - http://themaintainers.org/miii

# Questions?

philip.ashlock@gsa.gov