Building an Open Ecosystem for Data Discovery

Natasha Noy

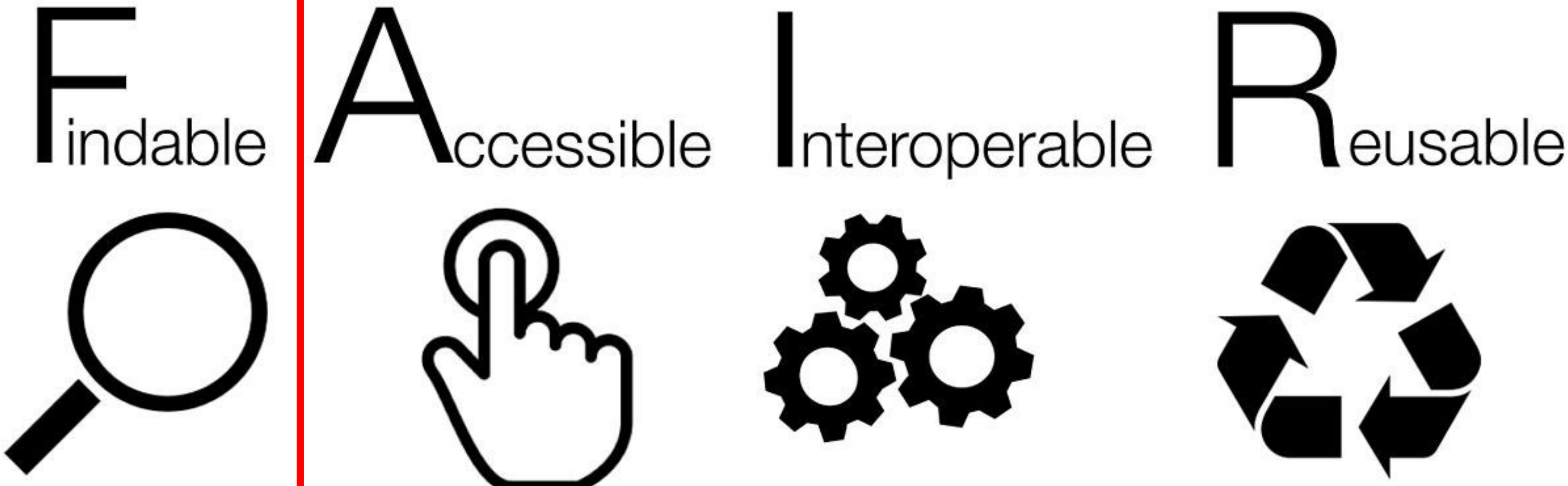
Google, Inc.

How do we publish data?

Source: United States public domain

It's the Web!

Source: Wikimedia commons; CC BY-SA 4.0



Source: Wikimedia commons; CC BY-SA 4.0

Why is finding data hard to do?

DataCite
FIND, ACCESS, AND REUSE DATA

1,660 Data Centers

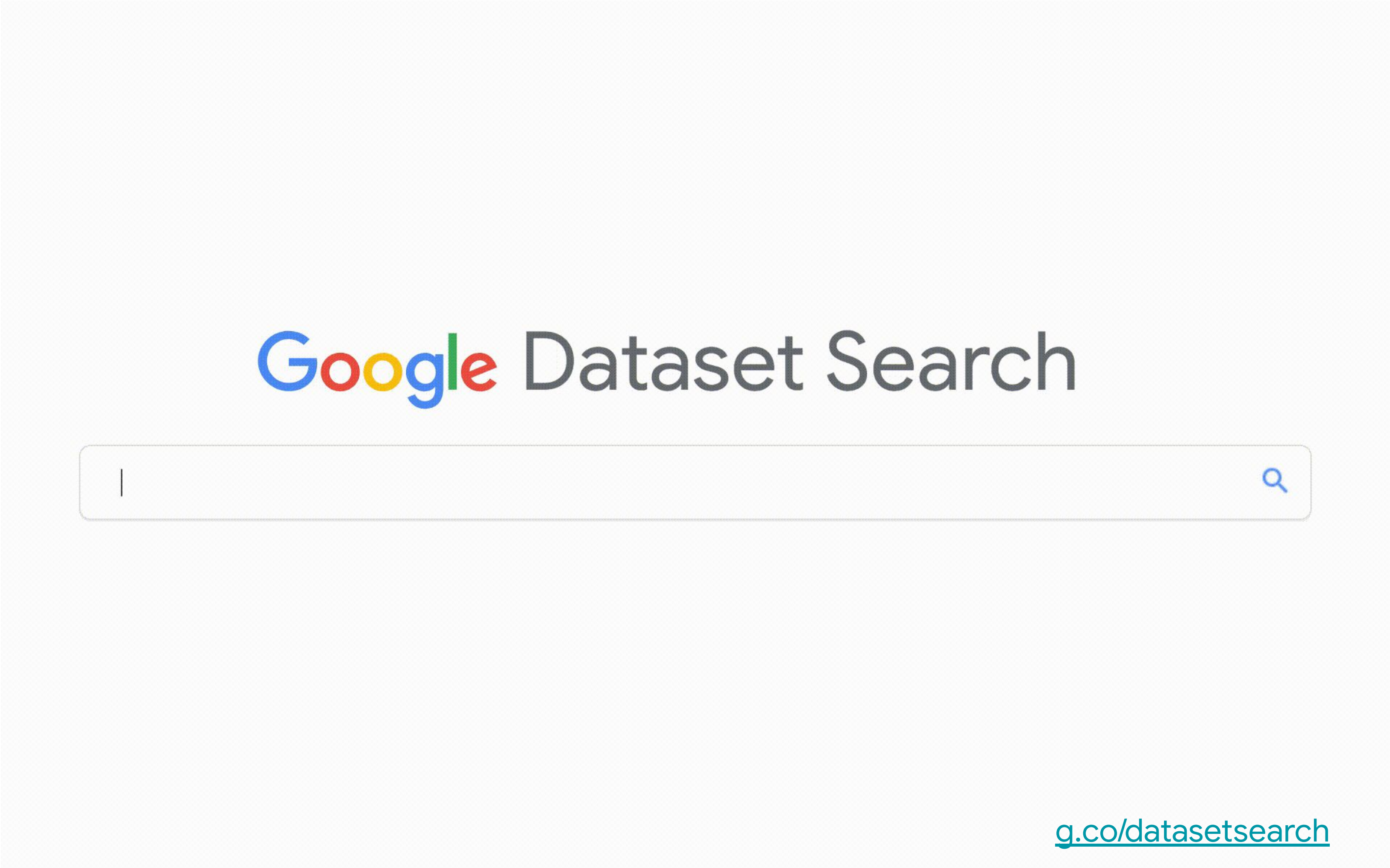
re3data.org*
REGISTRY OF RESEARCH DATA REPOSITORIES

Nature Scientific Data recommends 58 repositories

2,000 Data Repositories and Science Europe's Framework for Discipline-specific Research Data Management

- data.nodc.noaa.gov
- catalog.data.gov
- Kaggle
- Harvard Dataverse
- data.world
- www.europeandataportal.eu
- data.nasa.gov
- icpsr.umich.edu
- figshare.com
- zenodo.org
- data.opendatanetwork.com
- datadryad.org

What is Dataset Search?



Google Dataset Search

g.co/datasetsearch

Why do we need Dataset Search?

