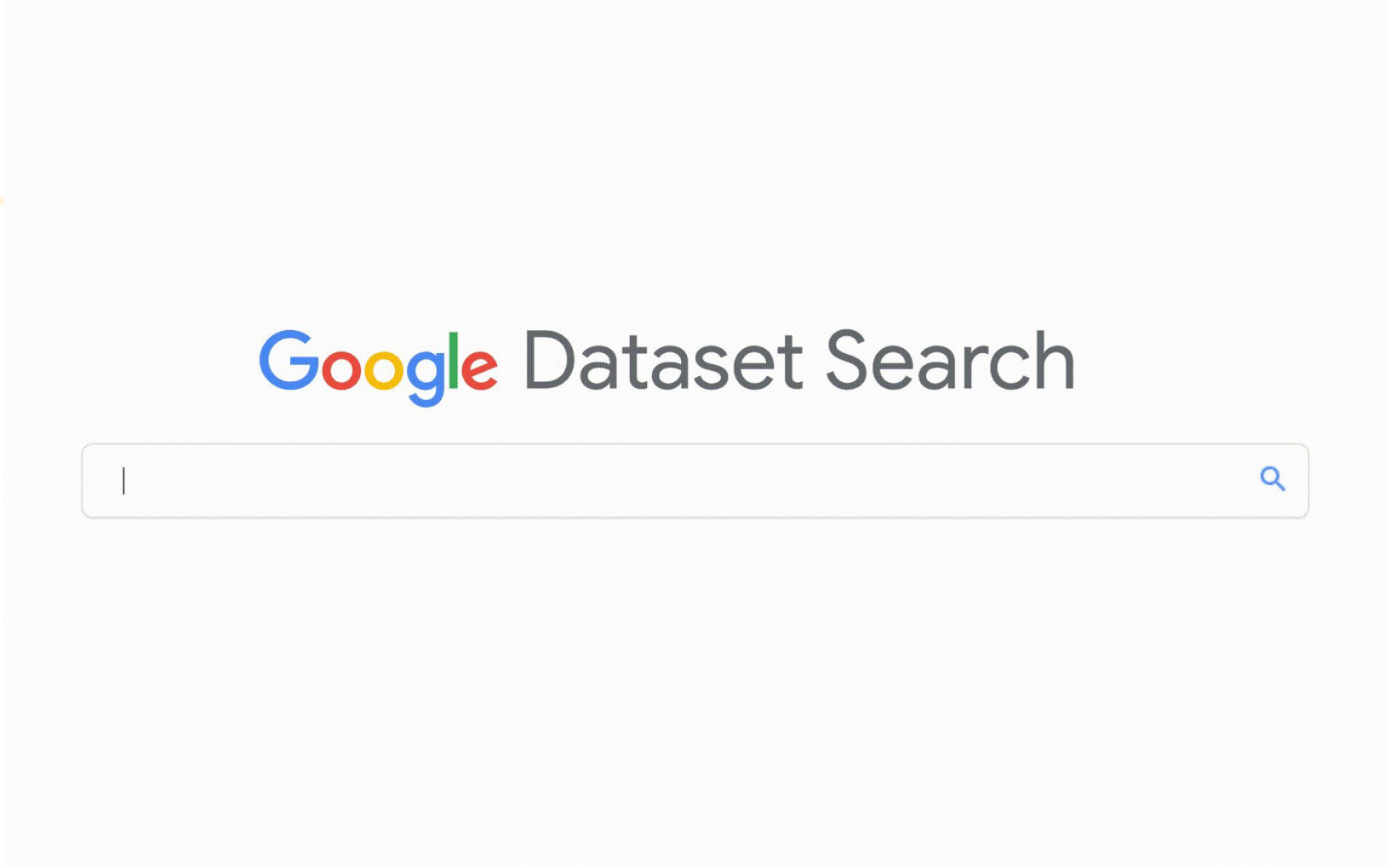
Facilitating Dataset Discovery in an Open Ecosystem

Natasha Noy
Google, Inc.


