Panel: The Art & Science of Data Visualization
#IOGDC

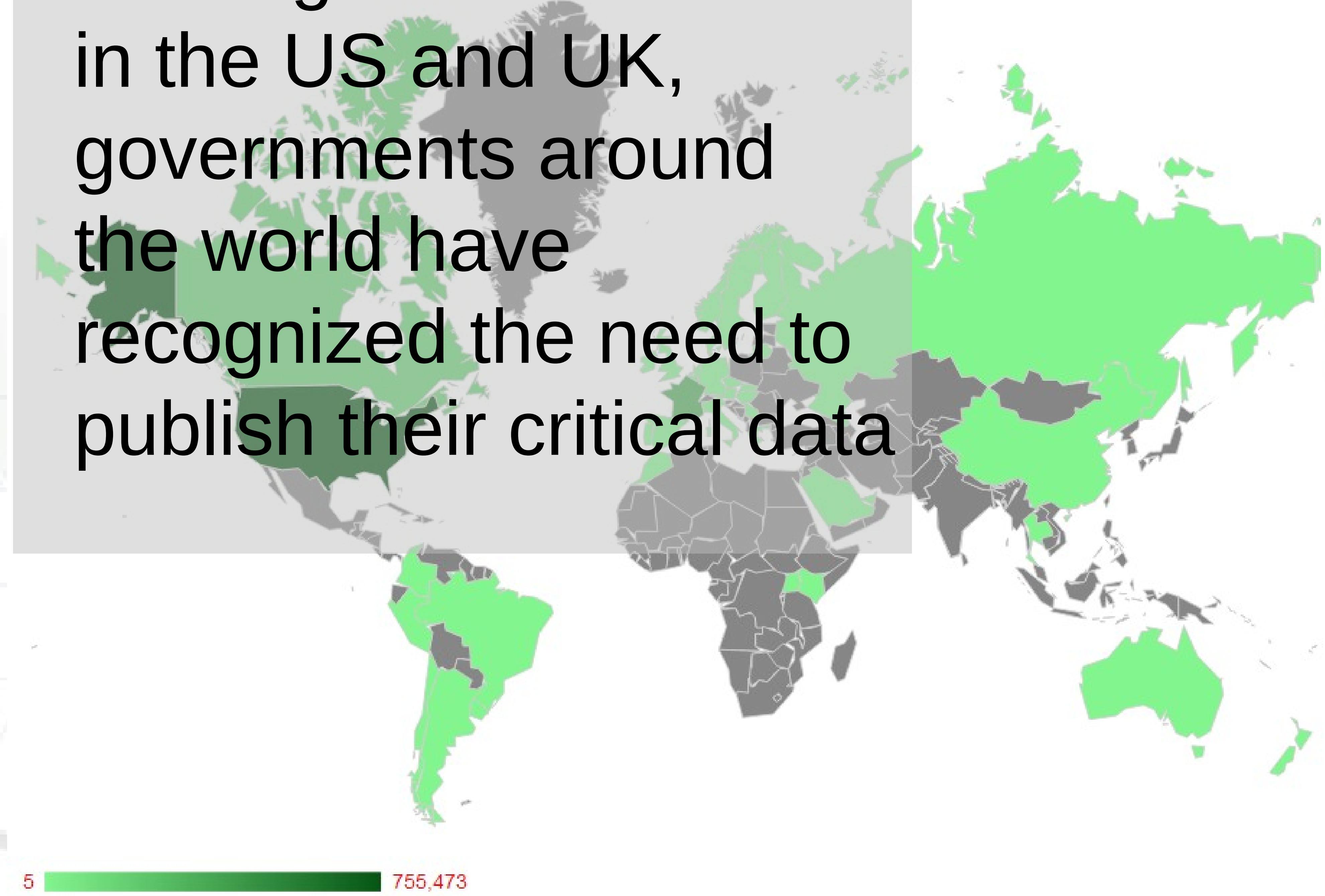
First they have to **find** it:
Getting Government Data Discovered and Used

John S. Erickson, Ph.D.
Tetherless World Constellation
Rensselaer Polytechnic Institute
Troy, New York, USA

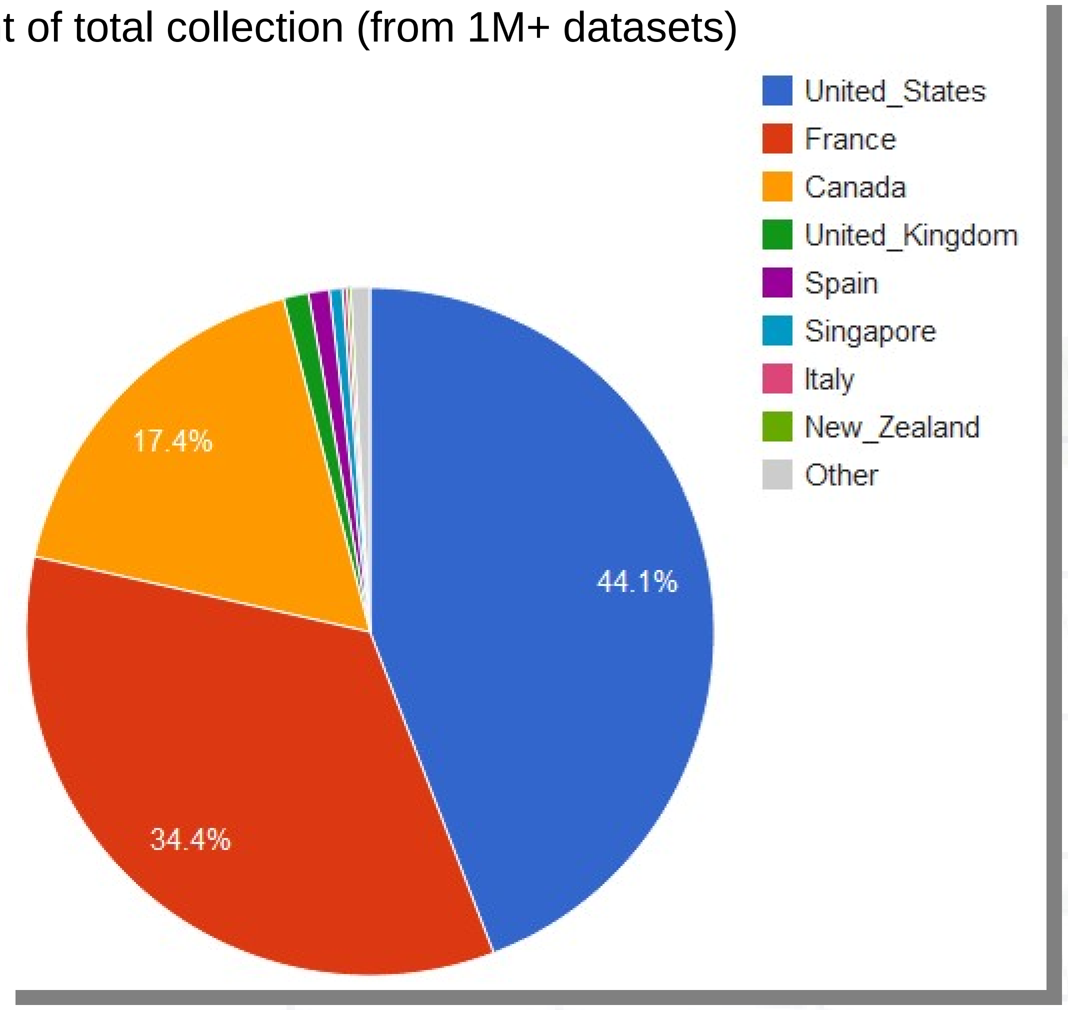
Twitter: **@olyerickson   #TWCRPI**

Open Government Data Around the World

Starting with efforts in the US and UK, governments around the world have recognized the need to publish their critical data



Percent of total collection (from 1M+ datasets)



Diverse Approaches to Open Gov't Data

- Government data initiatives have taken many forms
- GovData portals are widely varied in how they help users **discover** and **use relevant datasets**

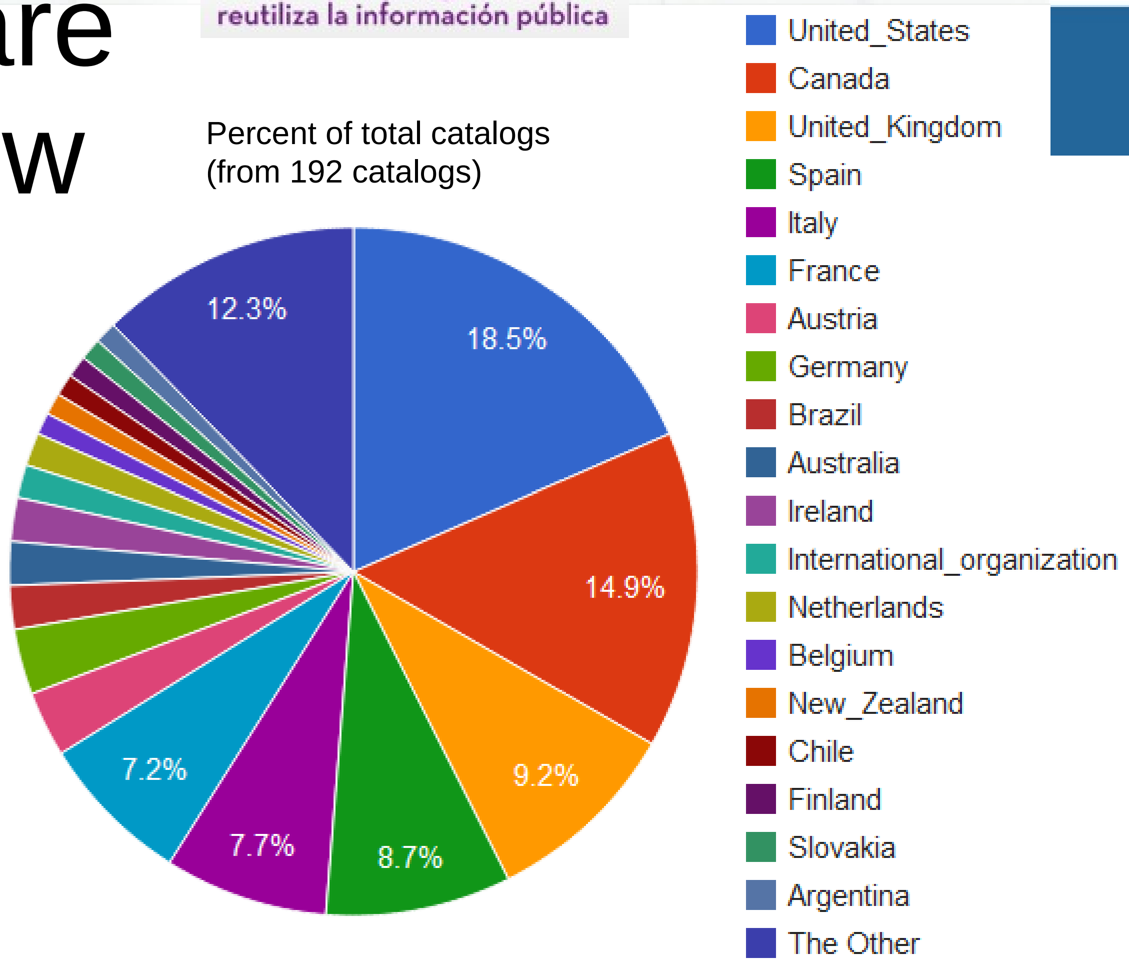
Gobierno de Chile

dati.gov.it
I dati aperti della PA

HM Government
data.gov.uk BETA
Opening up government

datos.gob.es
reutiliza la información pública

Percent of total catalogs
(from 192 catalogs)



