Searches and schemas

Hal Varian
Google, Inc.
FESAC presentation
June 12, 2020

Outline

- Schema.org
- Job search
- Dataset search
- Data commons
- Questions

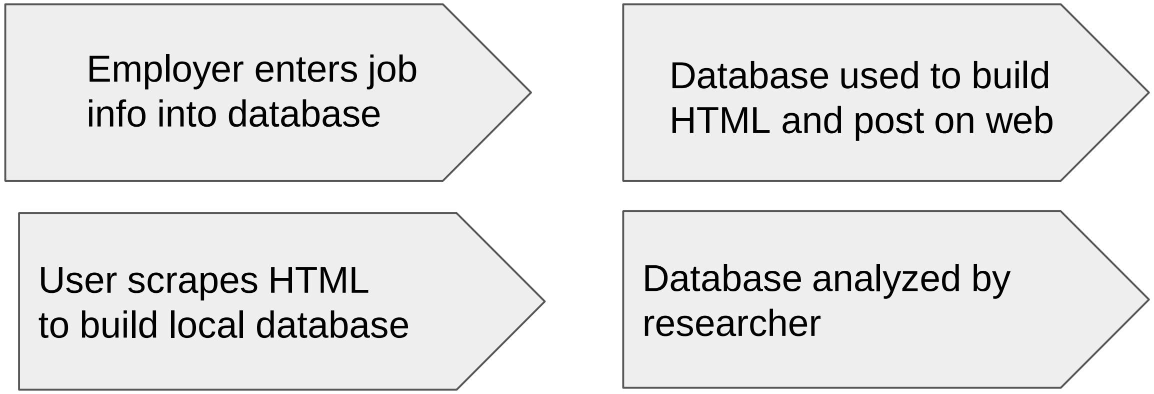
Outline

- Schema.org
- Job search
- Dataset search
- Data commons
- Questions

Scraping data (perhaps using burningglass® TECHNOLOGIES)

HTML is a markup language to format documents for human readers

It is now common to scrape data from the web, e.g., Computerization of White Collar Jobs. from UpJohn Institute, which scrapes online job postings.



Employer enters job info into database

Database used to build HTML and post on web

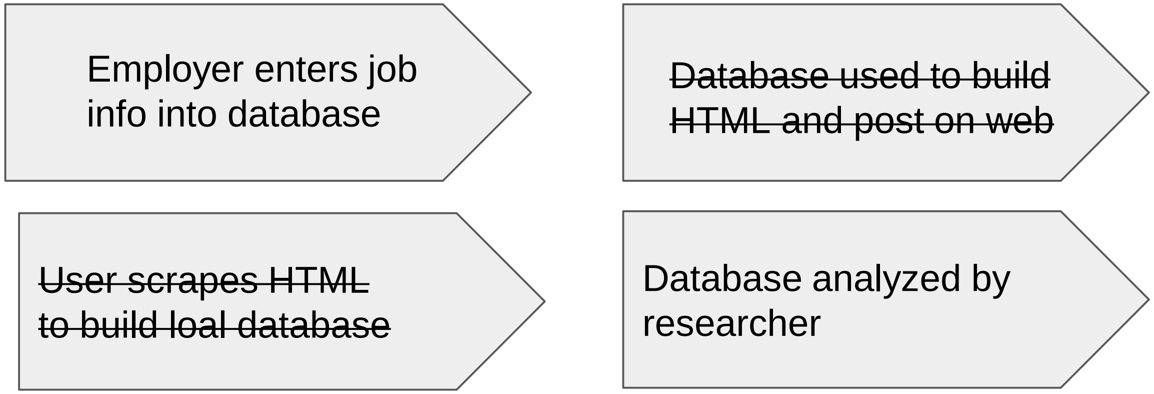
User scrapes HTML to build local database

Database analyzed by researcher

Scraping data (perhaps using burningglass® TECHNOLOGIES)

HTML is a markup language to format documents for human readers.

It is now common to scrape data from the web, e.g., Computerization of White Collar Jobs. from UpJohn Institute, which scrapes online job postings.



Employer enters job info into database

~~Database used to build HTML and post on web~~

~~User scrapes HTML to build loal database~~

Database analyzed by researcher

- **Q:** How make human readable document also machine readable?
- **A:** Embed metadata into the HTML.
- For that we need standards for the metadata.

schema.org

“Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond.”

