Discovering Earth Observation data

Despina-Athanasia Pantazi

Knowledge Technologies
2022-23

Using Google for dataset discovery

- Is there a land cover dataset produced by the European Environmental Agency covering the area of Chania, Crete, Greece?
- Google it!

Results

3

Let us pose a different query

- Is there a land cover dataset produced by the European Environment Agency?

Results

Google Dataset Search

Results

Google Dataset Search

- Datasets that are indexed using **schema.org**, as proposed by Google, show up.

- Enables users to find datasets stored across the Web by doing a simple keyword search.

- Uncovers information about datasets hosted in thousands of repositories across the web, making these datasets universally accessible and useful.

How does Dataset Search find datasets?

Authors need to add metadata in **schema.org** to each page that describes a dataset.

Schema.org:

- Founded by Google, Microsoft, Yahoo!, Yandex. Currently schema.org vocabularies are developed by an open community process.

- Provides a unique structured data markup schema to annotate a variety of topics.
  - Tags added to HTML as JSON-LD, Microdata, or RDFa.

- On-page markup allows search engines to understand information included in web pages.

Schema.org

```
<script type="application/ld+json">{
    "@context": "http://schema.org",
    "@type": ["ItemList", "Dataset"],
    "itemListOrder": "http://schema.org/ItemListOrderAscending",
    "numberOfItems": "7",
    "itemListElement": [{
        "@type": "ListItem", "position": 1,
        "item":{
            "@type" : "Dataset",
            "name": "GADM database of Global Administrative Areas",
            "alternateName": "GADM",
            "description": "GADM is a spatial database of the location ...",
            "author": "University of Athens",
            "sourceOrganization": "Robert Hijmans, in collaboration with ...",
            "copyrightYear": "2018",
            "keywords":["GADM", "Global Administrative Areas", "GADM 2015"],
            "spatialCoverage": "World", "temporalCoverage": "2015", "fileFormat": "7z",
            "isBasedOn": "GADM database of Global Administrative Areas 2015, Version 2.8",
            "isAccessibleForFree": true,
            "distribution": {
                "@type":"DataDownload", "encodingFormat":"7z",
                "contentUrl":"https://datahub.ckan.io/dataset/gadm"},
            "url": "https://datahub.ckan.io/dataset/gadm"}}, ... ]} </script>
```

Google Dataset Search

Google Dataset Search by the Numbers

❖ O. Benjelloun, S. Chen, N. Noy: **Google Dataset Search by the Numbers** (Jun 2020)
https://arxiv.org/abs/2006.06894

  ➢ As of March 2020, the corpus contained 28 million datasets from more than 3,700 sites

  ➢ The corpus is a reasonably representative snapshot of the datasets published on the Web, but there is no way of measuring how well the corpus covers all the datasets available on the Web

Google Dataset Search by the Numbers

❖ O. Benjelloun, S. Chen, N. Noy: **Google Dataset Search by the Numbers** (Jun 2020)
https://arxiv.org/abs/2006.06894

➢ **Licenses and access:** Only 34% of the datasets provide any licensing information

■ most of datasets available for free, almost always allowed reuse for both commercial and non-commercial purposes

➢ **Linked Data**: Fewer than 1% of datasets in the corpus are in linked data formats

■ there is plenty of shared data that the Semantic Web community produces, but the final step of describing it appears to be less common