Open Data
www.data.gc.ca

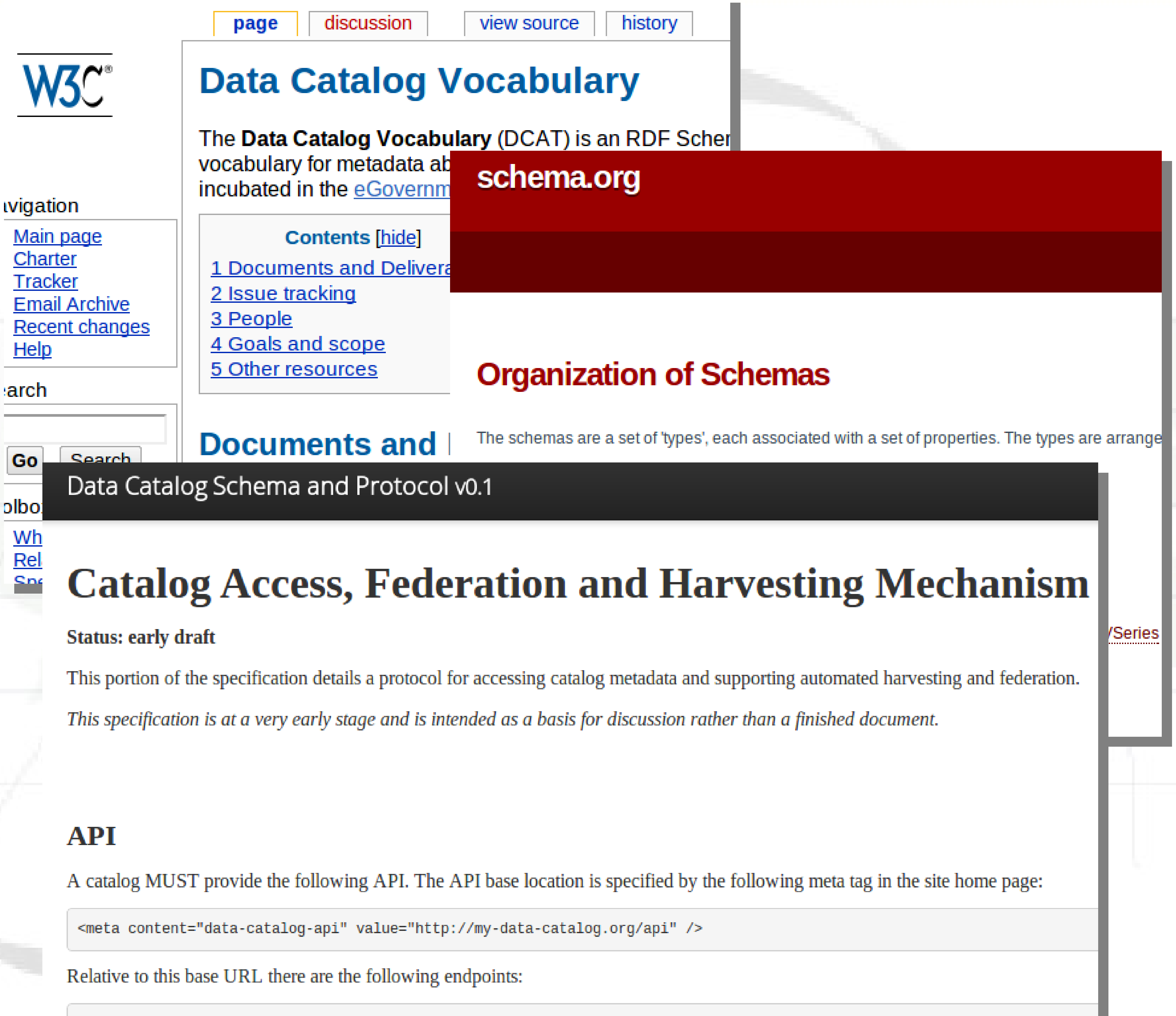
Data.go.kr
공유자원포털

data.gov.sg
discovering data, inspiring ideas

data.gouv.fr BETA
INNOVATION TRANSPARENCE . OUVERTURE

An Official Web Site of the United States Government
DATA.GOV
EMPOWERING PEOPLE

Federated Discovery of Government Data



page | discussion | view source | history

Data Catalog Vocabulary

The **Data Catalog Vocabulary** (DCAT) is an RDF Scher
vocabulary for metadata ab
incubated in the eGovernm

ıvigation

- Main page
- Charter
- Tracker
- Email Archive
- Recent changes
- Help

ːarch

Go | Search

olbo

- Wh
- Rel
- Spe

Contents [hide]

1 Documents and Delivera
2 Issue tracking
3 People
4 Goals and scope
5 Other resources

Documents and

schema.org

Organization of Schemas

The schemas are a set of 'types', each associated with a set of properties. The types are arrange

/Series

Data Catalog Schema and Protocol v0.1

Catalog Access, Federation and Harvesting Mechanism

Status: early draft

This portion of the specification details a protocol for accessing catalog metadata and supporting automated harvesting and federation.

This specification is at a very early stage and is intended as a basis for discussion rather than a finished document.

API

A catalog MUST provide the following API. The API base location is specified by the following meta tag in the site home page:

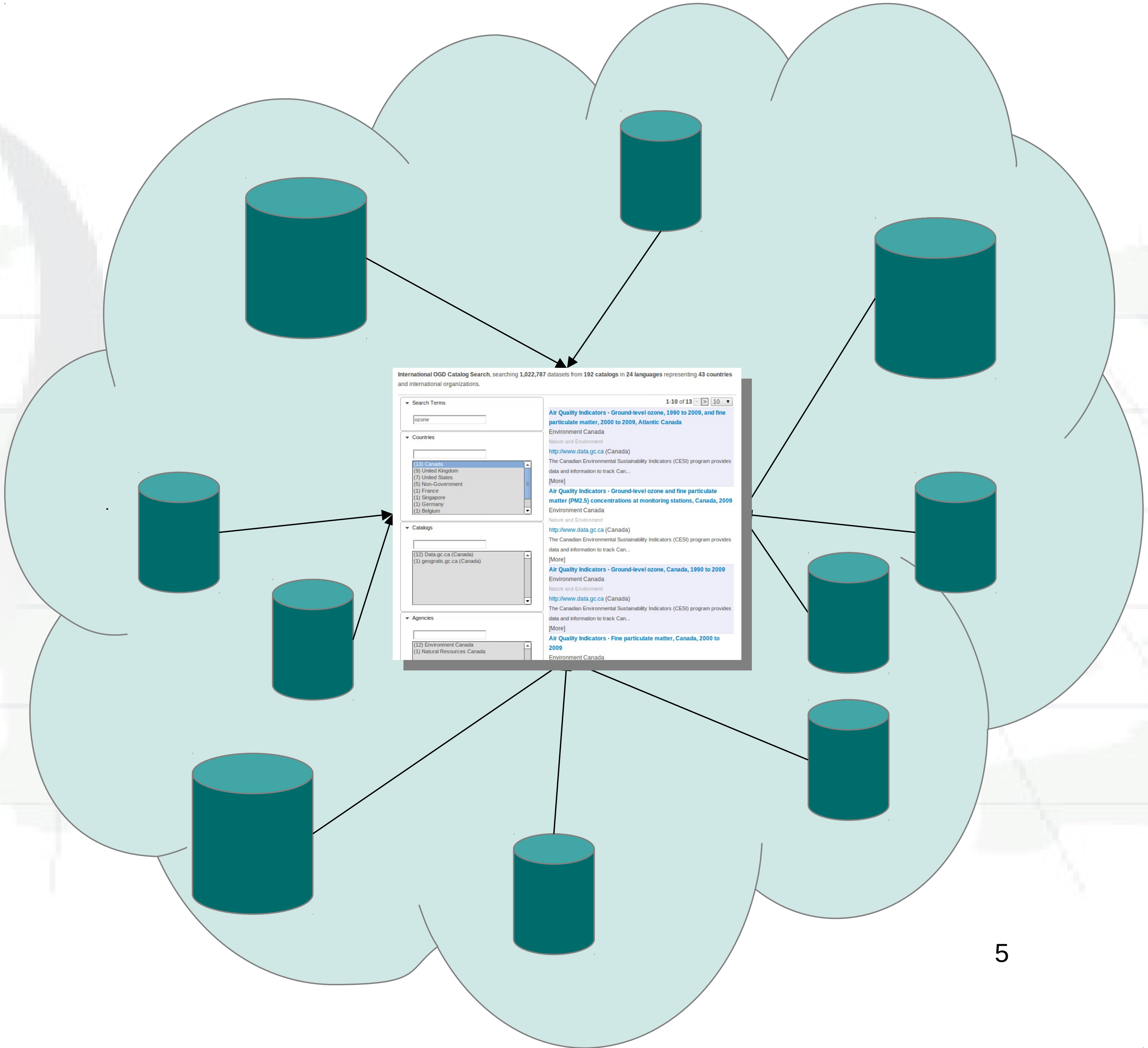
```
<meta content="data-catalog-api" value="http://my-data-catalog.org/api" />
```

Relative to this base URL there are the following endpoints:

Stakeholders have seen the need for **Federated discovery** across catalogs, **especially** from within major search engines including **Bing, Google, Yahoo!** and **Yandex**

Linked Data is Not Enough...

- Publishing open government data as Linked Data is **not enough**
- For OGD to be useful, datasets must be published using metadata, markup standards and presentation that aid **discovery** and **use**



International OGD Catalog Search, searching **1,022,787** datasets from **192 catalogs** in **24 languages** representing **43 countries** and international organizations.