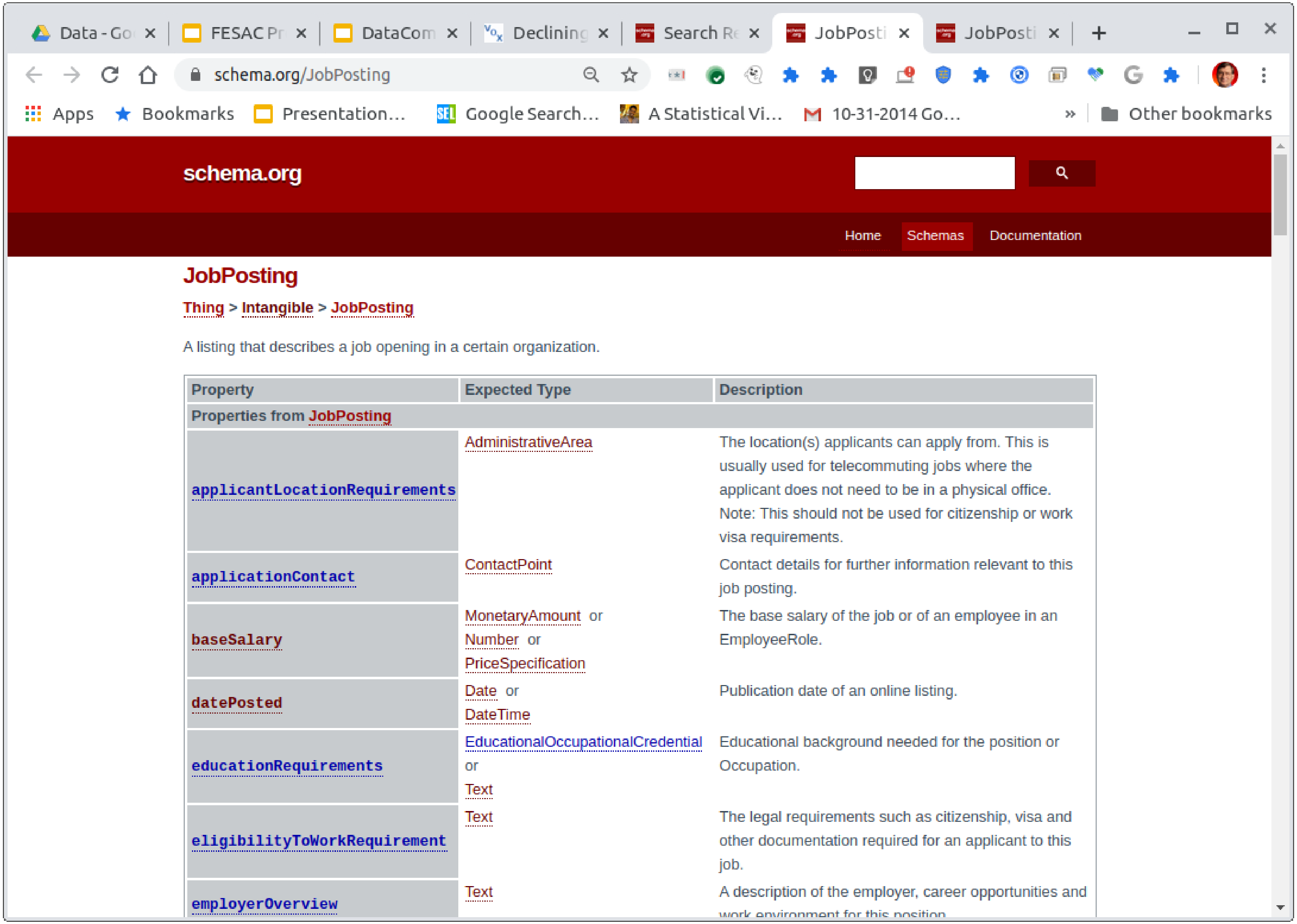
- Creative works: CreativeWork, Book, Movie, MusicRecording, Recipe, TVSeries ...
- Embedded non-text objects: AudioObject, ImageObject, VideoObject
- Event
- Health and medical types
- Organization
- Person
- Place, LocalBusiness, Restaurant ...
- Product, Offer, AggregateOffer
- Review, AggregateRating
- Action
- **JobPosting**
- **Dataset**

Schema.org is already widely used

- See CACM article on Schema.org: Evolution of Structured Data on the Web
- ~45% of pages in the Google search index have schema.org markup
- Most popular mark up use
  - 6 billion product listings across 5m domains
  - 96m job postings (including duplicates)
- ...and many more from comic books to recipes to mortgages, to events …
- Lesson: use schema.org standards if you want your data to be used