Google Dataset Search - An analysis of online datasets

❖ Distribution of dataset topics

Reference: https://ai.googleblog.com/2020/08/an-analysis-of-online-datasets-using.html
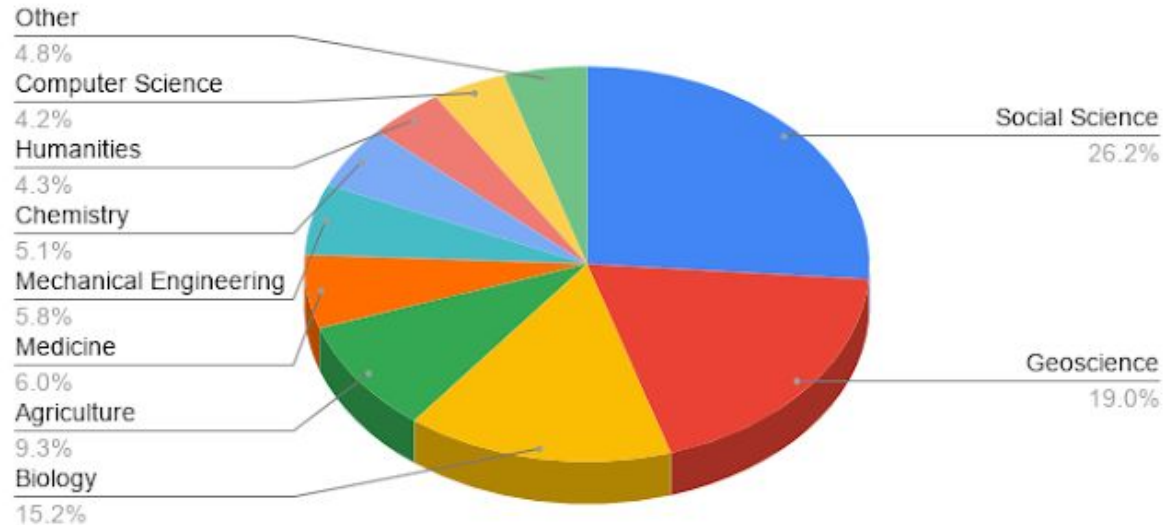
Google Dataset Search - An analysis of online datasets

❖ What do users access?

Reference: https://ai.googleblog.com/2020/08/an-analysis-of-online-datasets-using.html
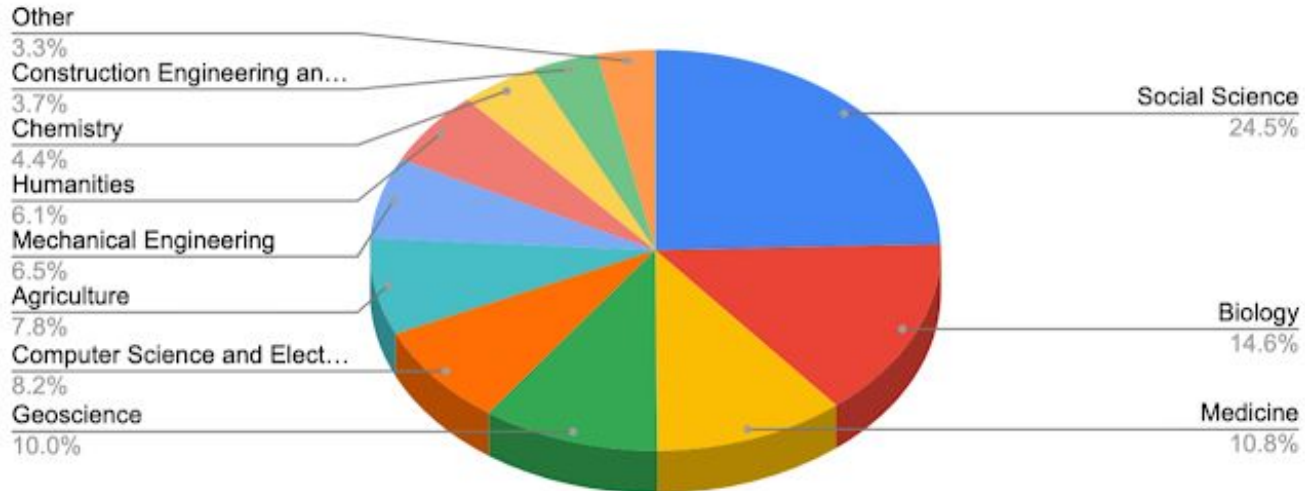
Google Dataset Search