Search Terms

ozone

Countries

(13) Canada
(9) United Kingdom
(7) United States
(5) Non-Government
(1) France
(1) Singapore
(1) Germany
(1) Belgium

Catalogs

(12) Data.gc.ca (Canada)
(1) geogratis.gc.ca (Canada)

Agencies

(12) Environment Canada
(1) Natural Resources Canada

1-10 of **13** < > 10

Air Quality Indicators - Ground-level ozone, 1990 to 2009, and fine particulate matter, 2000 to 2009, Atlantic Canada
Environment Canada
Nature and Environment
http://www.data.gc.ca (Canada)
The Canadian Environmental Sustainability Indicators (CESI) program provides data and information to track Can...
[More]

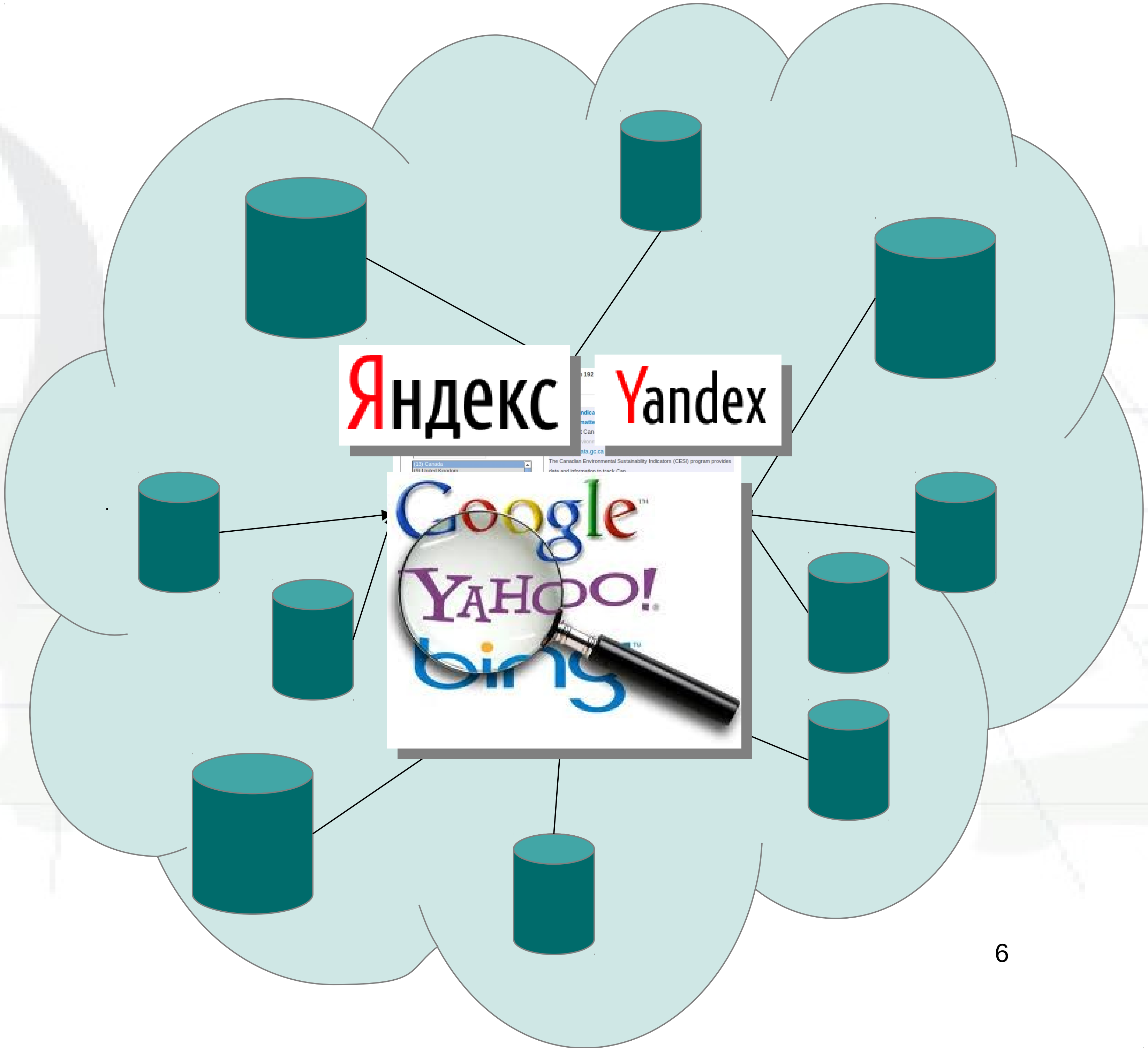
Air Quality Indicators - Ground-level ozone and fine particulate matter (PM2.5) concentrations at monitoring stations, Canada, 2009
Environment Canada
Nature and Environment
http://www.data.gc.ca (Canada)
The Canadian Environmental Sustainability Indicators (CESI) program provides data and information to track Can...
[More]

Air Quality Indicators - Ground-level ozone, Canada, 1990 to 2009
Environment Canada
Nature and Environment
http://www.data.gc.ca (Canada)
The Canadian Environmental Sustainability Indicators (CESI) program provides data and information to track Can...
[More]

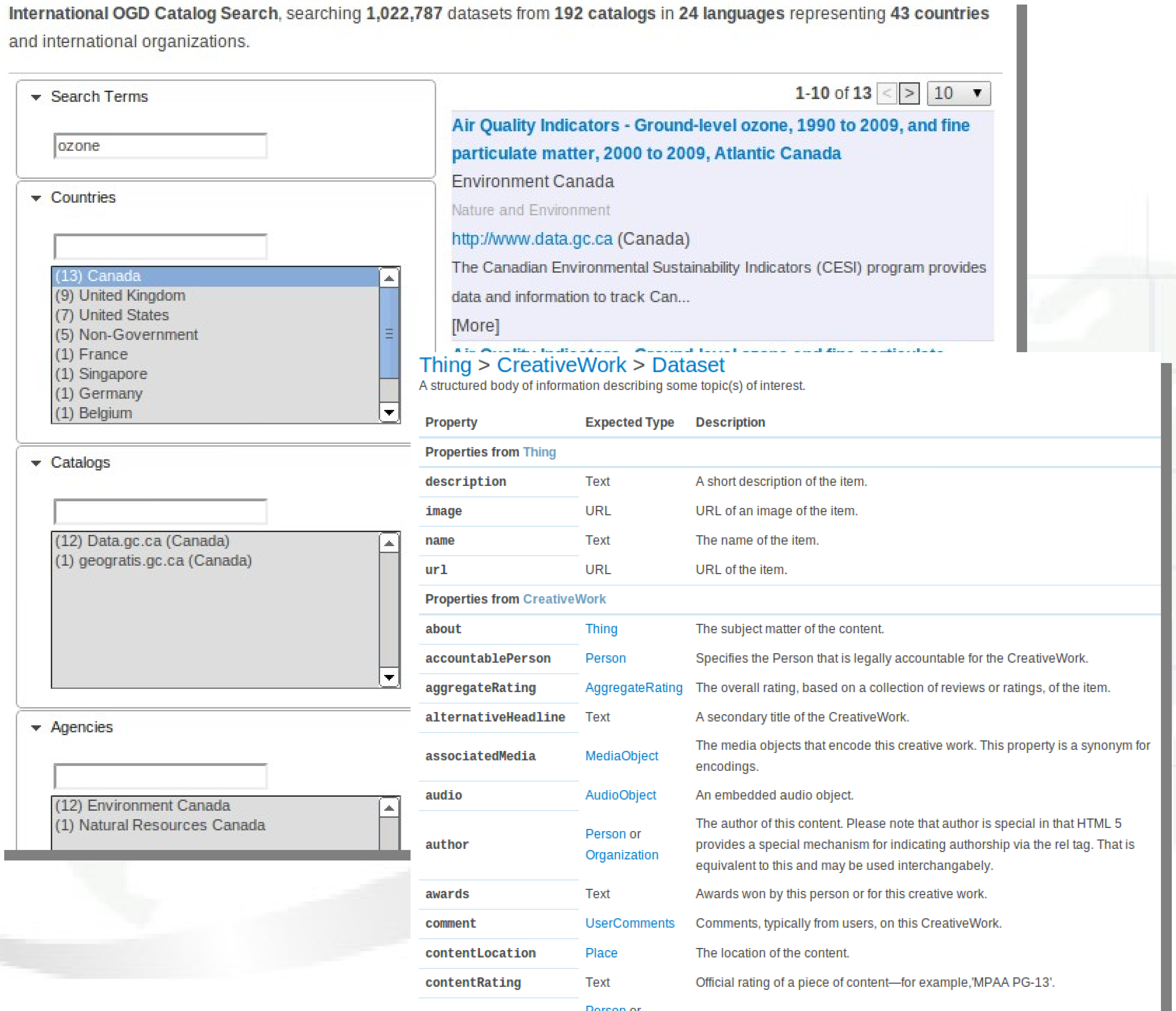
Air Quality Indicators - Fine particulate matter, Canada, 2000 to 2009
Environment Canada

Linked Data is Not Enough...

- Publishing open government data as Linked Data is **not enough**
- For OGD to be useful, datasets must be published using metadata, markup standards and presentation that aid **discovery** and **use**



Dataset Metadata for Discovery and Use



International OGD Catalog Search, searching **1,022,787** datasets from **192 catalogs** in **24 languages** representing **43 countries** and international organizations.

Air Quality Indicators - Ground-level ozone, 1990 to 2009, and fine particulate matter, 2000 to 2009, Atlantic Canada

Environment Canada

Nature and Environment

http://www.data.gc.ca (Canada)

The Canadian Environmental Sustainability Indicators (CESI) program provides data and information to track Can...

[More]

Thing > CreativeWork > Dataset

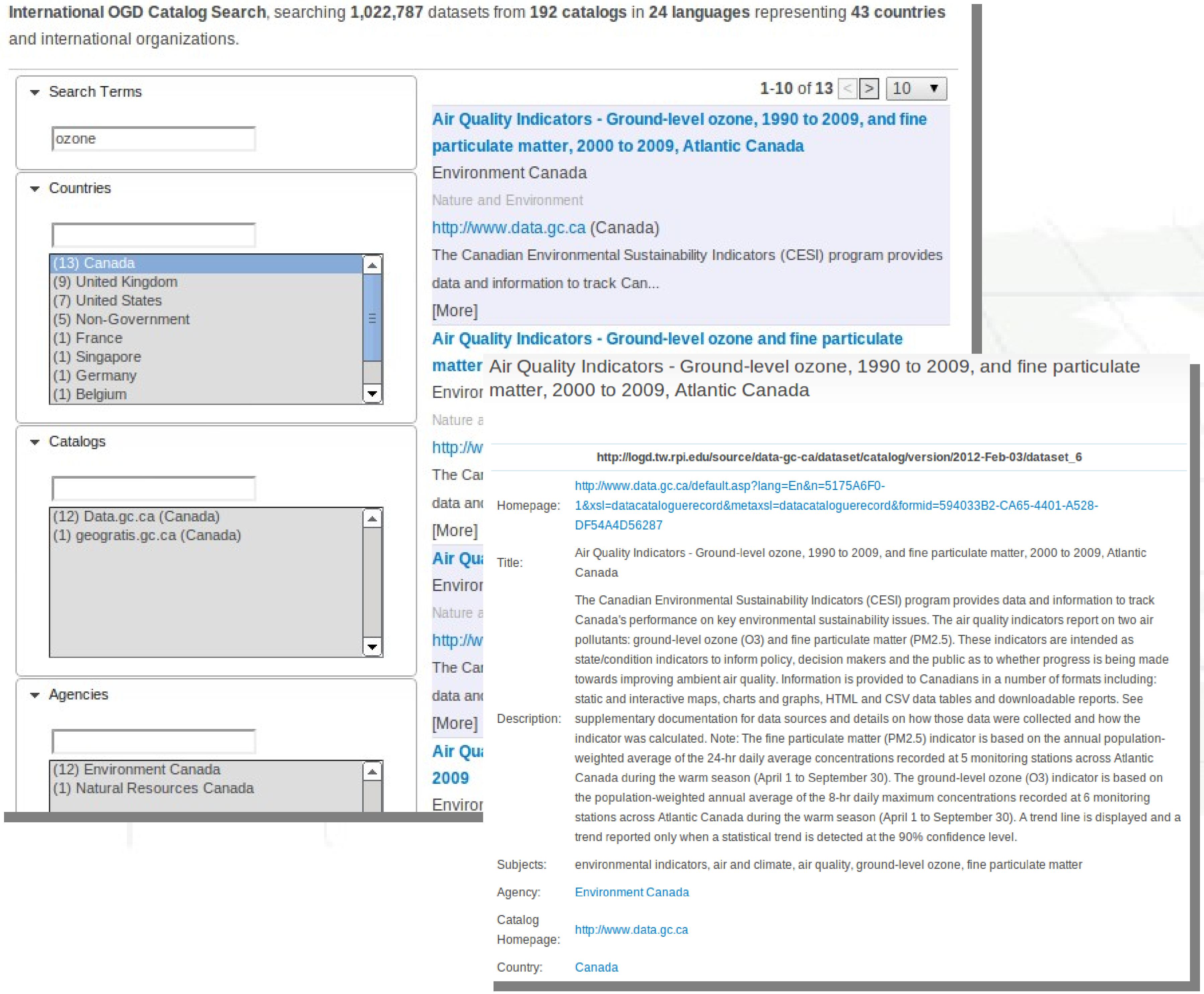
A structured body of information describing some topic(s) of interest.