Outline

- Schema.org
- Job search
- Dataset search
- Data commons
- Questions



JobPosting

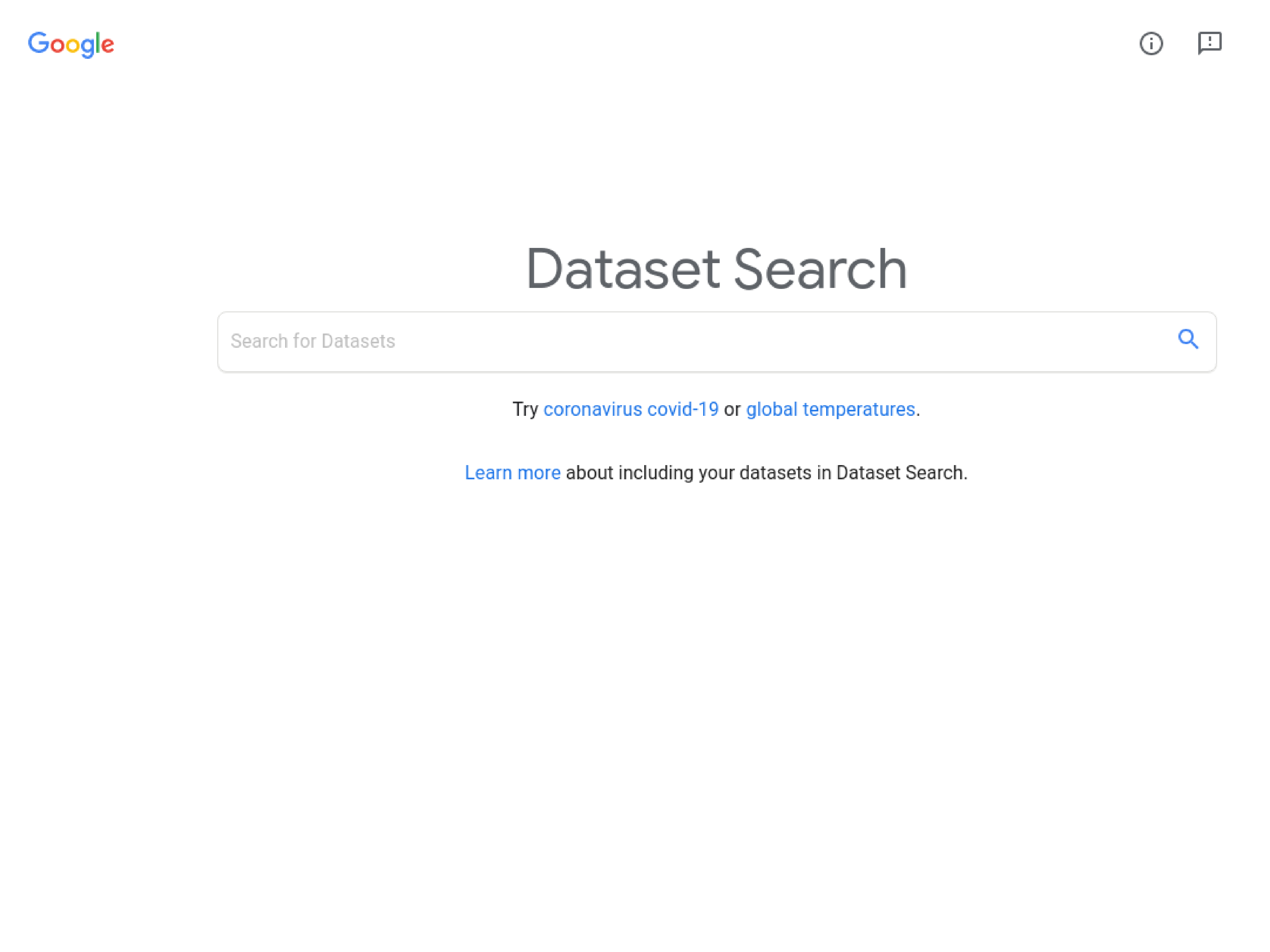
Thing > Intangible > JobPosting

A listing that describes a job opening in a certain organization.

Property	Expected Type	Description
Properties from JobPosting		
applicantLocationRequirements	AdministrativeArea	The location(s) applicants can apply from. This is usually used for telecommuting jobs where the applicant does not need to be in a physical office. Note: This should not be used for citizenship or work visa requirements.
applicationContact	ContactPoint	Contact details for further information relevant to this job posting.
baseSalary	MonetaryAmount or Number or PriceSpecification	The base salary of the job or of an employee in an EmployeeRole.
datePosted	Date or DateTime	Publication date of an online listing.
educationRequirements	EducationalOccupationalCredential or Text	Educational background needed for the position or Occupation.
eligibilityToWorkRequirement	Text	The legal requirements such as citizenship, visa and other documentation required for an applicant to this job.
employerOverview	Text	A description of the employer, career opportunities and work environment for this position.

Outline

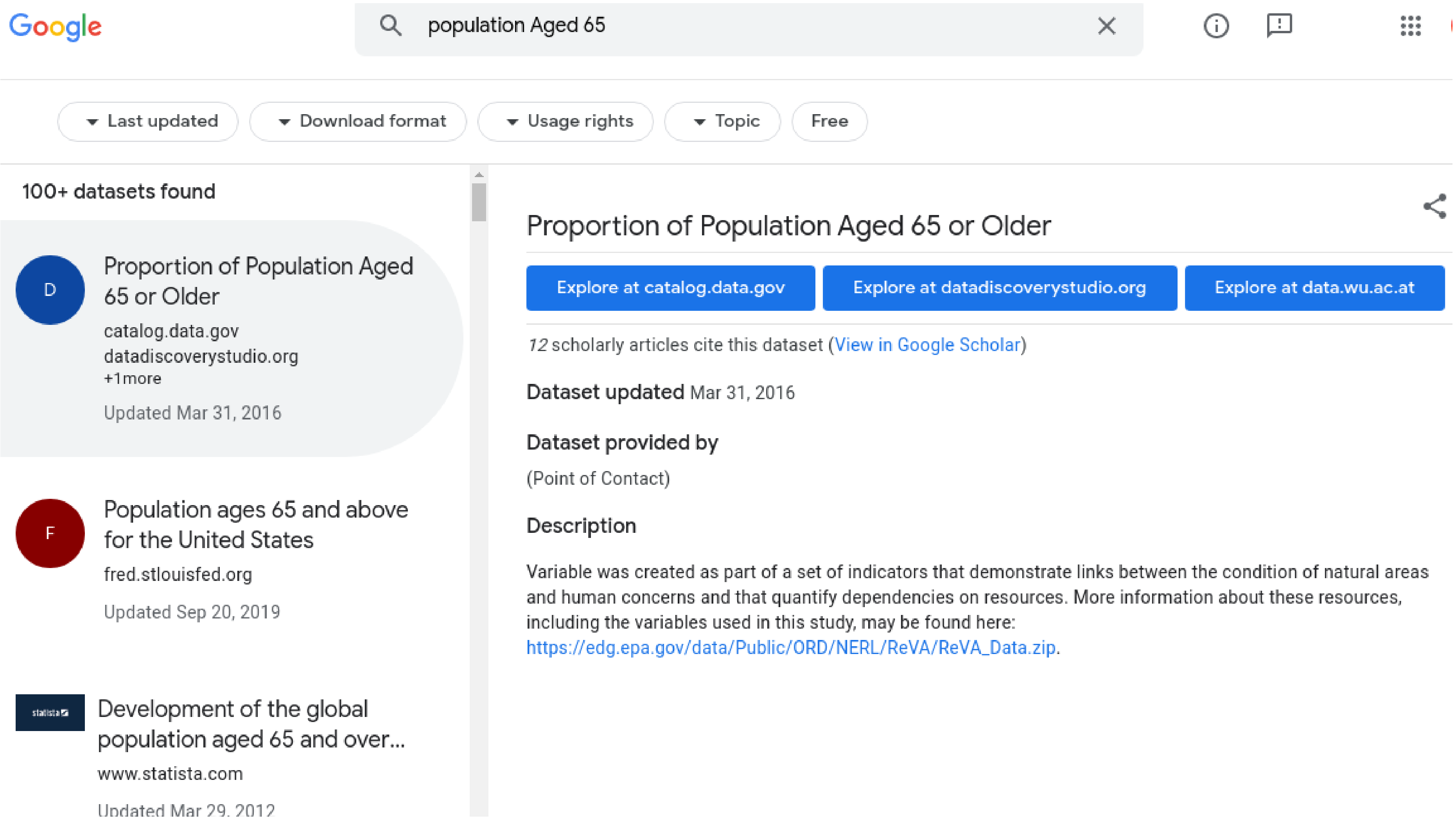
- Schema.org
- Job search
- Dataset search
- Data commons
- Questions