❖ T. Alrashed, D. Paparas , O. Benjelloun, Y. Sheng, and N. Noy: **Dataset or Not? A Study on the Veracity of Semantic Markup for Dataset Pages** (October 2021) https://research.google/pubs/pub50547/

➢ **Schema.org** has become **prevalent** on the Web as a way to express the semantics of Web page content

■ it is present on more than **30%** of Web pages

➢ We cannot always take **Schema.org/Dataset** markup at face value

■ pages may include this markup erroneously or for the purposes of search-engine optimization

Let us pose our original query to Dataset Search

Google   🔍 Is there a land cover dataset produced by the Eur ✕   ⓘ   🗨   **Sign in**

Saved datasets

Your search - **Is there a land cover dataset produced by the European Environment Agency covering the area of Chania, Crete, Greece?** - did not match any datasets.
Suggestions:

- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

Learn how you can add new datasets to our index.

EO dataset search - Copernicus Open Access Hub

➢ Provides complete, free and open access to Sentinel-1, Sentinel-2, Sentinel-3 and Sentinel-5P user products (https://scihub.copernicus.eu)

EO dataset search - Copernicus Open Access Hub

➢ Maintains an online archive of at least the latest month of products for immediate download via HTTP

➢ Synchronous access to latest data, asynchronous access to historic data

EO dataset search - EOWEB® GeoPortal

➢ A multi-mission web portal for interactive access to the German Aerospace Center (DLR) Earth observation data holdings

   ○ Combines classic discovery and order services for data held in the German Satellite Data Archive (D-SDA) with browse and download features via interoperable, OGC-compliant visualization and download services (https://eoweb.dlr.de/egp/)

EO dataset search - EOWEB® GeoPortal

EO dataset search - Earthdata Search

➢ Earthdata Search enables data discovery and access to more than 33,000 EO data collections from NASA's Earth Observation System Data and Information System (EOSDIS), U.S. and international partner agencies (search.earthdata.nasa.gov/)

- Users can:
  - **Search for EO data:** Earthdata Search uses the Common Metadata Repository (CMR) for sub-second search
  - **Preview EO data:** Using Global Imagery Browse Services (GIBS), enables high-performance data visualization
  - **Download and access EO data:** In addition to direct download, surfaces OPeNDAP services for simpler spatial and parameter subsetting

EO dataset search - Earthdata Search

➢ Many of the imagery layers are updated daily & are available within 3 hours of the observation

Copernicus Data and DIAS

Data and Information Access Services (DIAS):

- Five competitive platforms for quick access to a huge resource of Earth Observation data (satellite imagery) and Copernicus themed services.

- Easy and quick search, processing and sharing of satellite data.

- Quick access to satellite imagery via virtual machines.

- Data is free but computing power on a DIAS platform needs to be paid for.