There is **proliferation** of datasets and dataset repositories

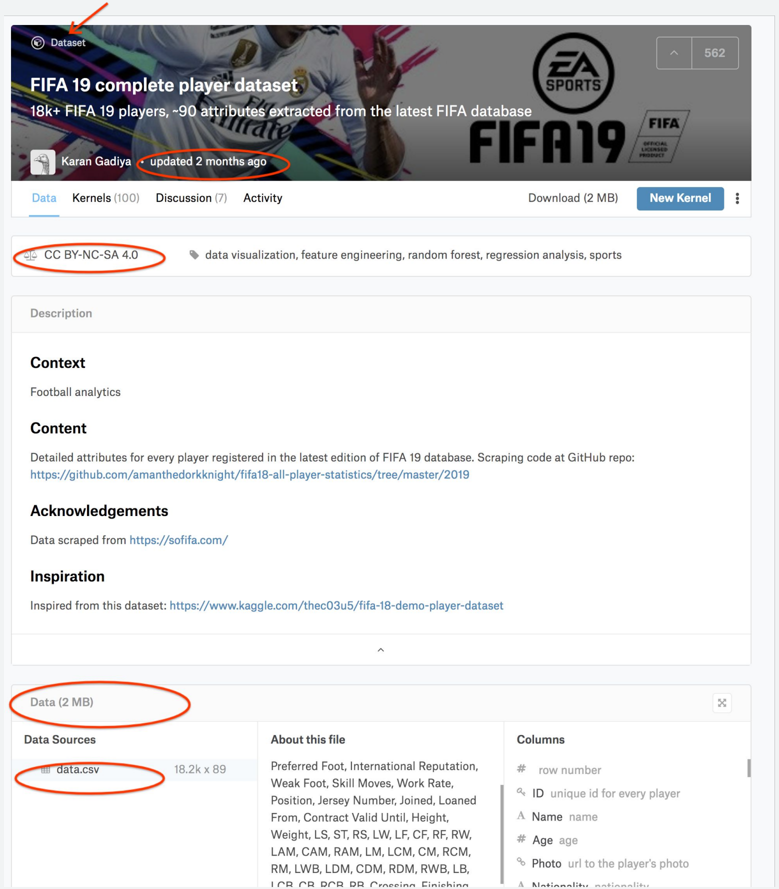
Dataset pages are often in the "**long tail**" of the Web

Data-publishing and research communities are very **specialized**

Where do we start?

Option 1: "Scrape" the metadata from various pages.

- Brittle:
  - Page layouts change
- We don't know where to look:
  - Should we look for a dataset on any web page?
- Somewhat pointless:
  - That metadata was structured in the first place



FIFA 19 complete player dataset

18k+ FIFA 19 players, ~90 attributes extracted from the latest FIFA database

Karan Gadiya • updated 2 months ago

CC BY-NC-SA 4.0

data visualization, feature engineering, random forest, regression analysis, sports

Description

Context

Football analytics

Content

Detailed attributes for every player registered in the latest edition of FIFA 19 database. Scraping code at GitHub repo: https://github.com/amanthedorkknight/fifa18-all-player-statistics/tree/master/2019

Acknowledgements

Data scraped from https://sofifa.com/

Inspiration

Inspired from this dataset: https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset

Data (2 MB)

Data Sources

data.csv 18.2k x 89

About this file

Preferred Foot, International Reputation, Weak Foot, Skill Moves, Work Rate, Position, Jersey Number, Joined, Loaned From, Contract Valid Until, Height, Weight, LS, ST, RS, LW, LF, CF, RF, RW, LAM, CAM, RAM, LM, LCM, CM, RCM, RM, LWB, LDM, CDM, RDM, RWB, LB, LCB, CB, RCB, RB, Crossing, Finishing

Columns

#		row number
ID		unique id for every player
A	Name	name
#	Age	age
Photo		url to the player's photo
A	Nationality	nationality

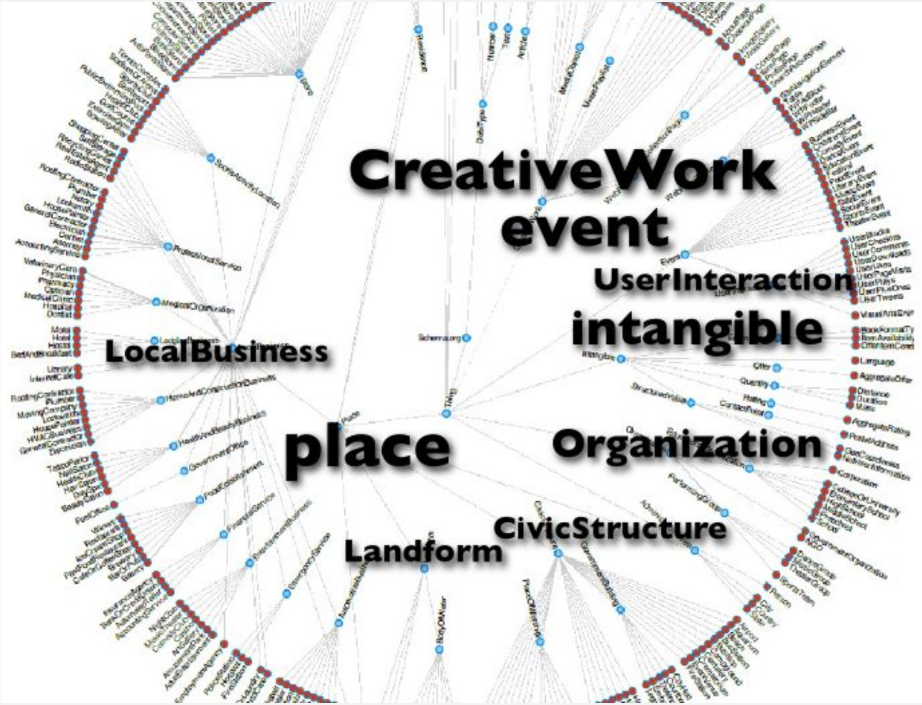
Part of the solution: Structured data (schema.org)

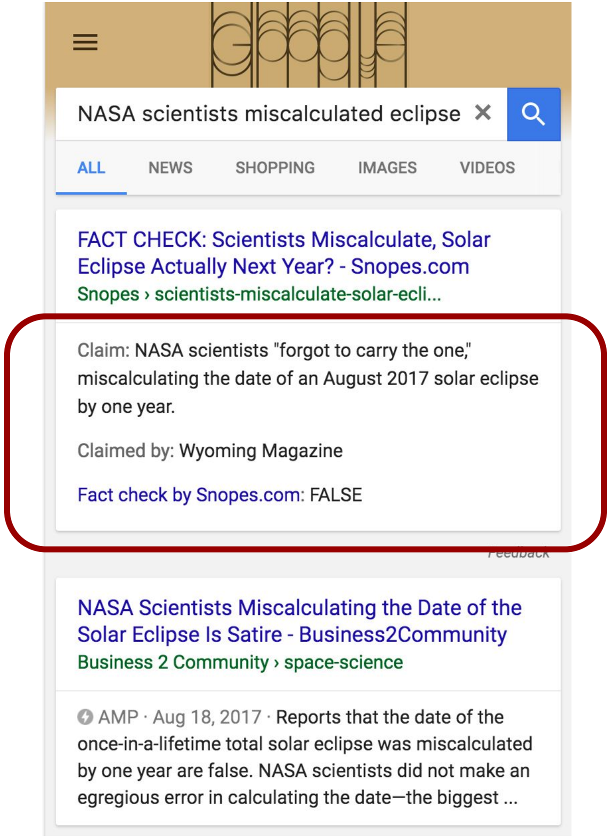
Structured data markup on the Web

- Founded by search engines in 2011
- Google Bing Yahoo! Yandex
- Big! widely used in the Web (⅓-ish)

Embedded in Web pages

Adoption driven by its use in real search products





NASA scientists miscalculated eclipse

ALL NEWS SHOPPING IMAGES VIDEOS

FACT CHECK: Scientists Miscalculate, Solar Eclipse Actually Next Year? - Snopes.com

Snopes › scientists-miscalculate-solar-ecli...