Google Dataset Search

Why Dataset Search?

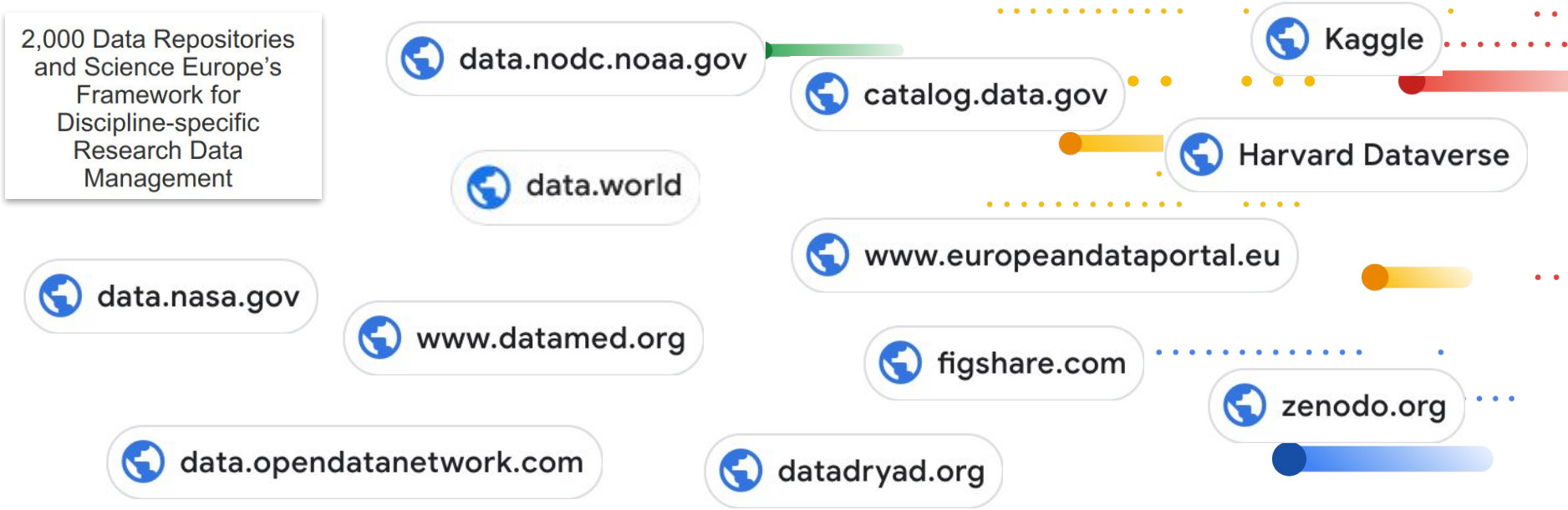
re3data.org*
REGISTRY OF RESEARCH DATA REPOSITORIES

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

DataCite
FIND, ACCESS, AND REUSE DATA

1,660 Data Centers

Nature Scientific Data recommends 58 repositories

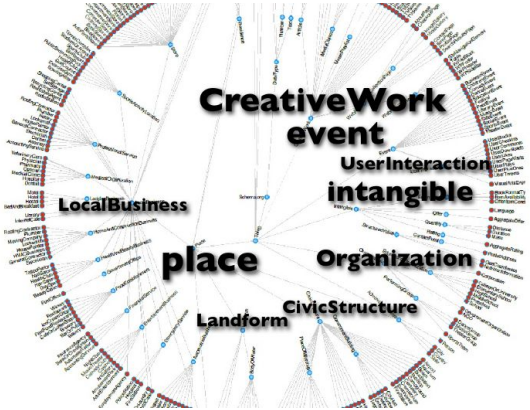


What is Dataset Search?

