Dataset Search

A Google approach to data discovery.

Chris Gorgolewski, Ph.D.
Google LLC

Open data...

... improves research quality

... saves money

... promotes innovation

Wilkinson et al. 2016

Findable

Accessible

Interoperable

Reusable

Wilkinson et al. 2016

Findable

F1. (meta)data are assigned a globally unique and eternally persistent identifier.

F2. data are described with rich metadata.

F3. (meta)data are registered or **indexed in a searchable resource**.

F4. metadata specify the data identifier.

Nature Scientific Data recommends 58 repositories

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

data.nodc.noaa.gov

catalog.data.gov

Kaggle

Harvard Dataverse

data.world

www.europeandataportal.eu

data.nasa.gov

www.datamed.org

figshare.com

zenodo.org

data.opendatanetwork.com

datadryad.org

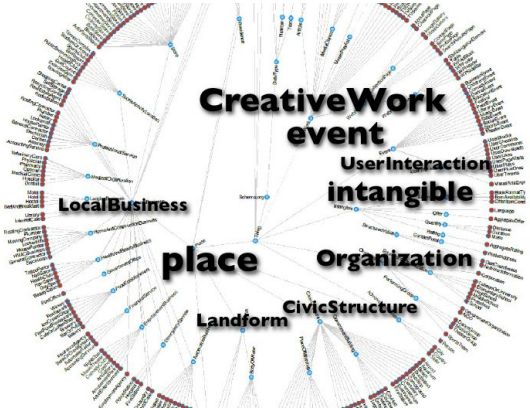
Our mission is to organize the world's information and make it universally accessible and useful.

What is Dataset Search?

Google Dataset Search Beta

Search for Datasets

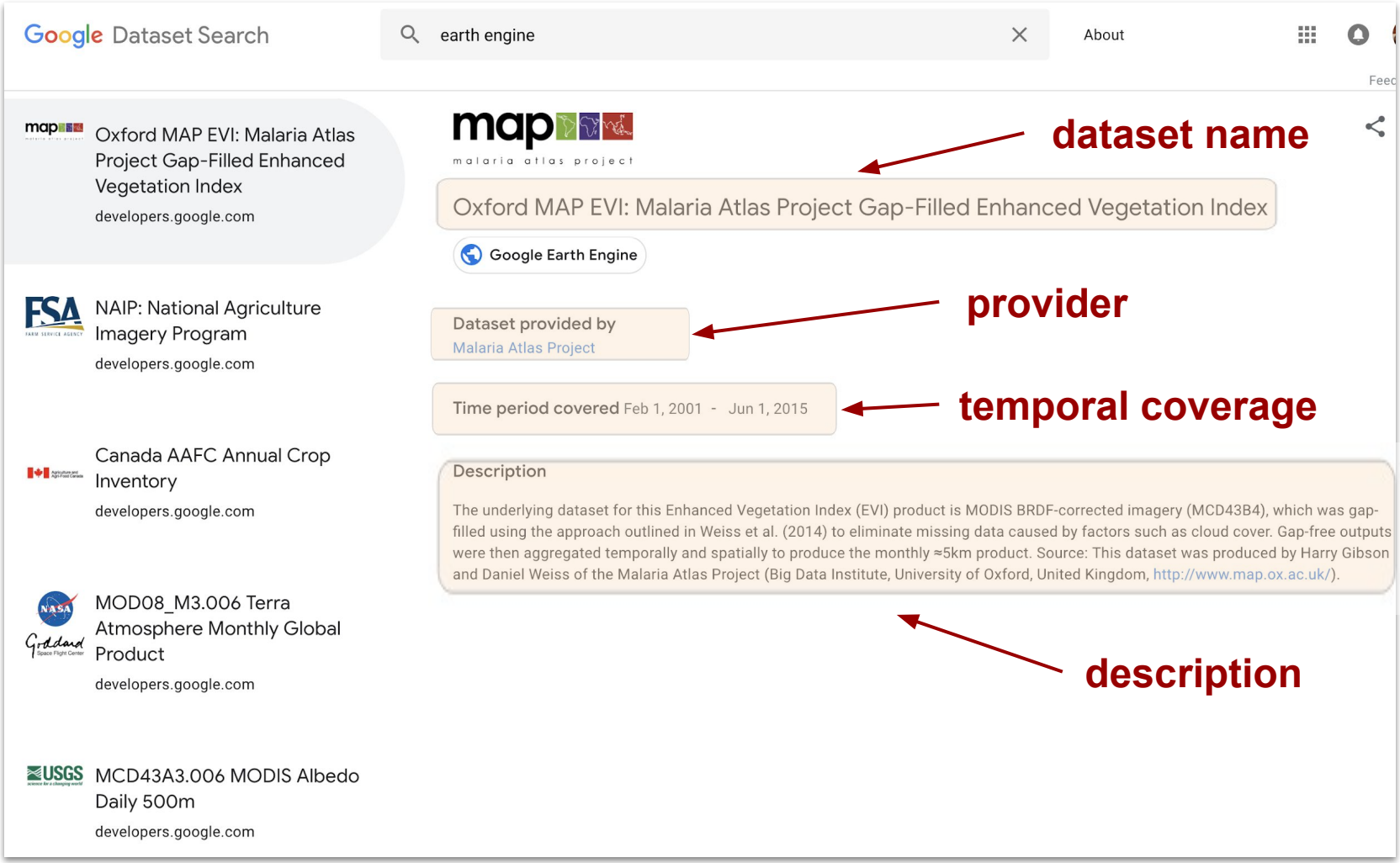
It's a search engine



It's a search engine over metadata

A search engine over **metadata**

- No need to access the actual data
- Supports special cases such as:
  - User agreements
  - Approval requirements
  - Fees

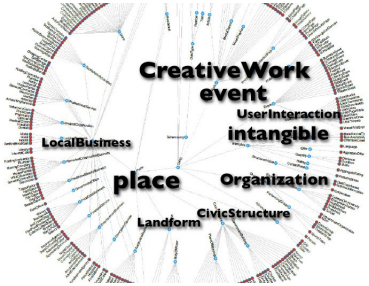

Google Dataset Search
earth engine
About
Oxford MAP EVI: Malaria Atlas Project Gap-Filled Enhanced Vegetation Index
developers.google.com
NAIP: National Agriculture Imagery Program
developers.google.com
Canada AAFC Annual Crop Inventory
developers.google.com
MOD08_M3.006 Terra Atmosphere Monthly Global 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?