Claim: NASA scientists "forgot to carry the one," miscalculating the date of an August 2017 solar eclipse by one year.

Claimed by: Wyoming Magazine

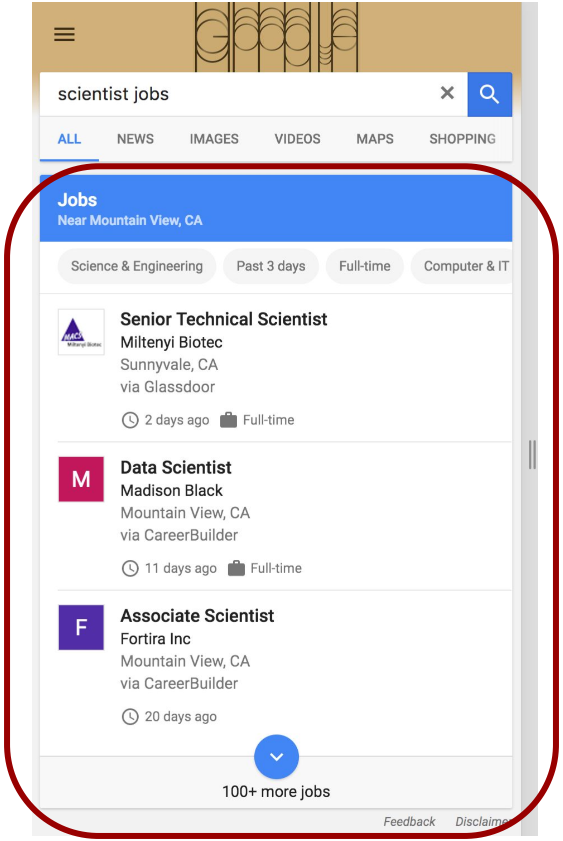
Fact check by Snopes.com: FALSE

Feedback

NASA Scientists Miscalculating the Date of the Solar Eclipse Is Satire - Business2Community

Business 2 Community › space-science

AMP · Aug 18, 2017 · Reports that the date of the once-in-a-lifetime total solar eclipse was miscalculated by one year are false. NASA scientists did not make an egregious error in calculating the date—the biggest ...



scientist jobs

ALL NEWS IMAGES VIDEOS MAPS SHOPPING

Jobs

Near Mountain View, CA

Science & Engineering | Past 3 days | Full-time | Computer & IT

Senior Technical Scientist
Miltenyi Biotec
Sunnyvale, CA
via Glassdoor
2 days ago · Full-time

Data Scientist
Madison Black
Mountain View, CA
via CareerBuilder
11 days ago · Full-time

Associate Scientist
Fortira Inc
Mountain View, CA
via CareerBuilder
20 days ago

100+ more jobs

Feedback Disclaimer

Why schema.org?

It's an open standard

Adoption driven by use in real search products

Embedded in HTML

Anybody can read and crawl this metadata

And build tools over it

It is really easy to add it

Our goal:
build a search engine for all datasets on the Web

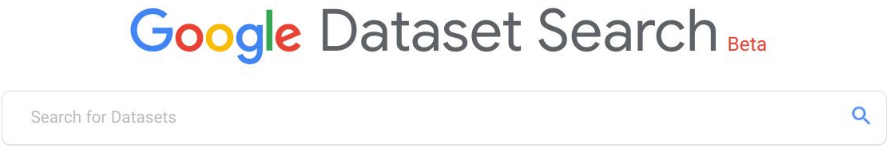
Open ecosystem: any data provider can join

Metadata: open
Data: can be open, require a license, etc.

Open standards: Web-friendly, community based

Inside Dataset Search

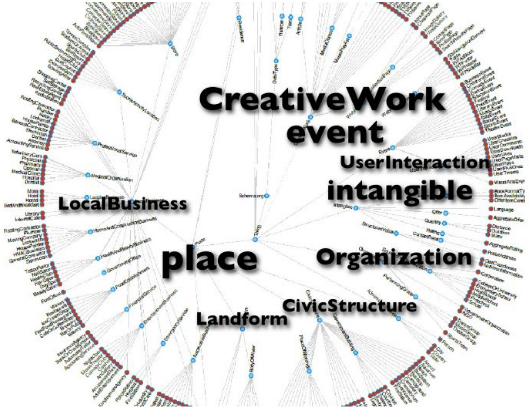
What is Dataset Search?