Dataset Search

Try coronavirus covid-19 or global temperatures.

Learn more about including your datasets in Dataset Search.



100+ datasets found

Proportion of Population Aged 65 or Older
catalog.data.gov
datadiscoverystudio.org
+1more
Updated Mar 31, 2016

Population ages 65 and above for the United States
fred.stlouisfed.org
Updated Sep 20, 2019

Development of the global population aged 65 and over...
www.statista.com
Updated Mar 29, 2012

Proportion of Population Aged 65 or Older

12 scholarly articles cite this dataset (View in Google Scholar)

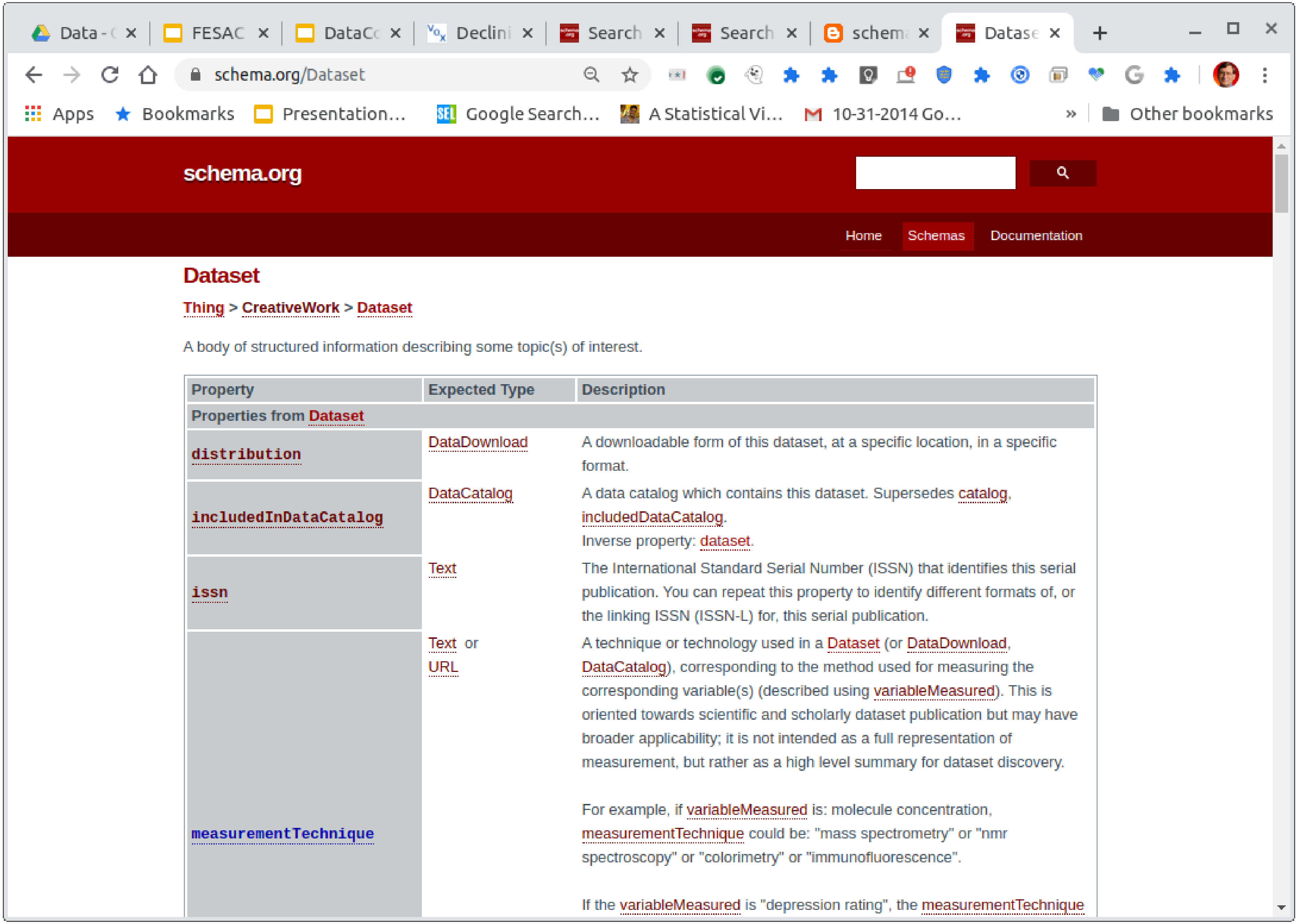
Dataset updated Mar 31, 2016

Dataset provided by

(Point of Contact)

Description

Variable was created as part of a set of indicators that demonstrate links between the condition of natural areas and human concerns and that quantify dependencies on resources. More information about these resources, including the variables used in this study, may be found here: https://edg.epa.gov/data/Public/ORD/NERL/ReVA/ReVA_Data.zip.