Copernicus Data and DIAS

THE DIAS & WHERE TO REACH THEM

WWW.CREODIAS.EU

WWW.SOBLOO.EU

WWW.MUNDIWEBSERVICES.COM

WWW.ONDA-DIAS.EU

WWW.WEKEO.EU
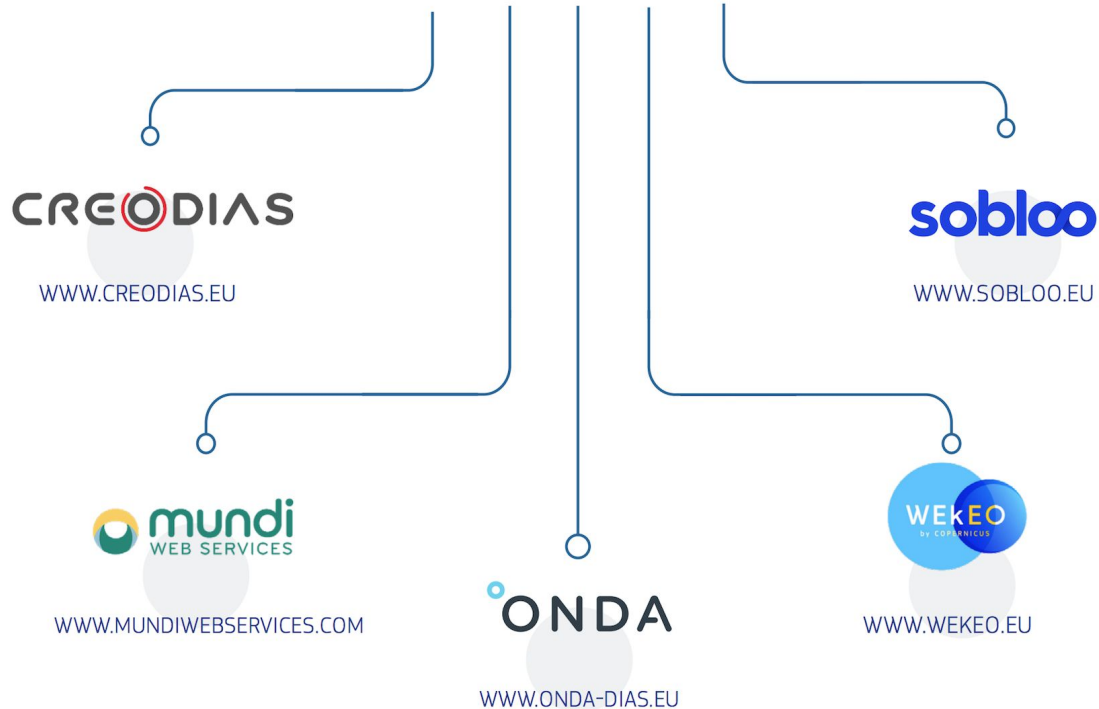
CREODIAS - Uses linked data in its catalogue!

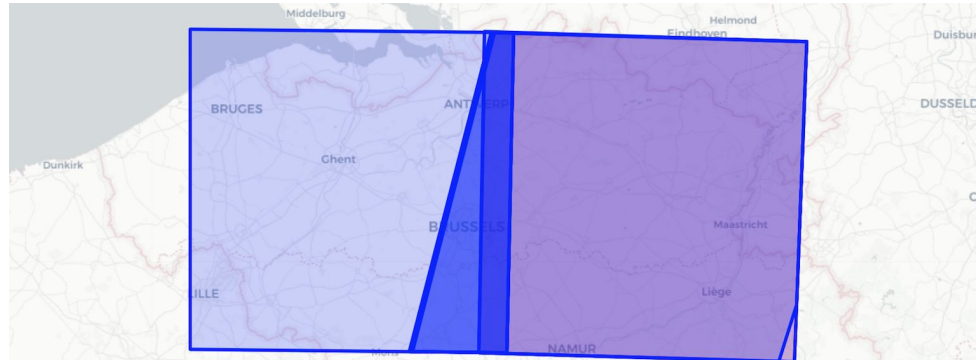**SPARQL interface:** provides extended search capabilities for linking metadata of all products stored in the repository with various information from the Internet

Example Query: Find all Sentinel-2 images in the area of Brussels

SPARQL Query:

```
SELECT DISTINCT ?title ?geom {
?feature <http://ws.eodias.eu/metadata/attribute#title>  ?title.
?feature <http://ws.eodias.eu/metadata/attribute#mission>

<http://ws.eodias.eu/metadata/mission/ Sentinel-2> .
    ?hex        <http://ws.eodias.eu/metadata/attribute#feature> ?feature.
    ?hex        <http://ws.creodias.eu/metadata/object/settlement >
                            <http://dbpedia.org/resource/ Brussels> .
    ?feature <http://ws.eodias.eu/metadata/attribute#geometry>  ?geom .

} LIMIT 100
```

CREODIAS

Current work

- We are developing an **extension to schema.org for Earth Observation data (eo.schema.org).**
- Main idea:
  - Use OGC 17-003: EO Dataset Metadata GeoJSON(-LD) Encoding Standard
- We are also developing an **annotation tool** that can be used to annotation Earth Observation datasets using this extension.