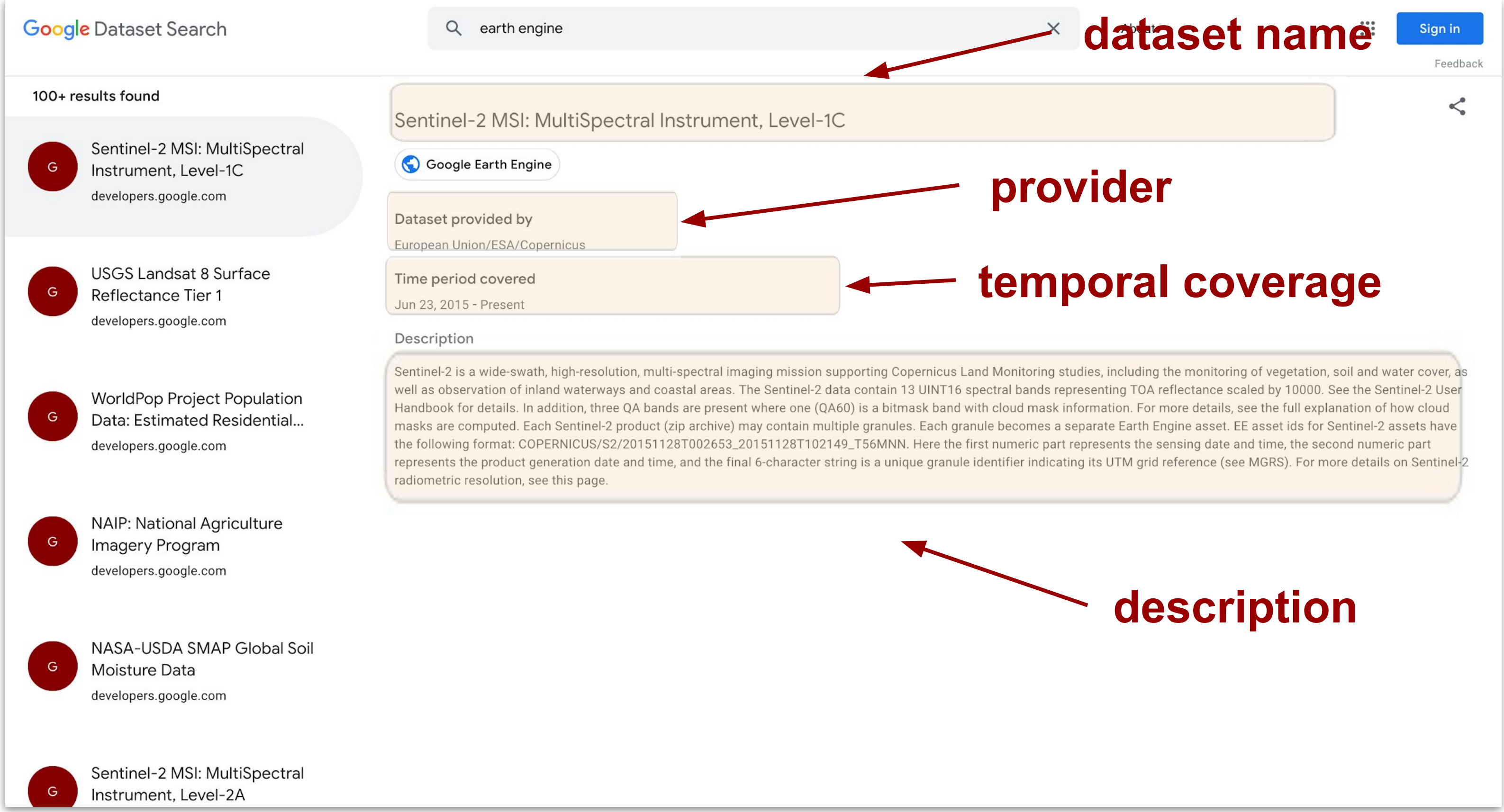
Google Dataset Search Beta

Search for Datasets

It's a search engine



It's a search engine over metadata



100+ results found

Sentinel-2 MSI: MultiSpectral Instrument, Level-1C
developers.google.com

USGS Landsat 8 Surface Reflectance Tier 1
developers.google.com

WorldPop Project Population Data: Estimated Residential...
developers.google.com

NAIP: National Agriculture Imagery Program
developers.google.com

NASA-USDA SMAP Global Soil Moisture Data
developers.google.com

Sentinel-2 MSI: MultiSpectral Instrument, Level-2A

Sentinel-2 MSI: MultiSpectral Instrument, Level-1C

Google Earth Engine

Dataset provided by
European Union/ESA/Copernicus

Time period covered
Jun 23, 2015 - Present

Description

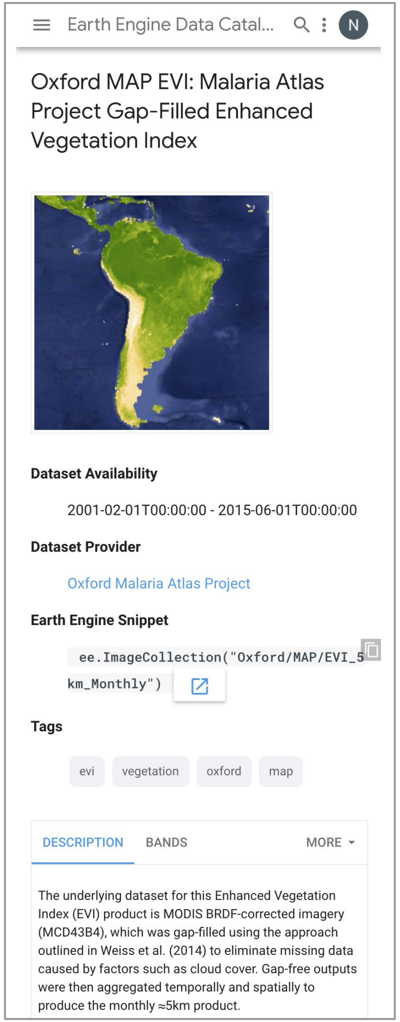
Sentinel-2 is a wide-swath, high-resolution, multi-spectral imaging mission supporting Copernicus Land Monitoring studies, including the monitoring of vegetation, soil and water cover, as well as observation of inland waterways and coastal areas. The Sentinel-2 data contain 13 UINT16 spectral bands representing TOA reflectance scaled by 10000. See the Sentinel-2 User Handbook for details. In addition, three QA bands are present where one (QA60) is a bitmask band with cloud mask information. For more details, see the full explanation of how cloud masks are computed. Each Sentinel-2 product (zip archive) may contain multiple granules. Each granule becomes a separate Earth Engine asset. EE asset ids for Sentinel-2 assets have the following format: COPERNICUS/S2/20151128T002653_20151128T102149_T56MNN. Here the first numeric part represents the sensing date and time, the second numeric part represents the product generation date and time, and the final 6-character string is a unique granule identifier indicating its UTM grid reference (see MGRS). For more details on Sentinel-2 radiometric resolution, see this page.

dataset name

provider

temporal coverage

description



Oxford MAP EVI: Malaria Atlas Project Gap-Filled Enhanced Vegetation Index

Dataset Availability

2001-02-01T00:00:00 - 2015-06-01T00:00:00

Dataset Provider

Oxford Malaria Atlas Project

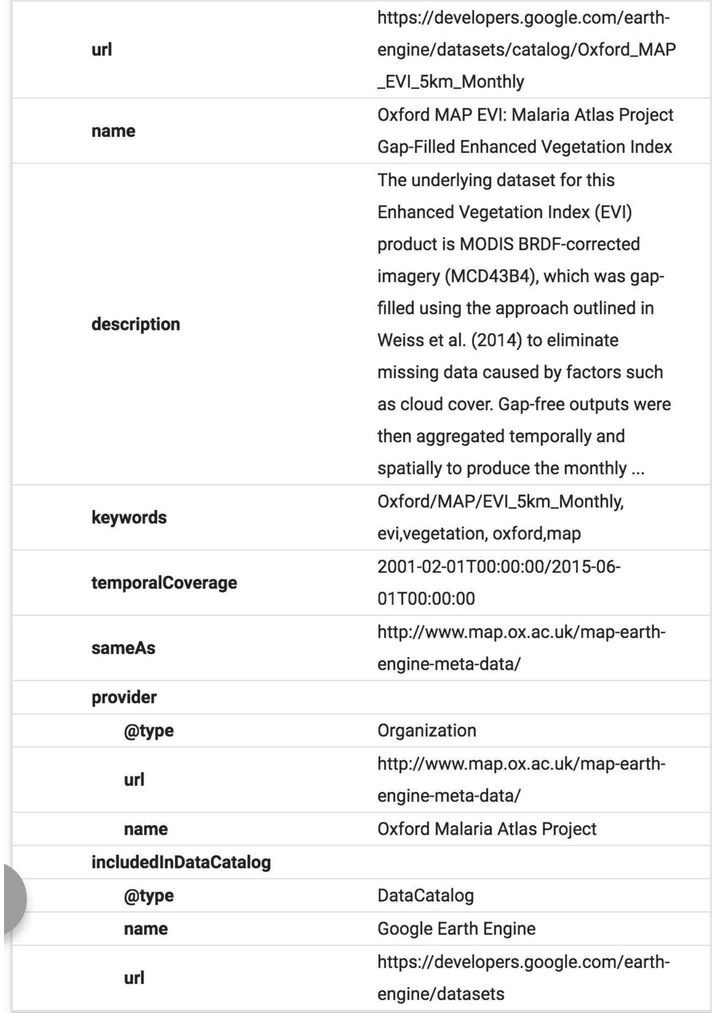
Earth Engine Snippet

```
ee.ImageCollection("Oxford/MAP/EVI_5km_Monthly")
```

Tags

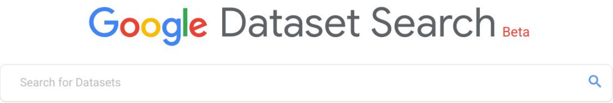
evi vegetation oxford map

The underlying dataset for this Enhanced Vegetation Index (EVI) product is MODIS BRDF-corrected imagery (MCD43B4), which was gap-filled using the approach outlined in Weiss et al. (2014) to eliminate missing data caused by factors such as cloud cover. Gap-free outputs were then aggregated temporally and spatially to produce the monthly ≈5km product.