Dataset

Thing > CreativeWork > Dataset

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

Property	Expected Type	Description
Properties from Dataset		
distribution	DataDownload	A downloadable form of this dataset, at a specific location, in a specific format.
includedInDataCatalog	DataCatalog	A data catalog which contains this dataset. Supersedes catalog, includedDataCatalog. Inverse property: dataset.
issn	Text	The International Standard Serial Number (ISSN) that identifies this serial publication. You can repeat this property to identify different formats of, or the linking ISSN (ISSN-L) for, this serial publication.
measurementTechnique	Text or URL	A technique or technology used in a Dataset (or DataDownload, DataCatalog), corresponding to the method used for measuring the corresponding variable(s) (described using variableMeasured). This is oriented towards scientific and scholarly dataset publication but may have broader applicability; it is not intended as a full representation of measurement, but rather as a high level summary for dataset discovery. For example, if variableMeasured is: molecule concentration, measurementTechnique could be: "mass spectrometry" or "nmr spectroscopy" or "colorimetry" or "immunofluorescence". If the variableMeasured is "depression rating", the measurementTechnique

Metadata about datasets

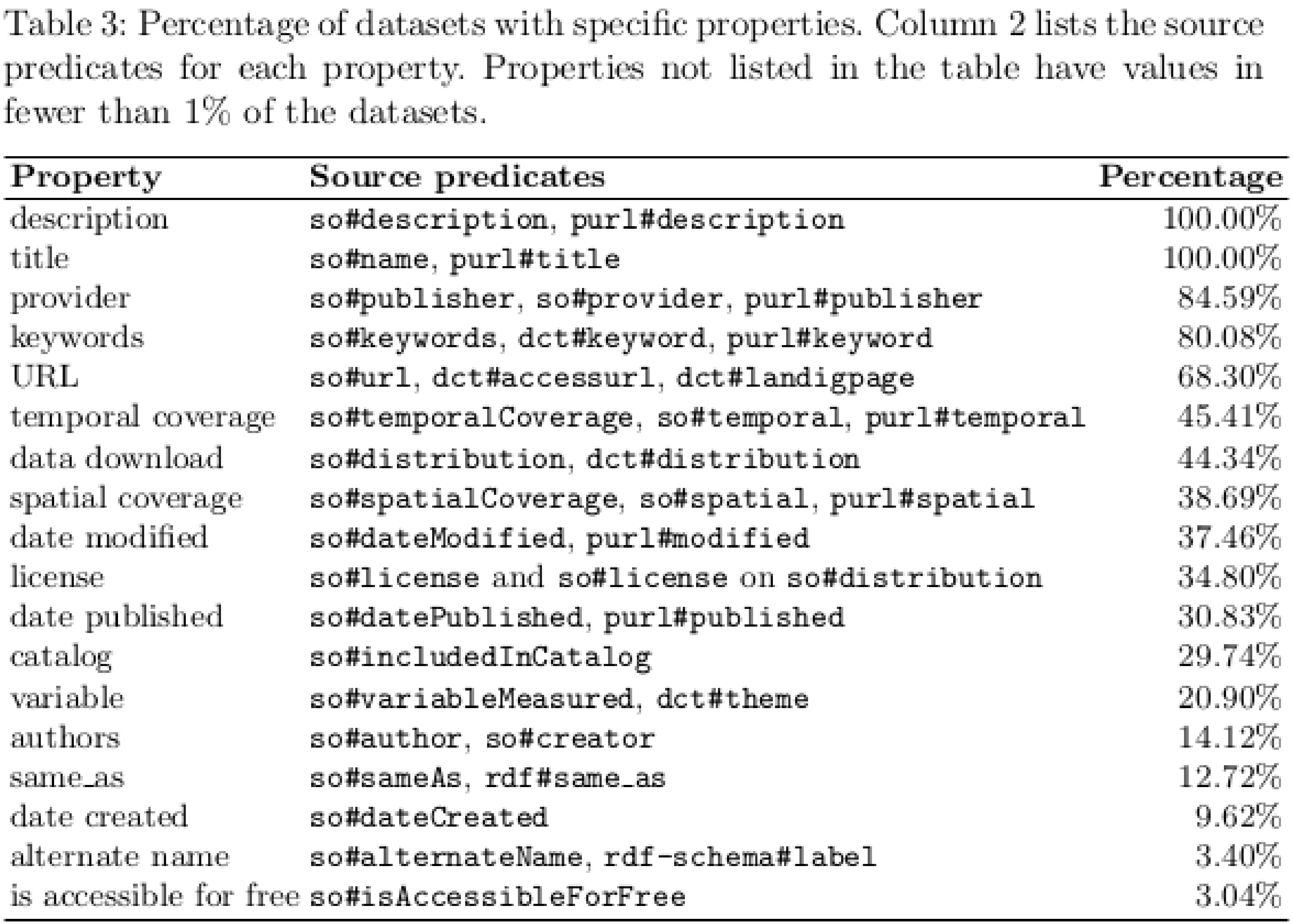


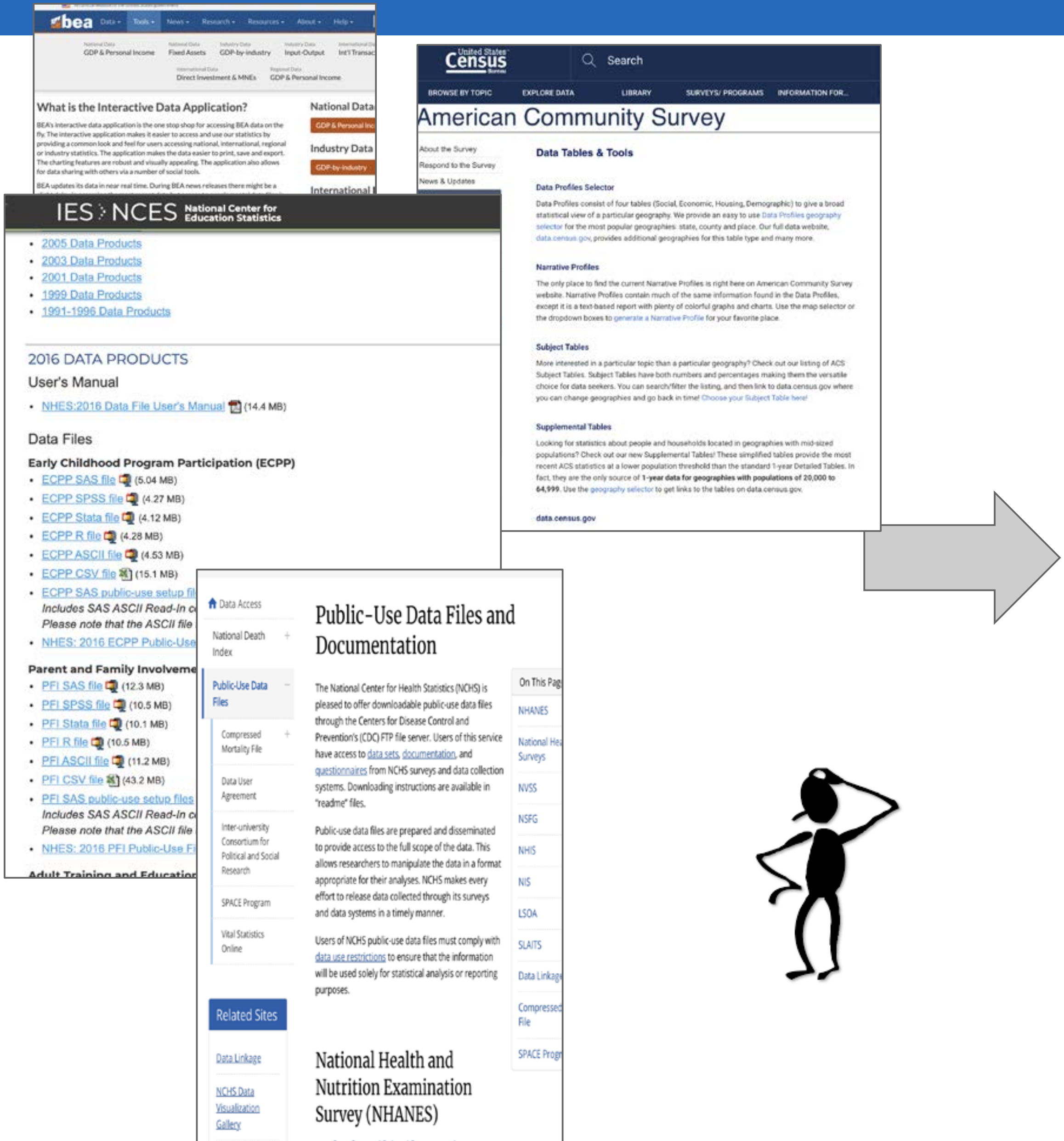
Table 3: Percentage of datasets with specific properties. Column 2 lists the source predicates for each property. Properties not listed in the table have values in fewer than 1% of the datasets.

Property	Source predicates	Percentage
description	`so#description`, `purl#description`	100.00%
title	`so#name`, `purl#title`	100.00%
provider	`so#publisher`, `so#provider`, `purl#publisher`	84.59%