Google Dataset Search Beta

Search for Datasets

It's a search engine



It's a search engine over metadata

Google Search

Products › Search › Guides

Dataset

Contents

Our approach to dataset discovery

Example

Guidelines

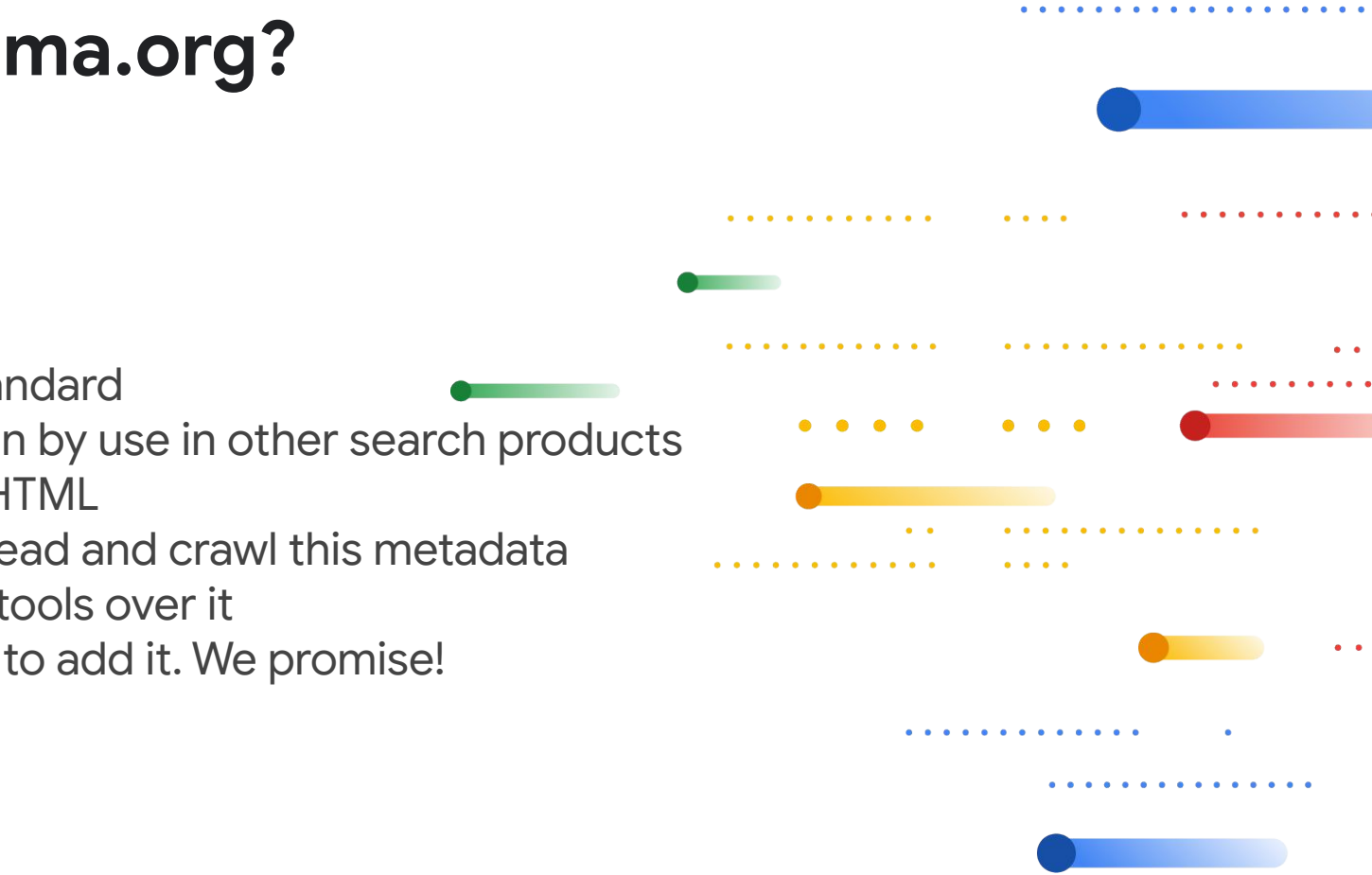
Sitemap best practices

...

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

Why schema.org?



- It's an open standard
- Adoption driven by use in other 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!

Schema.org - so simple it fits in a tweet

```
<script type="application/ld+json">
{
  "@context": "https://schema.org",
  "@type": "Dataset",
  "name": "My dataset",
  "description": "Description of my dataset"
}
</script>
```

Open ecosystem: any data provider can join

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

Open standards: Web-friendly, community based

20,000,000
datasets

4,000
data repositories

Thank You

Resources

g.co/DatasetSearch

- Developer documentation: https://developers.google.com/search/docs/data-types/dataset
- Mailing list for data providers: https://groups.google.com/forum/#!forum/datasetsearch-announce

Source: https://pixabay.com/photos/confused-hands-up-unsure-perplexed-2681507/