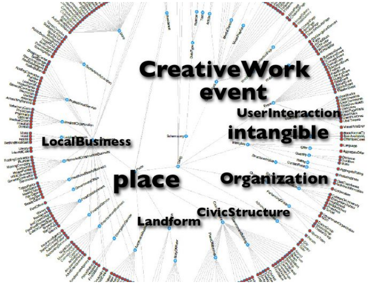


url	https://developers.google.com/earth-engine/datasets/catalog/Oxford_MAP_EVI_5km_Monthly
name	Oxford MAP EVI: Malaria Atlas Project Gap-Filled Enhanced Vegetation Index
description	The underlying dataset for this Enhanced Vegetation Index (EVI) product is MODIS BRDF-corrected imagery (MCD43B4), which was gap-filled using the approach outlined in Weiss et al. (2014) to eliminate missing data caused by factors such as cloud cover. Gap-free outputs were then aggregated temporally and spatially to produce the monthly ...
keywords	Oxford/MAP/EVI_5km_Monthly, evi,vegetation, oxford,map
temporalCoverage	2001-02-01T00:00:00/2015-06-01T00:00:00
sameAs	http://www.map.ox.ac.uk/map-earth-engine-meta-data/
provider	
@type	Organization
url	http://www.map.ox.ac.uk/map-earth-engine-meta-data/
name	Oxford Malaria Atlas Project
includedInDataCatalog	
@type	DataCatalog
name	Google Earth Engine
url	https://developers.google.com/earth-engine/datasets

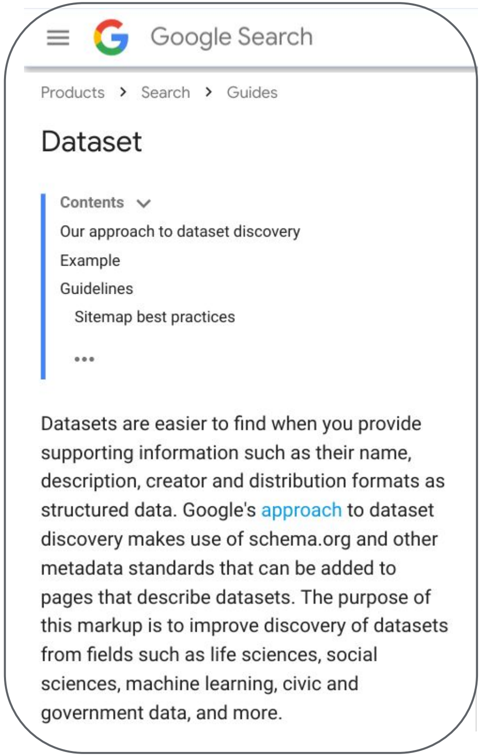
What is Dataset Search?



It's a search engine



It's a search engine over metadata



Dataset

Contents

Our approach to dataset discovery

Example

Guidelines

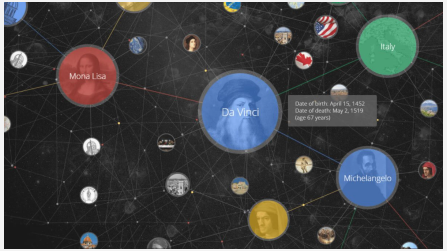
Sitemap best practices

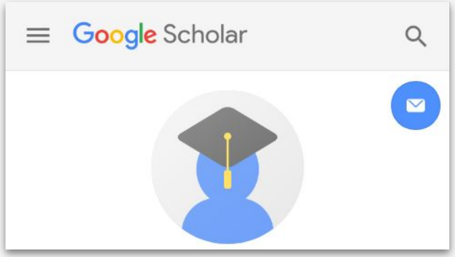
...

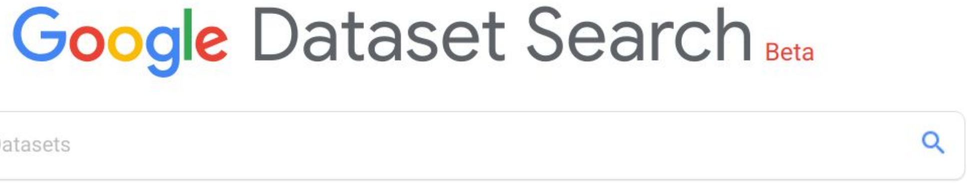
Datasets are easier to find when you provide supporting information such as their name, description, creator and distribution formats as structured data. Google's approach to dataset discovery makes use of schema.org and other metadata standards that can be added to pages that describe datasets. The purpose of this markup is to improve discovery of datasets from fields such as life sciences, social sciences, machine learning, civic and government data, and more.