keywords	`so#keywords`, `dct#keyword`, `purl#keyword`	80.08%
URL	`so#url`, `dct#accessurl`, `dct#landigpage`	68.30%
temporal coverage	`so#temporalCoverage`, `so#temporal`, `purl#temporal`	45.41%
data download	`so#distribution`, `dct#distribution`	44.34%
spatial coverage	`so#spatialCoverage`, `so#spatial`, `purl#spatial`	38.69%
date modified	`so#dateModified`, `purl#modified`	37.46%
license	`so#license` and `so#license` on `so#distribution`	34.80%
date published	`so#datePublished`, `purl#published`	30.83%
catalog	`so#includedInCatalog`	29.74%
variable	`so#variableMeasured`, `dct#theme`	20.90%
authors	`so#author`, `so#creator`	14.12%
same_as	`so#sameAs`, `rdf#same_as`	12.72%
date created	`so#dateCreated`	9.62%
alternate name	`so#alternateName`, `rdf-schema#label`	3.40%
is accessible for free	`so#isAccessibleForFree`	3.04%

Outline

- Schema.org
- Job search
- Dataset search
- Data commons
- Questions

We want to make it easier to find and analyze data



bea Data ▾ Tools ▾ News ▾ Research ▾ Resources ▾ About ▾ Help ▾

National Data GDP & Personal Income | National Data Fixed Assets | Industry Data GDP-by-industry | Industry Data Input-Output | International Data Int'l Transac

International Data Direct Investment & MNEs | Regional Data GDP & Personal Income

What is the Interactive Data Application?

BEA's interactive data application is the one stop shop for accessing BEA data on the fly. The interactive application makes it easier to access and use our statistics by providing a common look and feel for users accessing national, international, regional or industry statistics. The application makes the data easier to print, save and export. The charting features are robust and visually appealing. The application also allows for data sharing with others via a number of social tools.