Property	Expected Type	Description
Properties from Thing		
`description`	Text	A short description of the item.
`image`	URL	URL of an image of the item.
`name`	Text	The name of the item.
`url`	URL	URL of the item.
Properties from CreativeWork		
`about`	Thing	The subject matter of the content.
`accountablePerson`	Person	Specifies the Person that is legally accountable for the CreativeWork.
`aggregateRating`	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item.
`alternativeHeadline`	Text	A secondary title of the CreativeWork.
`associatedMedia`	MediaObject	The media objects that encode this creative work. This property is a synonym for encodings.
`audio`	AudioObject	An embedded audio object.
`author`	Person or Organization	The author of this content. Please note that author is special in that HTML 5 provides a special mechanism for indicating authorship via the rel tag. That is equivalent to this and may be used interchangabely.
`awards`	Text	Awards won by this person or for this creative work.
`comment`	UserComments	Comments, typically from users, on this CreativeWork.
`contentLocation`	Place	The location of the content.
`contentRating`	Text	Official rating of a piece of content—for example,'MPAA PG-13'.
	Person or	

Recent work at TWC RPI demonstrates the value of applying emerging **standards** for **uniformly describing** government datasets and catalogs

International Open Government Dataset Search

TWC's **IOGDS** application is an aggregated catalog of more than 1M datasets from over 192 dataset catalogs from governments at every level around the world



International OGD Catalog Search, searching **1,022,787** datasets from **192 catalogs** in **24 languages** representing **43 countries** and international organizations.

Air Quality Indicators - Ground-level ozone, 1990 to 2009, and fine particulate matter, 2000 to 2009, Atlantic Canada
Environment Canada
Nature and Environment
http://www.data.gc.ca (Canada)
The Canadian Environmental Sustainability Indicators (CESI) program provides data and information to track Can...
[More]

Air Quality Indicators - Ground-level ozone and fine particulate matter
Enviror
Nature a
http://w
The Car
data and
[More]

Air Qua
Enviror
Nature a
http://w
The Car
data and
[More]

Air Qua
2009
Enviror

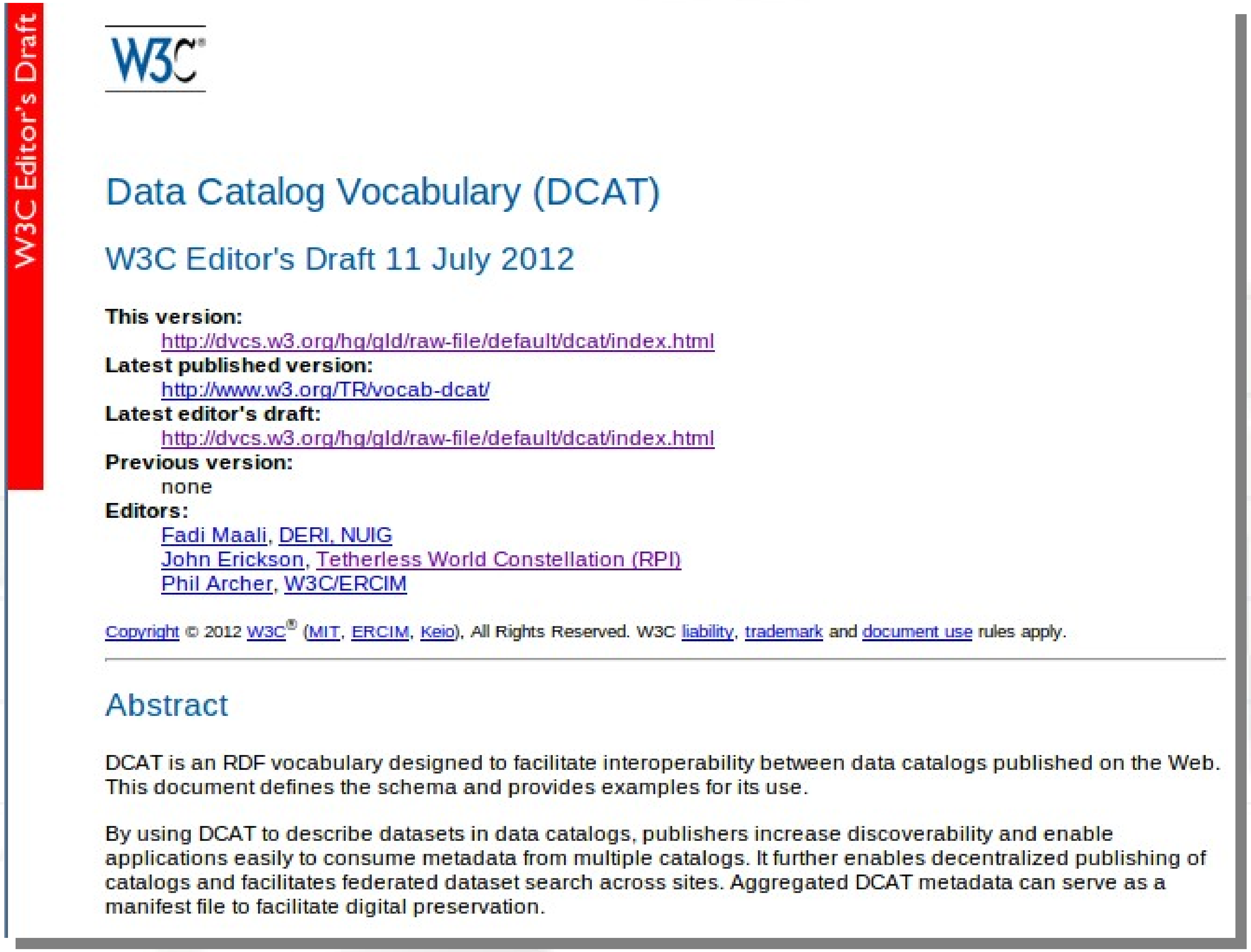
Air Quality Indicators - Ground-level ozone, 1990 to 2009, and fine particulate matter, 2000 to 2009, Atlantic Canada

http://logd.tw.rpi.edu/source/data-gc-ca/dataset/catalog/version/2012-Feb-03/dataset_6

Homepage:	http://www.data.gc.ca/default.asp?lang=En&n=5175A6F0-1&xsl=datacataloguerecord&metaxsl=datacataloguerecord&formid=594033B2-CA65-4401-A528-DF54A4D56287
Title:	Air Quality Indicators - Ground-level ozone, 1990 to 2009, and fine particulate matter, 2000 to 2009, Atlantic Canada
Description:	The Canadian Environmental Sustainability Indicators (CESI) program provides data and information to track Canada's performance on key environmental sustainability issues. The air quality indicators report on two air pollutants: ground-level ozone (O3) and fine particulate matter (PM2.5). These indicators are intended as state/condition indicators to inform policy, decision makers and the public as to whether progress is being made towards improving ambient air quality. Information is provided to Canadians in a number of formats including: static and interactive maps, charts and graphs, HTML and CSV data tables and downloadable reports. See supplementary documentation for data sources and details on how those data were collected and how the indicator was calculated. Note: The fine particulate matter (PM2.5) indicator is based on the annual population-weighted average of the 24-hr daily average concentrations recorded at 5 monitoring stations across Atlantic Canada during the warm season (April 1 to September 30). The ground-level ozone (O3) indicator is based on the population-weighted annual average of the 8-hr daily maximum concentrations recorded at 6 monitoring stations across Atlantic Canada during the warm season (April 1 to September 30). A trend line is displayed and a trend reported only when a statistical trend is detected at the 90% confidence level.
Subjects:	environmental indicators, air and climate, air quality, ground-level ozone, fine particulate matter
Agency:	Environment Canada
Catalog Homepage:	http://www.data.gc.ca
Country:	Canada

See: http://logd.tw.rpi.edu

International Open Government Dataset Search


W3C Editor's Draft

Data Catalog Vocabulary (DCAT)

W3C Editor's Draft 11 July 2012

This version:
http://dvcs.w3.org/hg/gld/raw-file/default/dcat/index.html
Latest published version:
http://www.w3.org/TR/vocab-dcat/
Latest editor's draft:
http://dvcs.w3.org/hg/gld/raw-file/default/dcat/index.html
Previous version:
none
Editors:
Fadi Maali, DERI, NUIG
John Erickson, Tetherless World Constellation (RPI)
Phil Archer, W3C/ERCIM

Abstract

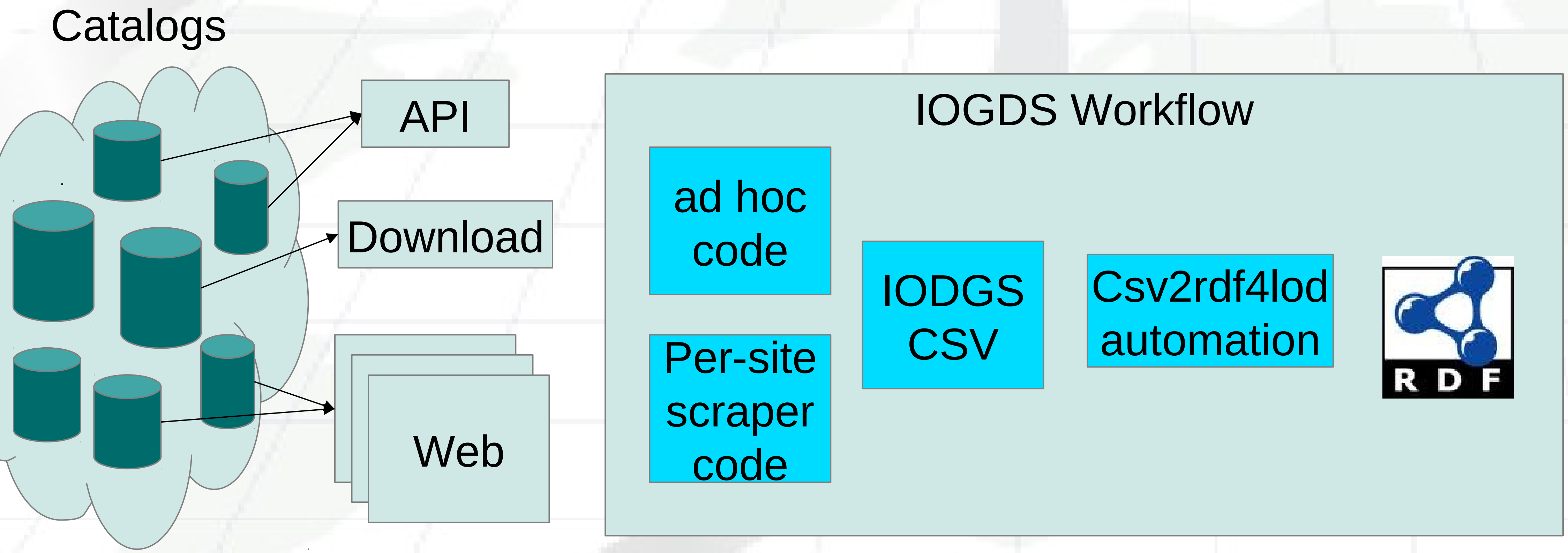
DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. This document defines the schema and provides examples for its use.