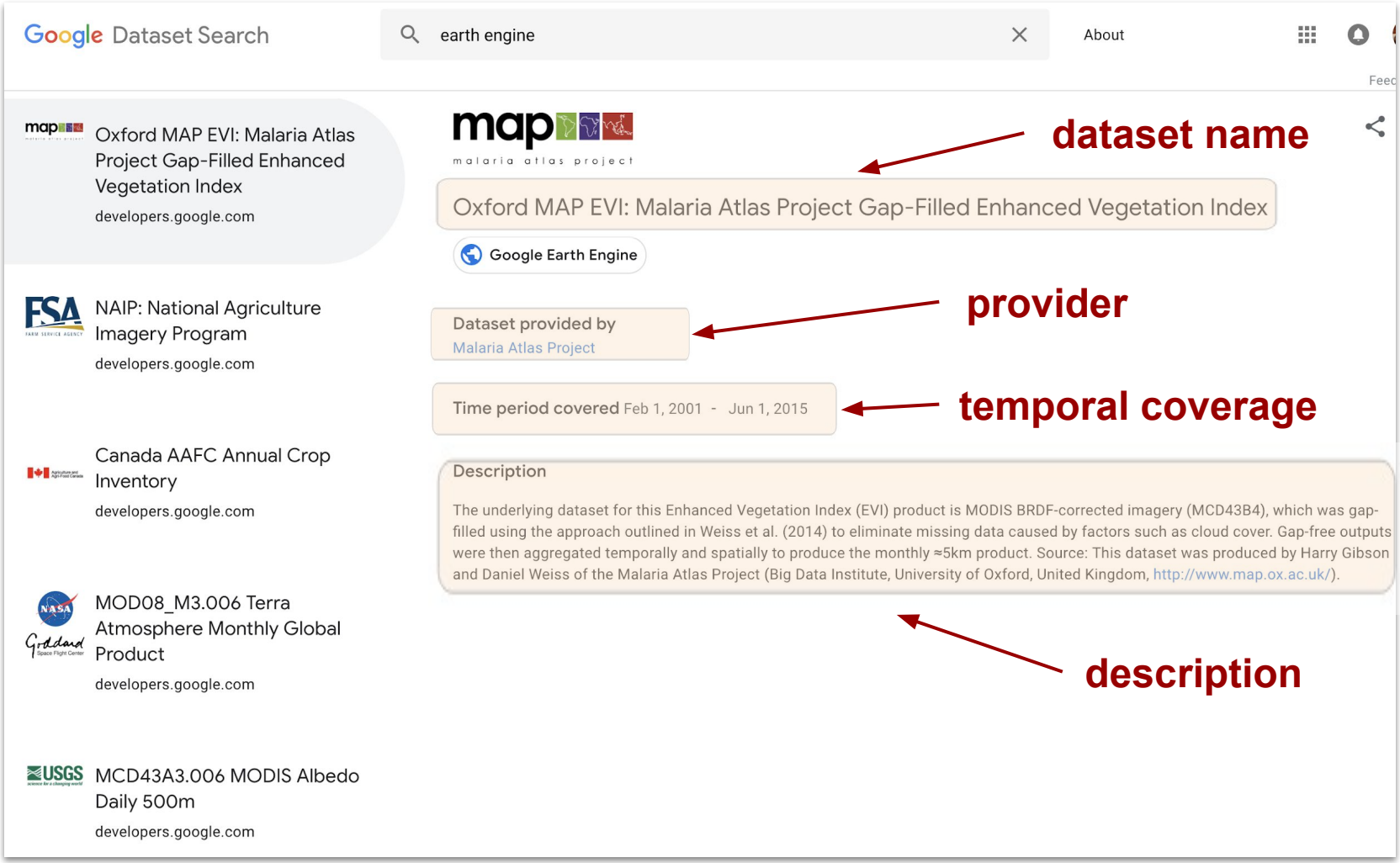
Google Dataset Search Beta

Search for Datasets

It's a search engine

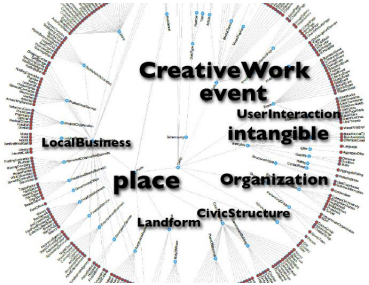


It's a search engine over metadata

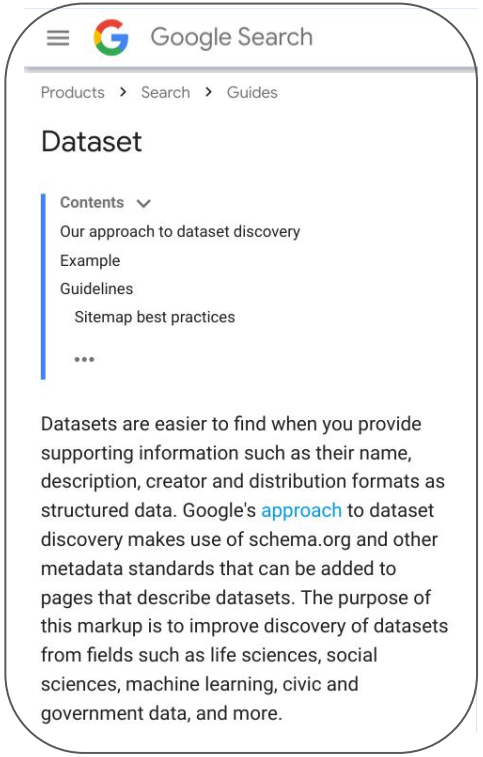

Google Dataset Search
earth engine
About
Oxford MAP EVI: Malaria Atlas Project Gap-Filled Enhanced Vegetation Index
developers.google.com
NAIP: National Agriculture Imagery Program
developers.google.com
Canada AAFC Annual Crop Inventory
developers.google.com
MOD08_M3.006 Terra Atmosphere Monthly Global Product
developers.google.com
MCD43A3.006 MODIS Albedo Daily 500m
developers.google.com
malaria atlas project
Oxford MAP EVI: Malaria Atlas Project Gap-Filled Enhanced Vegetation Index
Google Earth Engine
Dataset provided by
Malaria Atlas Project
Time period covered Feb 1, 2001 - Jun 1, 2015
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 ≈5km product. Source: This dataset was produced by Harry Gibson and Daniel Weiss of the Malaria Atlas Project (Big Data Institute, University of Oxford, United Kingdom, http://www.map.ox.ac.uk/).
dataset name
provider
temporal coverage
description

What is Dataset Search?