BEA updates its data in near real time. During BEA news releases there might be a

National Data

GDP & Personal Inc

Industry Data

GDP-by-industry

International

United States Census Bureau — Search

BROWSE BY TOPIC | EXPLORE DATA | LIBRARY | SURVEYS/ PROGRAMS | INFORMATION FOR...

American Community Survey

About the Survey

Respond to the Survey

News & Updates

Data Tables & Tools

Data Profiles Selector

Data Profiles consist of four tables (Social, Economic, Housing, Demographic) to give a broad statistical view of a particular geography. We provide an easy to use Data Profiles geography selector for the most popular geographies: state, county and place. Our full data website, data.census.gov, provides additional geographies for this table type and many more.

Narrative Profiles

The only place to find the current Narrative Profiles is right here on American Community Survey website. Narrative Profiles contain much of the same information found in the Data Profiles, except it is a text-based report with plenty of colorful graphs and charts. Use the map selector or the dropdown boxes to generate a Narrative Profile for your favorite place.

Subject Tables

More interested in a particular topic than a particular geography? Check out our listing of ACS Subject Tables. Subject Tables have both numbers and percentages making them the versatile choice for data seekers. You can search/filter the listing, and then link to data.census.gov where you can change geographies and go back in time! Choose your Subject Table here!

Supplemental Tables

Looking for statistics about people and households located in geographies with mid-sized populations? Check out our new Supplemental Tables! These simplified tables provide the most recent ACS statistics at a lower population threshold than the standard 1-year Detailed Tables. In fact, they are the only source of **1-year data for geographies with populations of 20,000 to 64,999**. Use the geography selector to get links to the tables on data.census.gov.

data.census.gov

IES › NCES National Center for Education Statistics

- 2005 Data Products
- 2003 Data Products
- 2001 Data Products
- 1999 Data Products
- 1991-1996 Data Products

2016 DATA PRODUCTS

User's Manual

- NHES:2016 Data File User's Manual (14.4 MB)

Data Files

Early Childhood Program Participation (ECPP)

- ECPP SAS file (5.04 MB)
- ECPP SPSS file (4.27 MB)
- ECPP Stata file (4.12 MB)
- ECPP R file (4.28 MB)
- ECPP ASCII file (4.53 MB)
- ECPP CSV file (15.1 MB)
- ECPP SAS public-use setup fil
  Includes SAS ASCII Read-in c
  Please note that the ASCII file
- NHES: 2016 ECPP Public-Use

Parent and Family Involveme

- PFI SAS file (12.3 MB)
- PFI SPSS file (10.5 MB)
- PFI Stata file (10.1 MB)
- PFI R file (10.5 MB)
- PFI ASCII file (11.2 MB)
- PFI CSV file (43.2 MB)
- PFI SAS public-use setup files
  Includes SAS ASCII Read-in c
  Please note that the ASCII file
- NHES: 2016 PFI Public-Use Fi

Adult Training and Education

Data Access

National Death Index +

Public-Use Data Files –

Compressed Mortality File +

Data User Agreement

Inter-university Consortium for Political and Social Research

SPACE Program

Vital Statistics Online

Related Sites

Data Linkage

NCHS Data Visualization Gallery

Public-Use Data Files and Documentation

The National Center for Health Statistics (NCHS) is pleased to offer downloadable public-use data files through the Centers for Disease Control and Prevention's (CDC) FTP file server. Users of this service have access to data sets, documentation, and questionnaires from NCHS surveys and data collection systems. Downloading instructions are available in "readme" files.

Public-use data files are prepared and disseminated to provide access to the full scope of the data. This allows researchers to manipulate the data in a format appropriate for their analyses. NCHS makes every effort to release data collected through its surveys and data systems in a timely manner.

Users of NCHS public-use data files must comply with data use restrictions to ensure that the information will be used solely for statistical analysis or reporting purposes.

National Health and Nutrition Examination Survey (NHANES)