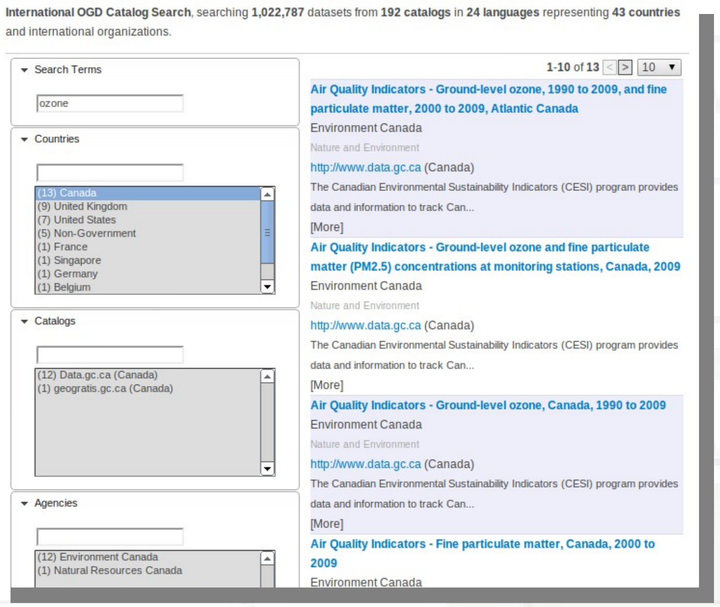
By using DCAT to describe datasets in data catalogs, publishers increase discoverability and enable applications easily to consume metadata from multiple catalogs. It further enables decentralized publishing of catalogs and facilitates federated dataset search across sites. Aggregated DCAT metadata can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

- Anticipates W3C DCAT RDF vocabulary
- Demos what a comprehensive **federated catalog** based on DCAT and aggregation API might look like

International Open Government Dataset Search

IOGDS is a multi-year effort based on downloading, scraping or accessing APIs, converting metadata to a proto-DCAT model, and publishing via endpoint and download





International OGD Catalog Search, searching **1,022,787** datasets from **192 catalogs** in **24 languages** representing **43 countries** and international organizations.

Search Terms

ozone

Countries

(13) Canada
(9) United Kingdom
(7) United States
(5) Non-Government
(1) France
(1) Singapore
(1) Germany
(1) Belgium

Catalogs

(12) Data.gc.ca (Canada)
(1) geogratis.gc.ca (Canada)

Agencies

(12) Environment Canada
(1) Natural Resources Canada

1-10 of **13** < > 10 ▼

Air Quality Indicators - Ground-level ozone, 1990 to 2009, and fine particulate matter, 2000 to 2009, Atlantic Canada
Environment Canada
Nature and Environment
http://www.data.gc.ca (Canada)
The Canadian Environmental Sustainability Indicators (CESI) program provides data and information to track Can...
[More]

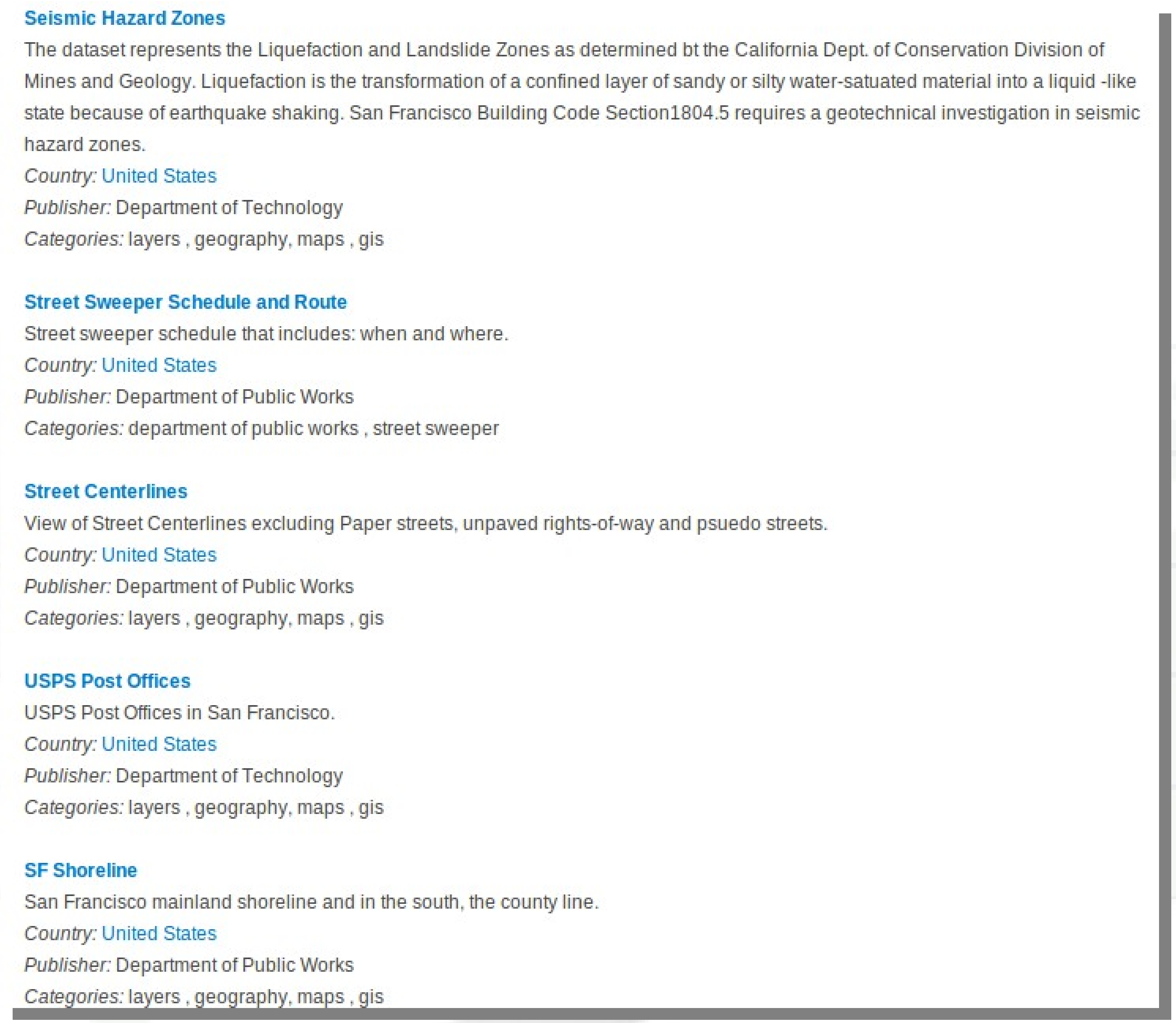
Air Quality Indicators - Ground-level ozone and fine particulate matter (PM2.5) concentrations at monitoring stations, Canada, 2009
Environment Canada
Nature and Environment
http://www.data.gc.ca (Canada)
The Canadian Environmental Sustainability Indicators (CESI) program provides data and information to track Can...
[More]

Air Quality Indicators - Ground-level ozone, Canada, 1990 to 2009
Environment Canada
Nature and Environment
http://www.data.gc.ca (Canada)
The Canadian Environmental Sustainability Indicators (CESI) program provides data and information to track Can...
[More]

Air Quality Indicators - Fine particulate matter, Canada, 2000 to 2009
Environment Canada

See: http://logd.tw.rpi.edu

Schema.org: Semantic Markup for Discovery



Seismic Hazard Zones

The dataset represents the Liquefaction and Landslide Zones as determined bt the California Dept. of Conservation Division of Mines and Geology. Liquefaction is the transformation of a confined layer of sandy or silty water-satuated material into a liquid -like state because of earthquake shaking. San Francisco Building Code Section1804.5 requires a geotechnical investigation in seismic hazard zones.

Country: United States
Publisher: Department of Technology
Categories: layers , geography, maps , gis

Street Sweeper Schedule and Route

Street sweeper schedule that includes: when and where.

Country: United States
Publisher: Department of Public Works
Categories: department of public works , street sweeper

Street Centerlines

View of Street Centerlines excluding Paper streets, unpaved rights-of-way and psuedo streets.

Country: United States
Publisher: Department of Public Works
Categories: layers , geography, maps , gis

USPS Post Offices

USPS Post Offices in San Francisco.