It's a search engine over metadata from data providers thousands of domains

millions of datasets

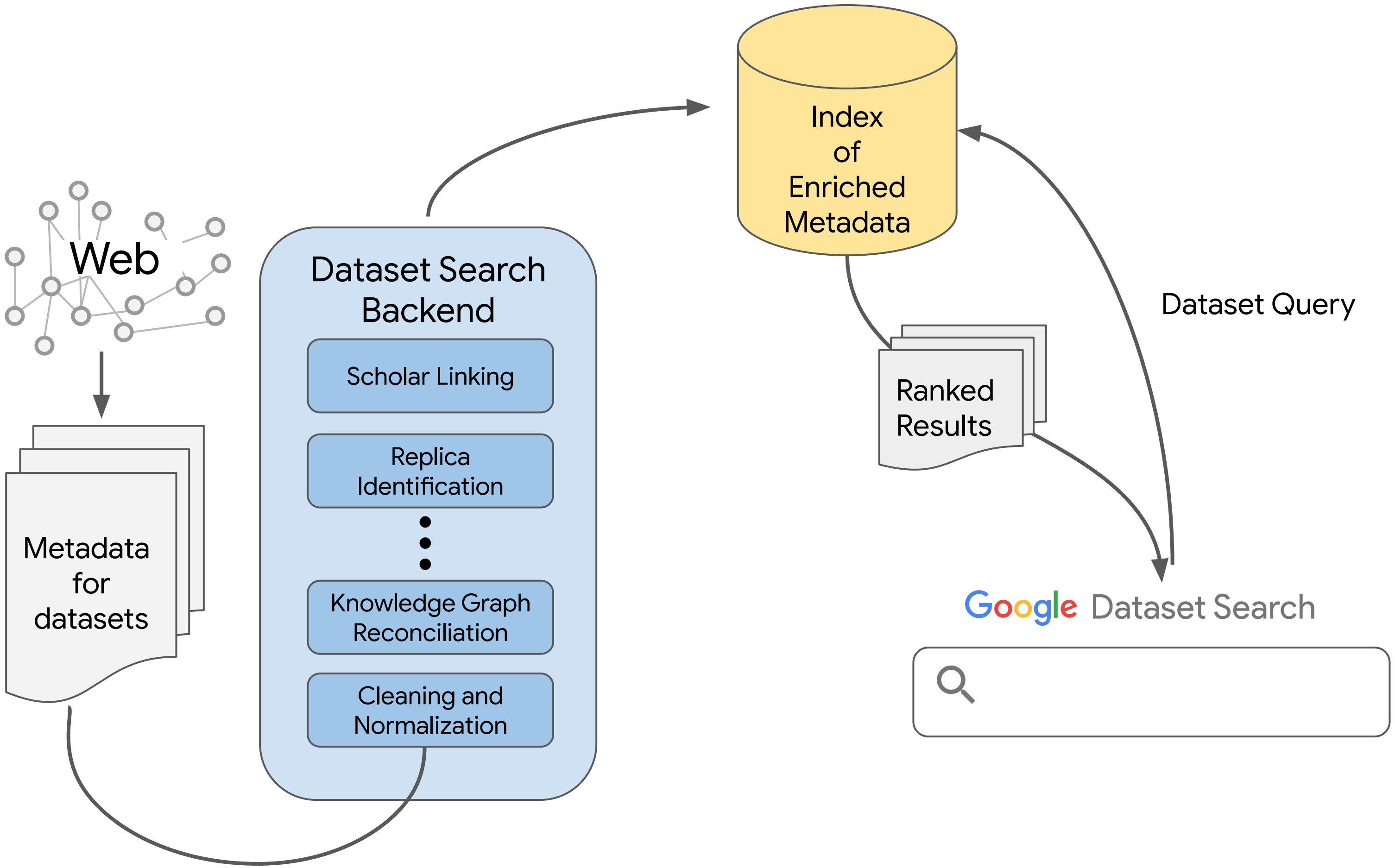






Google Dataset Search Beta

Search for Datasets



Web

Metadata for datasets

Dataset Search Backend

Scholar Linking

Replica Identification

Knowledge Graph Reconciliation

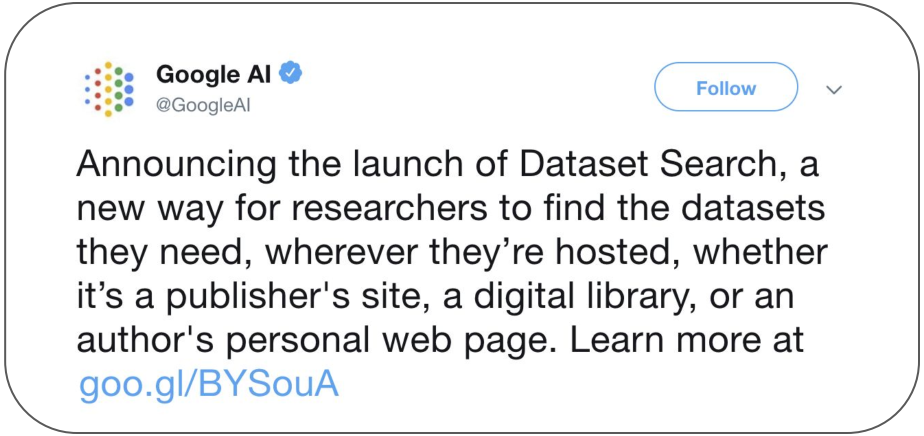
Cleaning and Normalization

Index of Enriched Metadata

Ranked Results

Dataset Query

Google Dataset Search



Announcing the launch of Dataset Search, a new way for researchers to find the datasets they need, wherever they're hosted, whether it's a publisher's site, a digital library, or an author's personal web page. Learn more at goo.gl/BYSouA

>3,800 repositories

>27M datasets

September 2018

August 2019

Lessons learned

Build an **ecosystem** first

    Don't jump to a heavy-weight technical solution

Open, non-proprietary standard is key

    When providers add metadata, it's not "just for Google"

Bootstrapping requires **influencers and incentives**

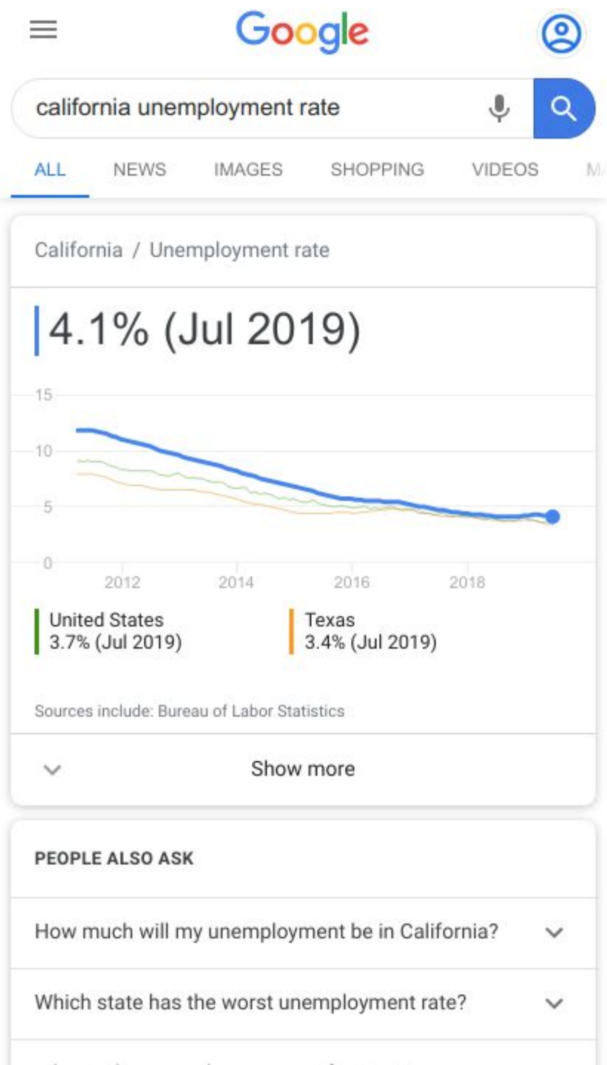
Beyond Metadata

Making statistics data more useful

Understand the data to enable search features

- Answers to factual questions
- Context about places, news events, issues
- Useful visualizations for comparison and insights

Requires open web-friendly formats for dataset **content**



California / Unemployment rate

4.1% (Jul 2019)

United States
3.7% (Jul 2019)

Texas
3.4% (Jul 2019)

Sources include: Bureau of Labor Statistics

PEOPLE ALSO ASK

How much will my unemployment be in California?

Which state has the worst unemployment rate?

DataSet Publishing Language (DSPL)

Schema.org-based format to describe public statistics datasets

Data: Time series and codelists, represented as CSV files (or triples)

Metadata:

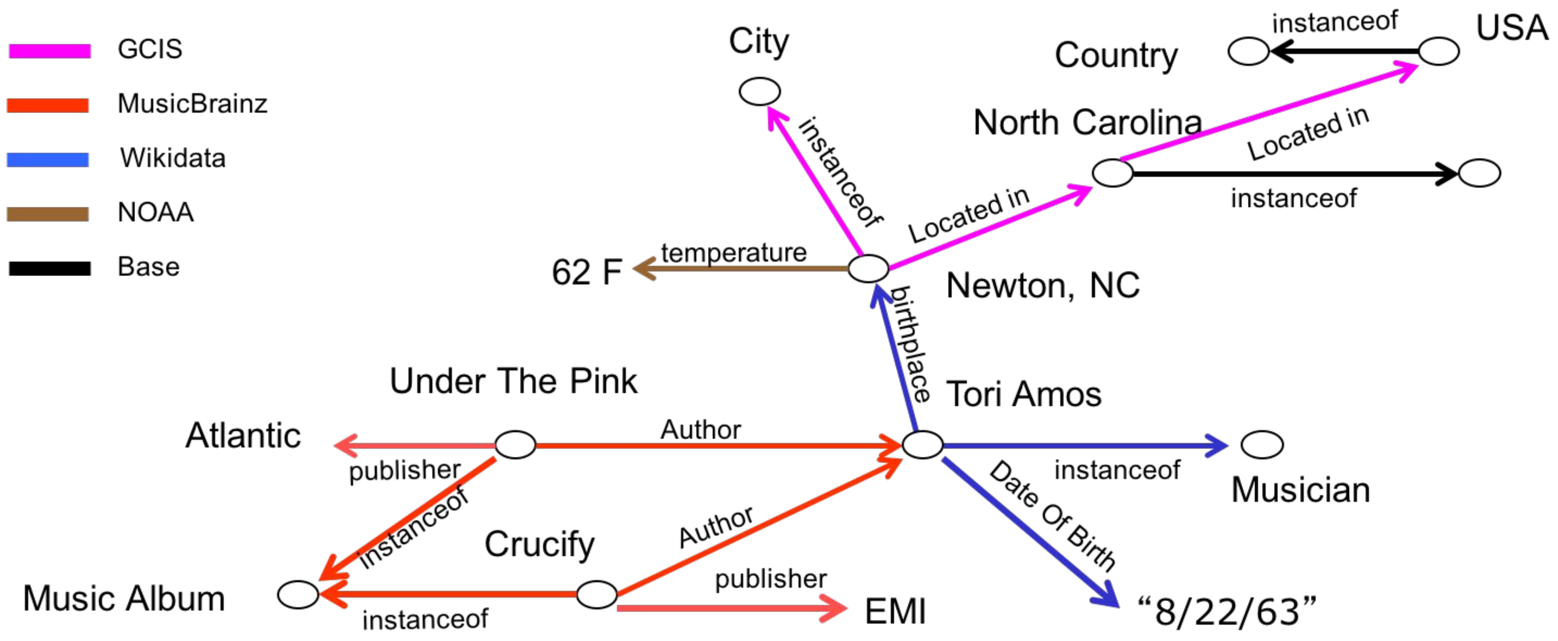
- schema.org/Dataset for general metadata
- Dimensions, Measures, Footnotes

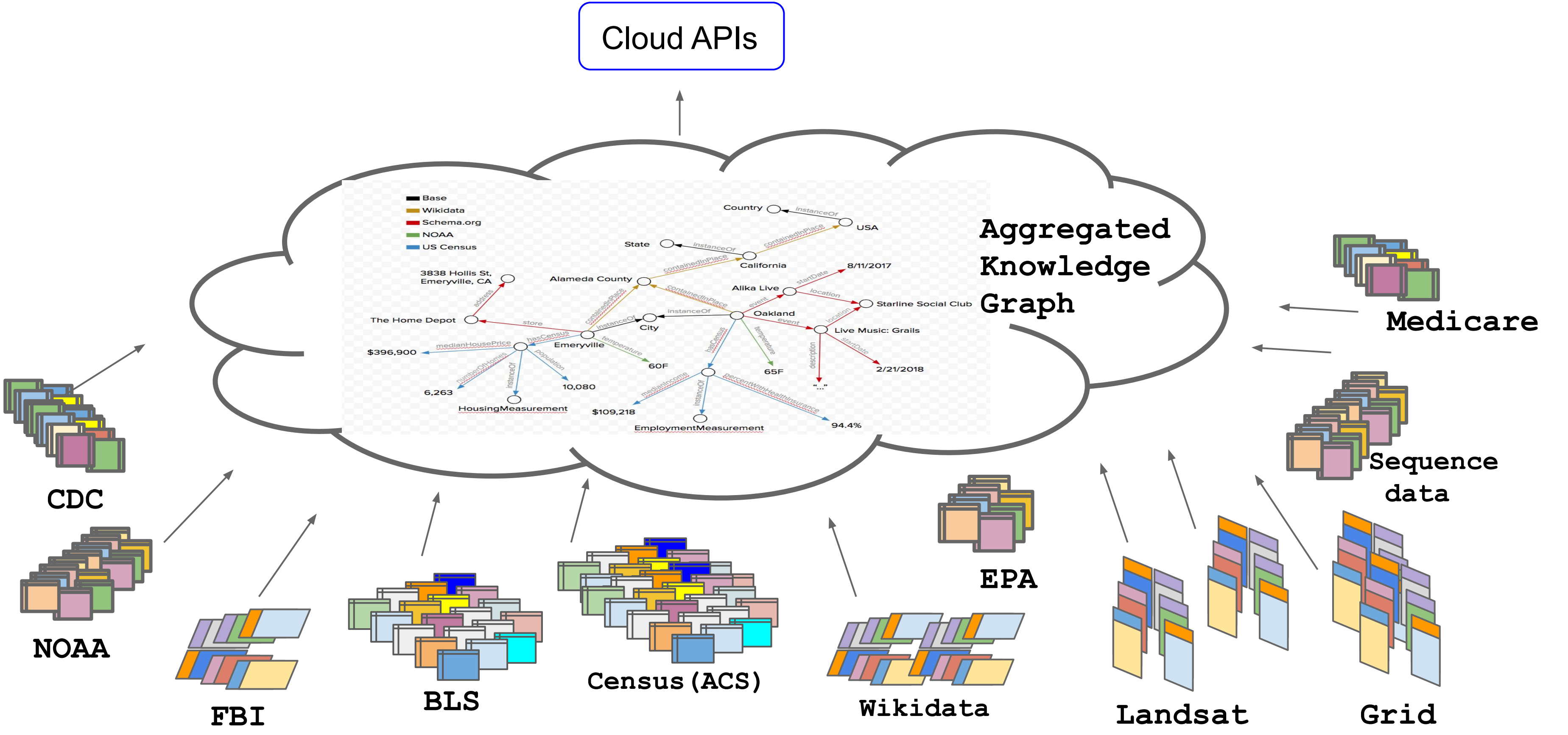
Data model is similar to SDMX, RDF Data Cube.

Documentation and samples at google.github.io/dspl/

Going further: Treat different sources as one database

datacommons.org





Cloud APIs

Aggregated Knowledge Graph

Medicare

Sequence data

CDC

NOAA

FBI

BLS

Census (ACS)