It's a search engine



It's a search engine over metadata

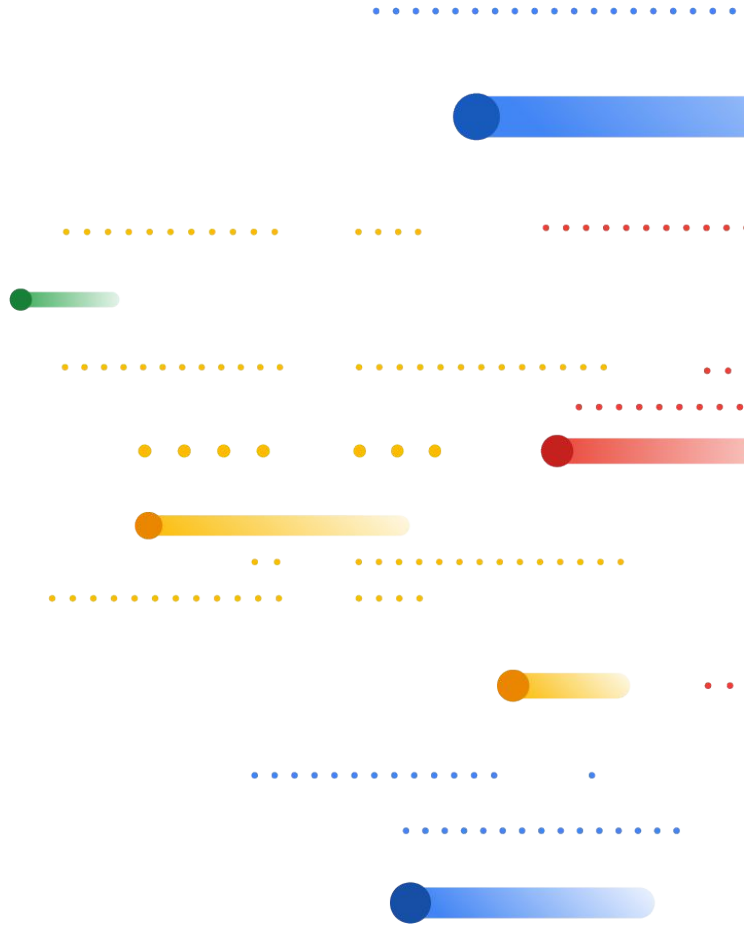


It's a search engine over metadata from data providers


<script type="application/ld+json">
{
"@context":"http://schema.org/",
"@type":"Dataset",
"name":"NCDC Storm Events Database",
"description":"Storm Data is provided by the National Weather Service (NWS) and contain statistics on...",
"url":"https://catalog.data.gov/dataset/ncdc-storm-events-database",
"sameAs":"https://gis.ncdc.noaa.gov/geoportal/catalog/search/resource/details.page?id=gov.noaa.ncdc:C00510",
"keywords":[
"ATMOSPHERE > ATMOSPHERIC PHENOMENA > CYCLONES",
"ATMOSPHERE > ATMOSPHERIC PHENOMENA > DROUGHT",
"ATMOSPHERE > ATMOSPHERIC PHENOMENA > FOG",
"ATMOSPHERE > ATMOSPHERIC PHENOMENA > FREEZE"
],
"creator":{
"@type":"Organization",
"url": "https://www.ncei.noaa.gov/",
"name":"OC/NOAA/NESDIS/NCEI > National Centers for Environmental Information, NESDIS, NOAA, U.S. Department of Commerce",
"contactPoint":{
"@type":"ContactPoint",
"contactType": "customer service",
"telephone":"+1-828-271-4800",
"email":"ncei.orders@noaa.gov"
}
},
"includedInDataCatalog":{
"@type":"DataCatalog",
"name":"data.gov"
},
"distribution":[
{
"@type":"DataDownload",
"encodingFormat":"CSV",
"contentUrl":"http://www.ncdc.noaa.gov/stormevents/ftp.jsp"
},
{
"@type":"DataDownload",
"encodingFormat":"XML",
"contentUrl":"http://gis.ncdc.noaa.gov/all-records/catalog/search/resource/details.page?id=gov.noaa.ncdc:C00510"
}
],
"temporalCoverage":"1950-01-01/2013-12-18",
"spatialCoverage":{
"@type":"Place",
"geo":{
"@type":"GeoShape",
"box":"18.0 -65.0 72.0 172.0"
}
}
}
</script>

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. We promise!



Where is my dataset?

Does the page have schema.org metadata?

Google Structured Data Testing Tool

Do you have a sitemap?

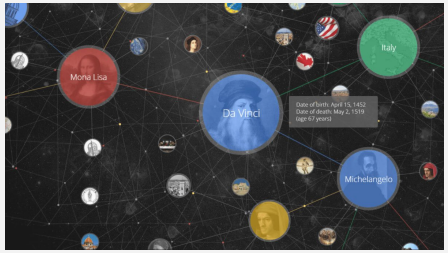
sitemaps.org

Has it been indexed?

Google Search Console

thousands of domains

millions of datasets



Google Dataset Search Beta

Search for Datasets

Beta

Google Dataset Search Beta

Search for Datasets

Next steps

Data providers

publish
structured metadata using community standards

Data consumers

cite
data properly, much as we cite scientific publications

Developers

contribute
to expanding metadata for datasets

Create a healthy data ecosystem