Country: United States
Publisher: Department of Technology
Categories: layers , geography, maps , gis

SF Shoreline

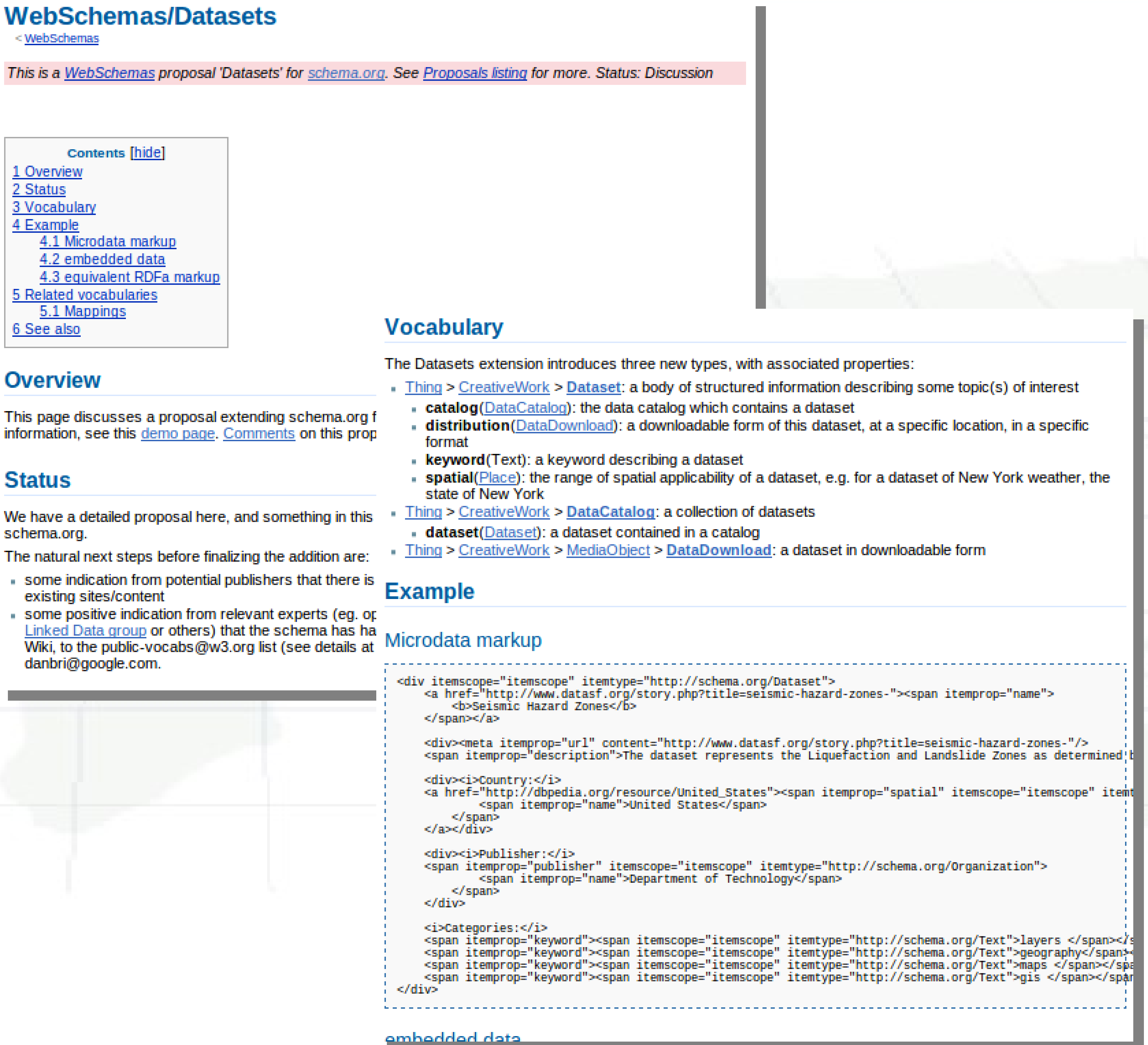
San Francisco mainland shoreline and in the south, the county line.

Country: United States
Publisher: Department of Public Works
Categories: layers , geography, maps , gis

TWC RPI has published dataset listings based on IOGDS using emerging microdata standards, esp. **schema.org** model endorsed by **Bing, Google, Yahoo!, Yandex...**

Schema.org **datasets** extension

- TWC RPI's schema.org **dataset extension** will enable government dataset catalogs to more easily be parsed and indexed by the major search engines...
- ...which will help users find relevant datasets!
- TWC's dataset extension entered public discussion June 2012



WebSchemas/Datasets

< WebSchemas

This is a WebSchemas proposal 'Datasets' for schema.org. See Proposals listing for more. Status: Discussion

Contents [hide]

Overview

This page discusses a proposal extending schema.org f
information, see this demo page. Comments on this prop

Status

We have a detailed proposal here, and something in this
schema.org.

The natural next steps before finalizing the addition are:

- some indication from potential publishers that there is
existing sites/content
- some positive indication from relevant experts (eg. op
Linked Data group or others) that the schema has ha
Wiki, to the public-vocabs@w3.org list (see details at
danbri@google.com.

Vocabulary

The Datasets extension introduces three new types, with associated properties:

- Thing > CreativeWork > Dataset: a body of structured information describing some topic(s) of interest
  - **catalog**(DataCatalog): the data catalog which contains a dataset
  - **distribution**(DataDownload): a downloadable form of this dataset, at a specific location, in a specific format
  - **keyword**(Text): a keyword describing a dataset
  - **spatial**(Place): the range of spatial applicability of a dataset, e.g. for a dataset of New York weather, the state of New York
- Thing > CreativeWork > DataCatalog: a collection of datasets
  - **dataset**(Dataset): a dataset contained in a catalog
- Thing > CreativeWork > MediaObject > DataDownload: a dataset in downloadable form

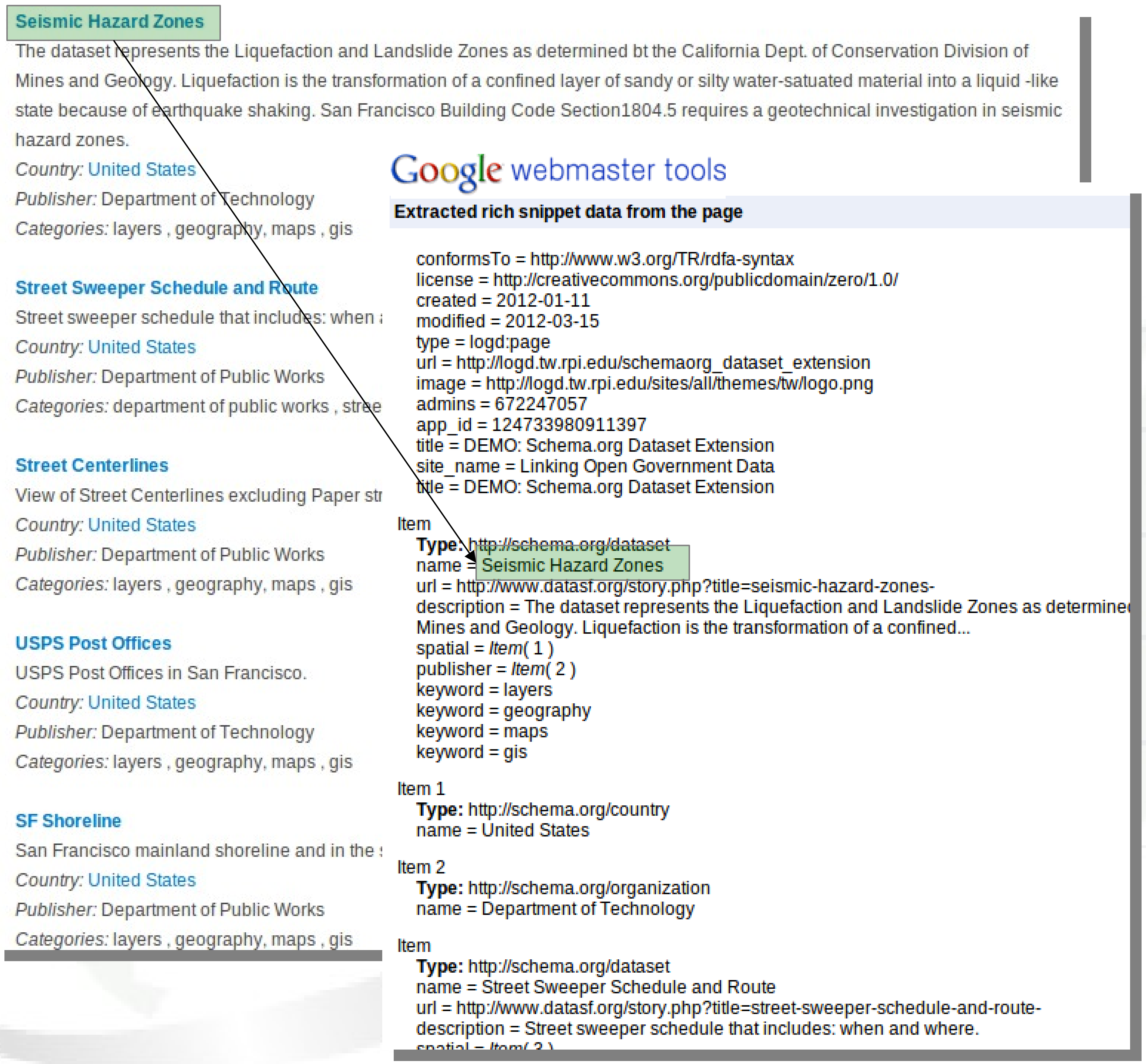
Example

Microdata markup

```
<div itemscope="itemscope" itemtype="http://schema.org/Dataset">
    <a href="http://www.datasf.org/story.php?title=seismic-hazard-zones-"><span itemprop="name">
        <b>Seismic Hazard Zones</b>
    </span></a>

    <div><meta itemprop="url" content="http://www.datasf.org/story.php?title=seismic-hazard-zones-"/>
    <span itemprop="description">The dataset represents the Liquefaction and Landslide Zones as determined b

    <div><i>Country:</i>
    <a href="http://dbpedia.org/resource/United_States"><span itemprop="spatial" itemscope="itemscope" item
            <span itemprop="name">United States</span>
        </span>
    </a></div>

    <div><i>Publisher:</i>
    <span itemprop="publisher" itemscope="itemscope" itemtype="http://schema.org/Organization">
            <span itemprop="name">Department of Technology</span>
        </span>
    </div>

    <i>Categories:</i>
    <span itemprop="keyword"><span itemscope="itemscope" itemtype="http://schema.org/Text">layers </span></s
    <span itemprop="keyword"><span itemscope="itemscope" itemtype="http://schema.org/Text">geography</span>
    <span itemprop="keyword"><span itemscope="itemscope" itemtype="http://schema.org/Text">maps </span></spa
    <span itemprop="keyword"><span itemscope="itemscope" itemtype="http://schema.org/Text">gis </span></span
</div>
```

embedded data

Schema.org **datasets** extension



Seismic Hazard Zones

The dataset represents the Liquefaction and Landslide Zones as determined bt the California Dept. of Conservation Division of Mines and Geology. Liquefaction is the transformation of a confined layer of sandy or silty water-satuated material into a liquid -like state because of earthquake shaking. San Francisco Building Code Section1804.5 requires a geotechnical investigation in seismic hazard zones.

Country: United States

Publisher: Department of Technology

Categories: layers , geography, maps , gis

Street Sweeper Schedule and Route

Street sweeper schedule that includes: when a

Country: United States

Publisher: Department of Public Works

Categories: department of public works , stree

Street Centerlines

View of Street Centerlines excluding Paper str

Country: United States

Publisher: Department of Public Works

Categories: layers , geography, maps , gis

USPS Post Offices

USPS Post Offices in San Francisco.