On This Pag

NHANES

National Hea Surveys

NVSS

NSFG

NHIS

NIS

LSOA

SLAITS

Data Linkage

Compressed File

SPACE Progr

From search for dataset, download, clean, normalize, join …

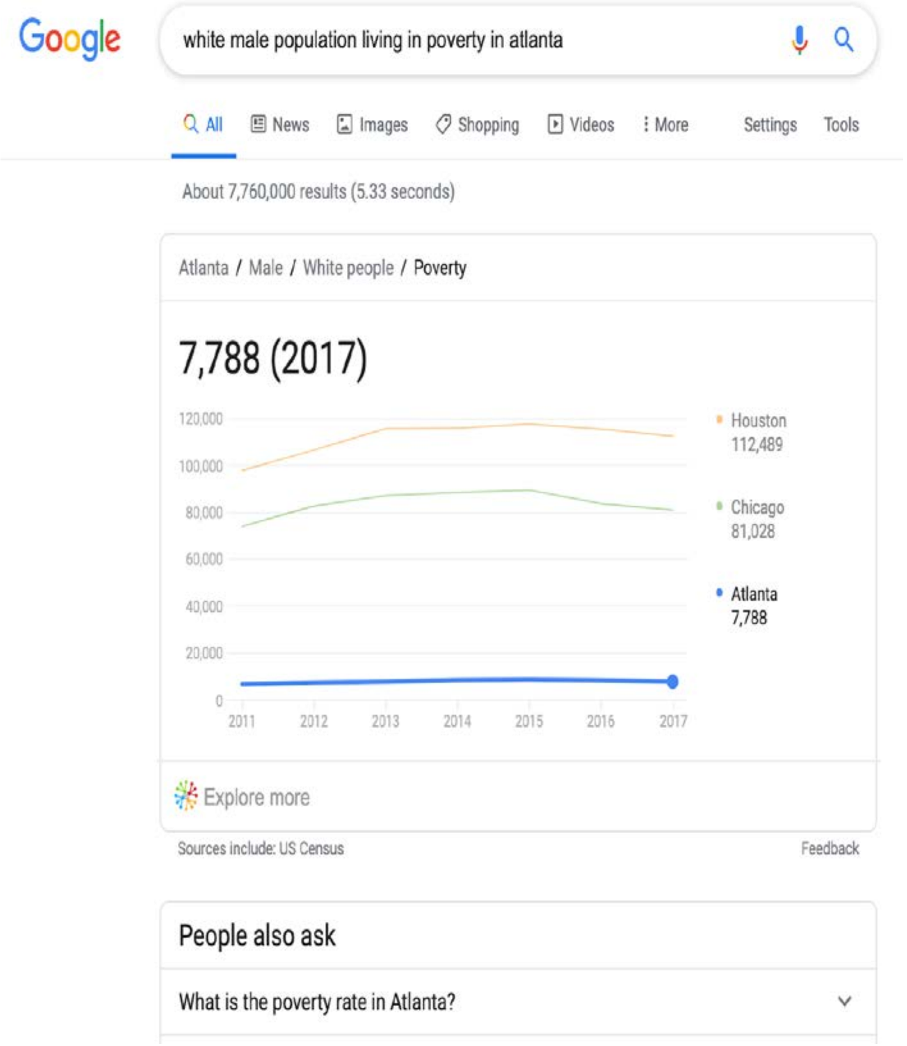
to

Just ask Google

Coming soon: In Google Search

- **Stats seeking queries** for O(1000) variables from Demographics, Education, Employment, Health, Commerce, Disasters, Crime, and Housing

- **Line charts**
- **US data** from federal government; international data from World Bank, OECD



white male population living in poverty in atlanta

About 7,760,000 results (5.33 seconds)

Atlanta / Male / White people / Poverty

7,788 (2017)

Sources include: US Census

People also ask

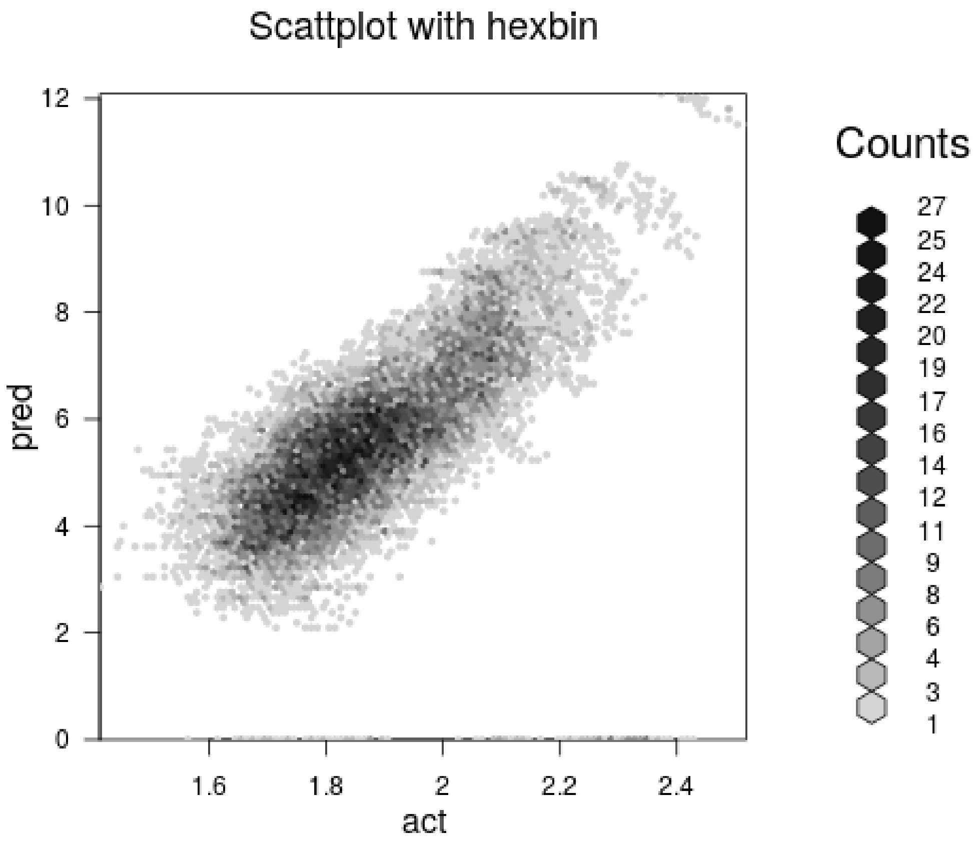
What is the poverty rate in Atlanta?

Important goal: Joining data across providers

Sites like FRED, Census, etc. allow joining data across their data sets by hosting the data and using common standards. But it is the common standards that really matter, since this allows joins across different distinct providers.

Start with dates, move on to geos.

E.g. I recently joined “county health rankings” (125 datasets) with “JHU covid confirmed cases by county” to build a predictive model at the county level.



Outline

- Schema.org
- Job search
- Dataset search
- Data commons
- Questions

Questions

1. What are best practices wrt to populating dataset metadata?
2. We are updating schema.org. What additional schemas do you want?
3. Who are important international partners?
4. Would “Job Trends” or “Recipe Trends” (modeled on Google Trends) be useful?