Wikidata

EPA

Landsat

Grid

Schema.org proposal for representing Aggregate Statistical Data

Creating a data-publishing ecosystem

Key challenges

Make it easy for scientists to share data and metadata in a **meaningful** way

Understand **incentives** for publishing metadata and data in a reusable way

Enable shared descriptions of **biases** and other **experimental conditions**

Google

Technical foundations

Long-term **storage**

- Landing pages crawlable by search engines

Persistent **identifiers**

- DOI (doi.org)
- identifiers.org

Structured **metadata**

- Web-friendly
- Standards-compliant
- schema.org

Clear **license description**

Source: pexels.com

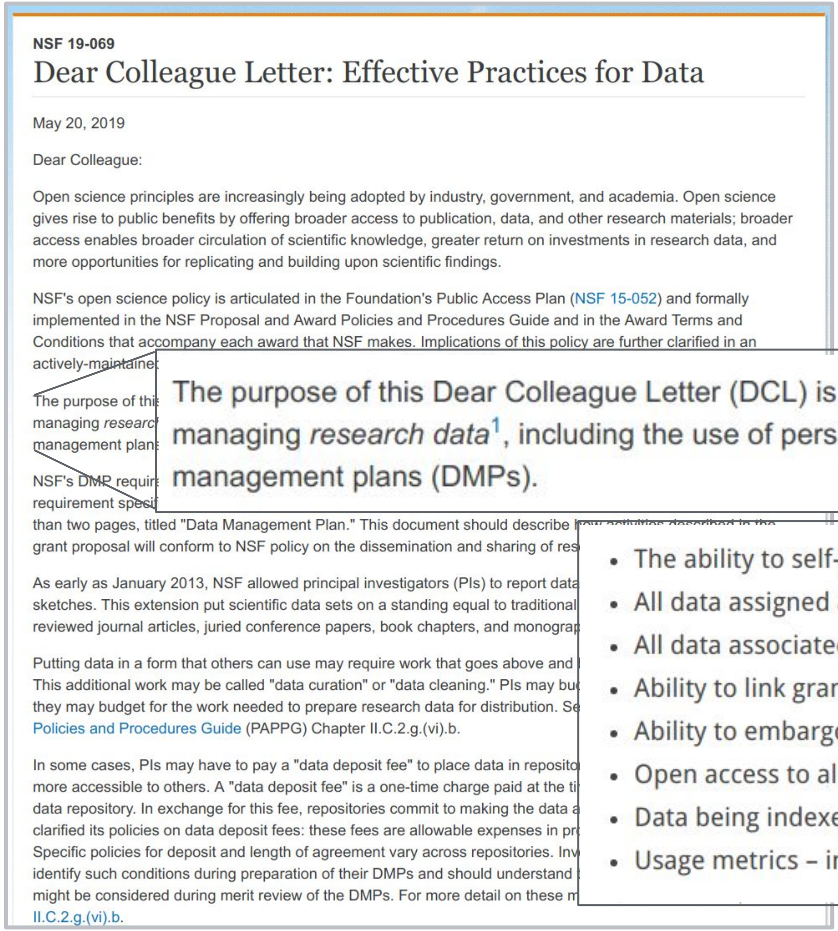
Incentives

Guidance and requirements from regulatory and funding agencies

Rewards and credit for publishing widely reused and cited data

Funding for the technical infrastructure that builds the foundation

Source: pexels.com



NSF 19-069

Dear Colleague Letter: Effective Practices for Data

May 20, 2019

Dear Colleague:

Open science principles are increasingly being adopted by industry, government, and academia. Open science gives rise to public benefits by offering broader access to publication, data, and other research materials; broader access enables broader circulation of scientific knowledge, greater return on investments in research data, and more opportunities for replicating and building upon scientific findings.

NSF's open science policy is articulated in the Foundation's Public Access Plan (NSF 15-052) and formally implemented in the NSF Proposal and Award Policies and Procedures Guide and in the Award Terms and Conditions that accompany each award that NSF makes. Implications of this policy are further clarified in an actively-maintaine

The purpose of thi
managing *researc*
management plan

NSF's DMP requir
requirement speci
than two pages, titled "Data Management Plan." This document should describe
grant proposal will conform to NSF policy on the dissemination and sharing of res

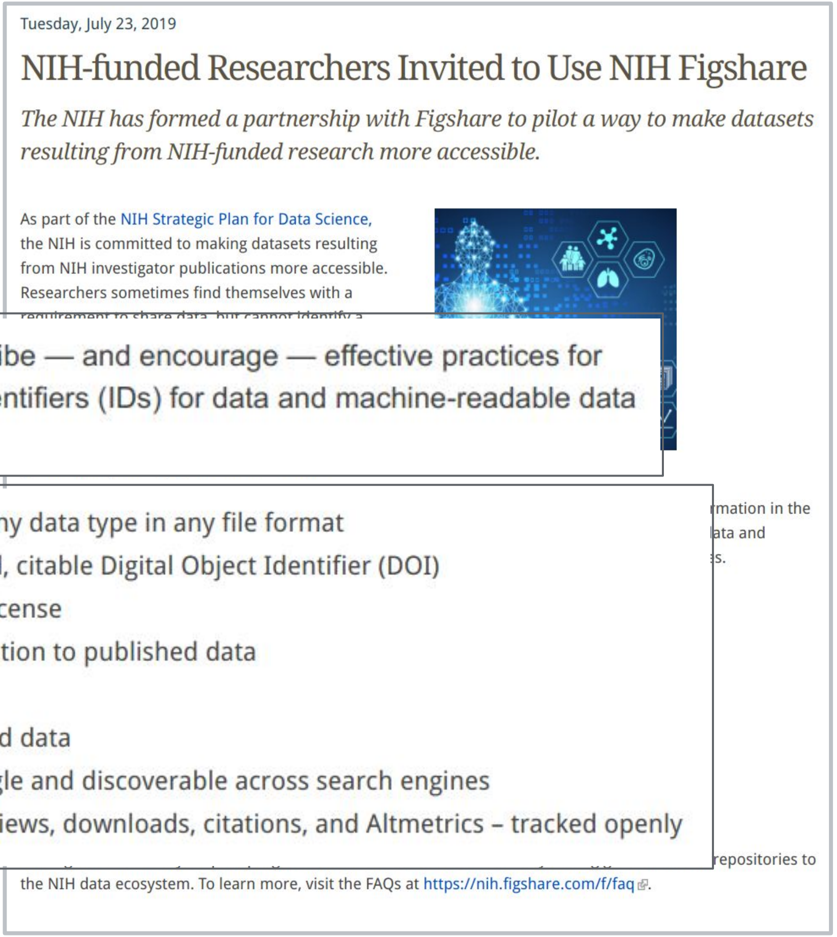
As early as January 2013, NSF allowed principal investigators (PIs) to report data
sketches. This extension put scientific data sets on a standing equal to traditional
reviewed journal articles, juried conference papers, book chapters, and monograp

Putting data in a form that others can use may require work that goes above and
This additional work may be called "data curation" or "data cleaning." PIs may bu
they may budget for the work needed to prepare research data for distribution. Se
Policies and Procedures Guide (PAPPG) Chapter II.C.2.g.(vi).b.

In some cases, PIs may have to pay a "data deposit fee" to place data in reposito
more accessible to others. A "data deposit fee" is a one-time charge paid at the ti
data repository. In exchange for this fee, repositories commit to making the data a
clarified its policies on data deposit fees: these fees are allowable expenses in pr
Specific policies for deposit and length of agreement vary across repositories. Inv
identify such conditions during preparation of their DMPs and should understand
might be considered during merit review of the DMPs. For more detail on these m
II.C.2.g.(vi).b.

The purpose of this Dear Colleague Letter (DCL) is to describe — and encourage — effective practices for managing *research data*[1], including the use of persistent identifiers (IDs) for data and machine-readable data management plans (DMPs).

https://www.nsf.gov/pubs/2019/nsf19069/nsf19069.jsp



Tuesday, July 23, 2019

NIH-funded Researchers Invited to Use NIH Figshare

The NIH has formed a partnership with Figshare to pilot a way to make datasets resulting from NIH-funded research more accessible.

As part of the NIH Strategic Plan for Data Science, the NIH is committed to making datasets resulting from NIH investigator publications more accessible. Researchers sometimes find themselves with a requirement to share data, but cannot identify a

rmation in the
ata and
s.

- The ability to self-publish any data type in any file format
- All data assigned a branded, citable Digital Object Identifier (DOI)
- All data associated with a license
- Ability to link grant information to published data
- Ability to embargo data
- Open access to all published data
- Data being indexed in Google and discoverable across search engines
- Usage metrics – including views, downloads, citations, and Altmetrics – tracked openly

repositories to
the NIH data ecosystem. To learn more, visit the FAQs at https://nih.figshare.com/f/faq.

https://datascience.nih.gov/news/nih-funded-researchers-invited-use-nih-figshare