Country: United States

Publisher: Department of Technology

Categories: layers , geography, maps , gis

SF Shoreline

San Francisco mainland shoreline and in the s

Country: United States

Publisher: Department of Public Works

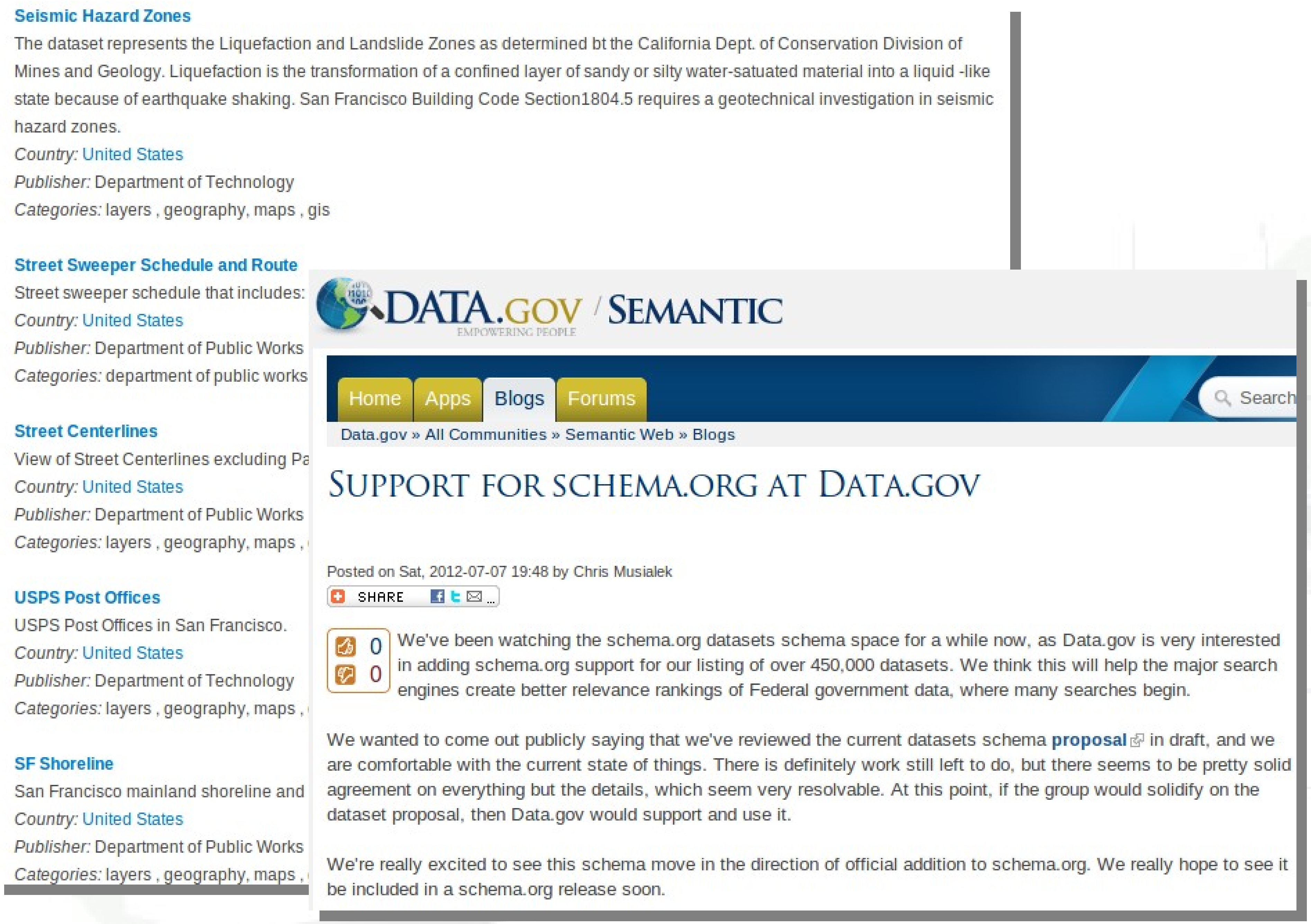
Categories: layers , geography, maps , gis

Extracted rich snippet data from the page

```
    conformsTo = http://www.w3.org/TR/rdfa-syntax
    license = http://creativecommons.org/publicdomain/zero/1.0/
    created = 2012-01-11
    modified = 2012-03-15
    type = logd:page
    url = http://logd.tw.rpi.edu/schemaorg_dataset_extension
    image = http://logd.tw.rpi.edu/sites/all/themes/tw/logo.png
    admins = 672247057
    app_id = 124733980911397
    title = DEMO: Schema.org Dataset Extension
    site_name = Linking Open Government Data
    title = DEMO: Schema.org Dataset Extension

Item
    Type: http://schema.org/dataset
    name = Seismic Hazard Zones
    url = http://www.datasf.org/story.php?title=seismic-hazard-zones-
    description = The dataset represents the Liquefaction and Landslide Zones as determined
    Mines and Geology. Liquefaction is the transformation of a confined...
    spatial = Item( 1 )
    publisher = Item( 2 )
    keyword = layers
    keyword = geography
    keyword = maps
    keyword = gis

Item 1
    Type: http://schema.org/country
    name = United States

Item 2
    Type: http://schema.org/organization
    name = Department of Technology

Item
    Type: http://schema.org/dataset
    name = Street Sweeper Schedule and Route
    url = http://www.datasf.org/story.php?title=street-sweeper-schedule-and-route-
    description = Street sweeper schedule that includes: when and where.
    spatial = Item( 3 )
```

The schema.org **datasets** extension enables **relevant datasets** to be more easily **discovered** by a range of stakeholders including researchers, data journalists, bloggers and developers

Schema.org **datasets** extension



Seismic Hazard Zones
The dataset represents the Liquefaction and Landslide Zones as determined bt the California Dept. of Conservation Division of Mines and Geology. Liquefaction is the transformation of a confined layer of sandy or silty water-satuated material into a liquid -like state because of earthquake shaking. San Francisco Building Code Section1804.5 requires a geotechnical investigation in seismic hazard zones.
Country: United States
Publisher: Department of Technology
Categories: layers , geography, maps , gis

Street Sweeper Schedule and Route
Street sweeper schedule that includes:
Country: United States
Publisher: Department of Public Works
Categories: department of public works

Street Centerlines
View of Street Centerlines excluding Pa
Country: United States
Publisher: Department of Public Works
Categories: layers , geography, maps ,

USPS Post Offices
USPS Post Offices in San Francisco.
Country: United States
Publisher: Department of Technology
Categories: layers , geography, maps ,

SF Shoreline
San Francisco mainland shoreline and
Country: United States
Publisher: Department of Public Works
Categories: layers , geography, maps ,

Data.gov » All Communities » Semantic Web » Blogs

SUPPORT FOR SCHEMA.ORG AT DATA.GOV

Posted on Sat, 2012-07-07 19:48 by Chris Musialek

We've been watching the schema.org datasets schema space for a while now, as Data.gov is very interested in adding schema.org support for our listing of over 450,000 datasets. We think this will help the major search engines create better relevance rankings of Federal government data, where many searches begin.

We wanted to come out publicly saying that we've reviewed the current datasets schema **proposal** in draft, and we are comfortable with the current state of things. There is definitely work still left to do, but there seems to be pretty solid agreement on everything but the details, which seem very resolvable. At this point, if the group would solidify on the dataset proposal, then Data.gov would support and use it.

We're really excited to see this schema move in the direction of official addition to schema.org. We really hope to see it be included in a schema.org release soon.

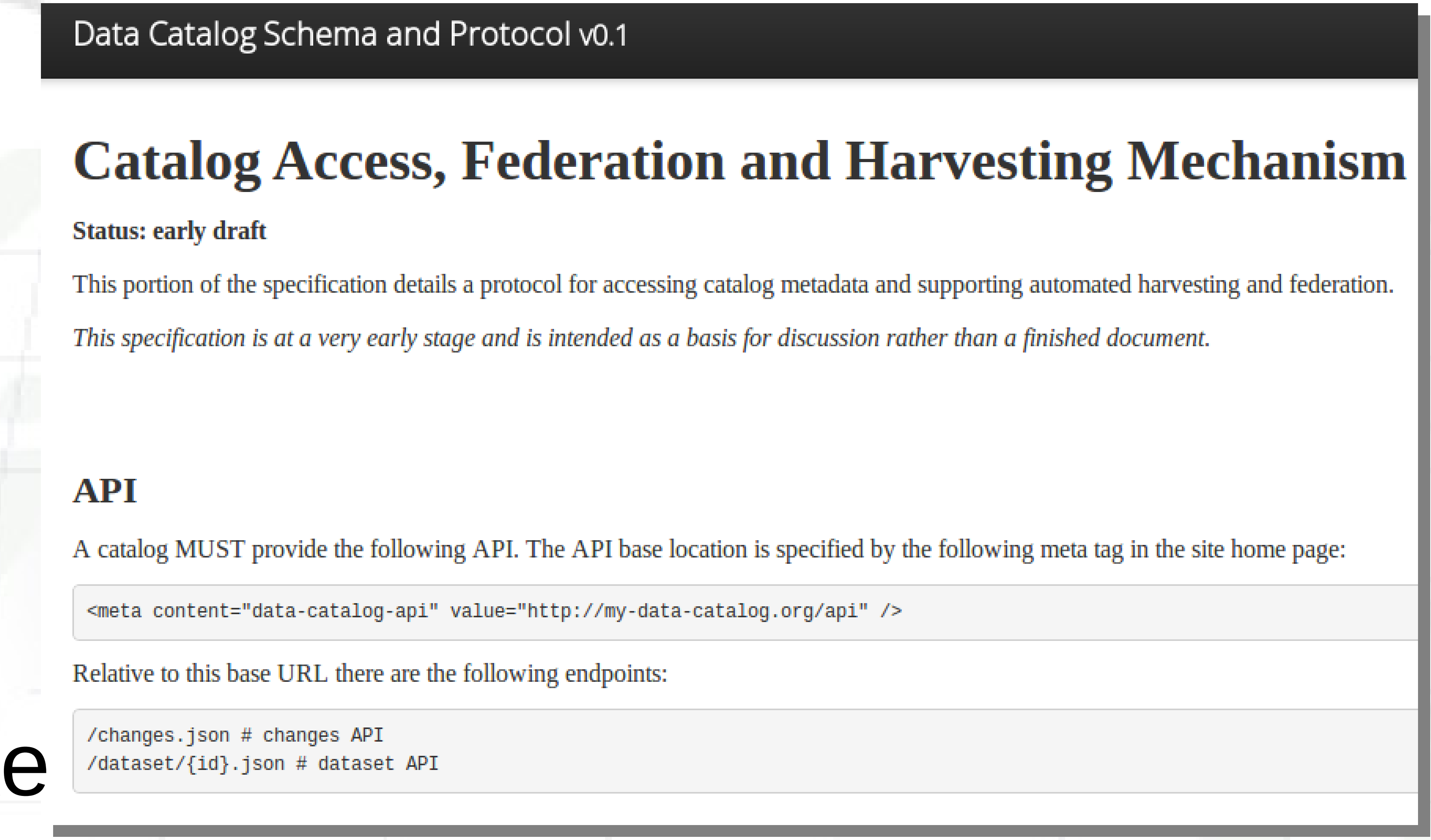
"...we've reviewed the current datasets schema proposal in draft, and we are comfortable with the current state of things...

"...At this point, if the group would solidify on the dataset proposal, then **Data.gov would support and use it.**

---Chris Musialek

CKAN Data Catalog Scheme & Protocol

- API-based catalog federation is also possible
- ckan announced DCAT-based query/federation API
- enables OAI-PMH-like harvesting and more



Data Catalog Schema and Protocol v0.1

Catalog Access, Federation and Harvesting Mechanism

Status: early draft

This portion of the specification details a protocol for accessing catalog metadata and supporting automated harvesting and federation.

This specification is at a very early stage and is intended as a basis for discussion rather than a finished document.

API

A catalog MUST provide the following API. The API base location is specified by the following meta tag in the site home page:

```
<meta content="data-catalog-api" value="http://my-data-catalog.org/api" />
```

Relative to this base URL there are the following endpoints:

```
/changes.json # changes API
/dataset/{id}.json # dataset API
```

Other Thoughts...

- Geo-based discovery:
  What data is available by geo-selection?
- Provenance-based discovery:
  How do I get the data that someone else used? “Get the Data”
- Community/social-based discovery:
  Dude, check out this data! (Linked Data perfect for this...

Other Thoughts...

- Geo-based discovery:
  What data is available by geo-selection?

DATA.GOV *Geo Viewer*

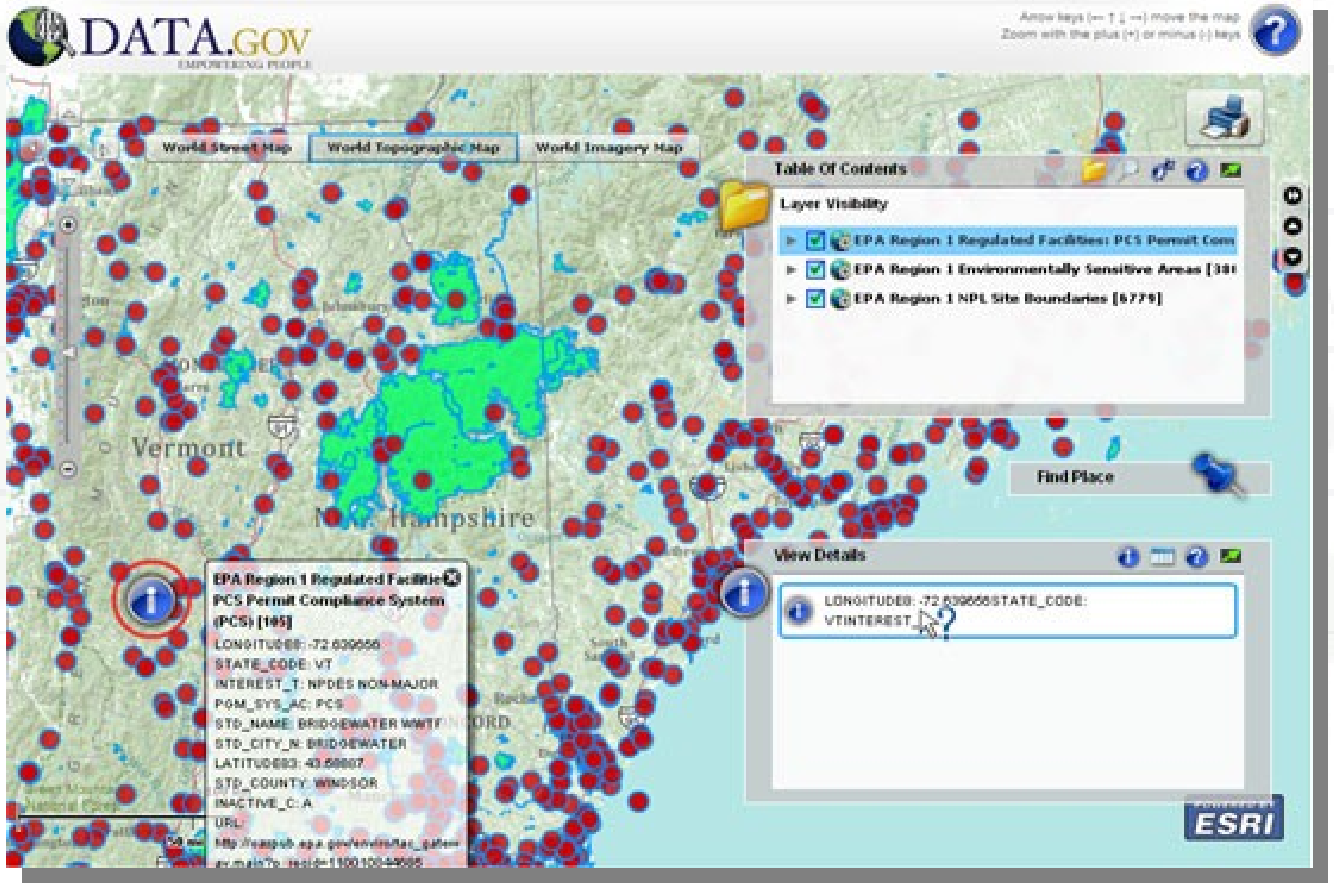


Table Of Contents

Layer Visibility

- ☑ EPA Region 1 Regulated Facilities: PCS Permit Com
- ☑ EPA Region 1 Environmentally Sensitive Areas [38
- ☑ EPA Region 1 NPL Site Boundaries [6779]

EPA Region 1 Regulated Facilitie
PCS Permit Compliance System
(PCS) [105]
LONGITUDE8: -72.639656
STATE_CODE: VT
INTEREST_T: NPDES NON-MAJOR
PGM_SYS_AC: PCS
STD_NAME: BRIDGEWATER WWTF
STD_CITY_N: BRIDGEWATER
LATITUDE83: 43.66807
STD_COUNTY: WINDSOR
INACTIVE_C: A
URL:
http://oaspub.epa.gov/enviro/fac_gatew
ay.main?p_regid=110010046685

View Details

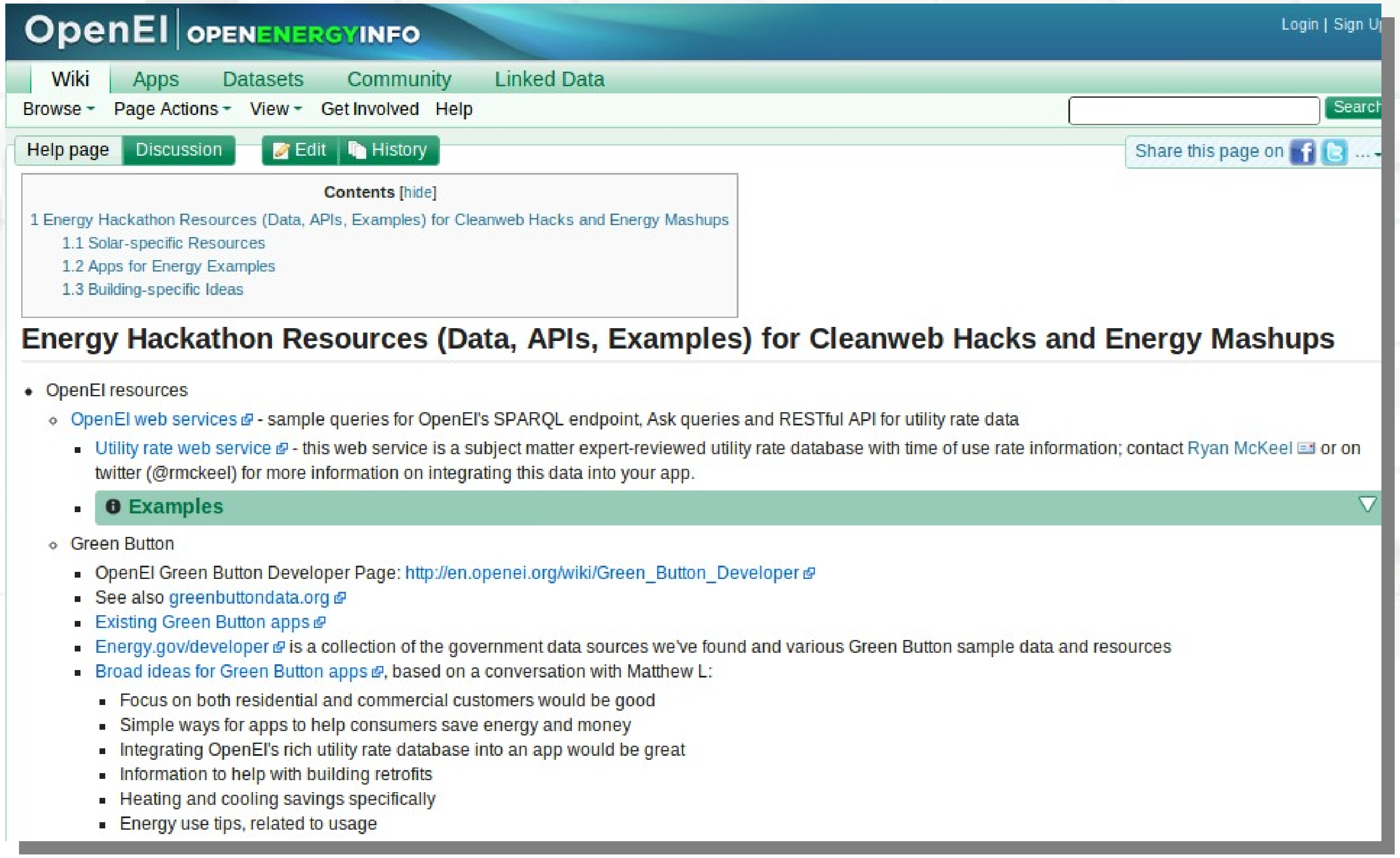
LONGITUDE8: -72.639656STATE_CODE:
VTINTEREST_

Other Thoughts...

- Community/social-based discovery:

Dude, check out this data!

OPENEI.org



Contents [hide]

1 Energy Hackathon Resources (Data, APIs, Examples) for Cleanweb Hacks and Energy Mashups
   1.1 Solar-specific Resources
   1.2 Apps for Energy Examples
   1.3 Building-specific Ideas

Energy Hackathon Resources (Data, APIs, Examples) for Cleanweb Hacks and Energy Mashups

- OpenEI resources
  - OpenEI web services - sample queries for OpenEI's SPARQL endpoint, Ask queries and RESTful API for utility rate data
    - Utility rate web service - this web service is a subject matter expert-reviewed utility rate database with time of use rate information; contact Ryan McKeel or on twitter (@rmckeel) for more information on integrating this data into your app.
    - Examples
  - Green Button
    - OpenEI Green Button Developer Page: http://en.openei.org/wiki/Green_Button_Developer
    - See also greenbuttondata.org
    - Existing Green Button apps
    - Energy.gov/developer is a collection of the government data sources we've found and various Green Button sample data and resources
    - Broad ideas for Green Button apps, based on a conversation with Matthew L:
      - Focus on both residential and commercial customers would be good
      - Simple ways for apps to help consumers save energy and money
      - Integrating OpenEI's rich utility rate database into an app would be great
      - Information to help with building retrofits
      - Heating and cooling savings specifically
      - Energy use tips, related to usage

Choose your own medicine...

but **do** expose your **metadata**

and **get** your catalogs **discovered**!