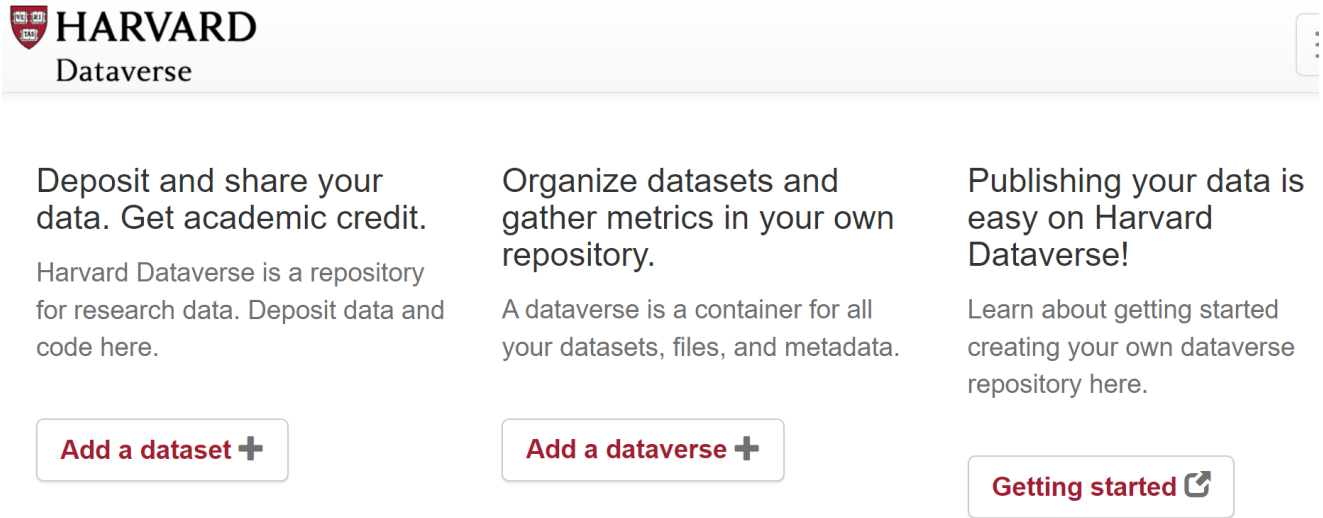
> Repositories

General Repositories

General repositories are platforms that contain/store hundreds/thousands of datasets related to different topics, and can be easily search through dedicated queries. Two families:

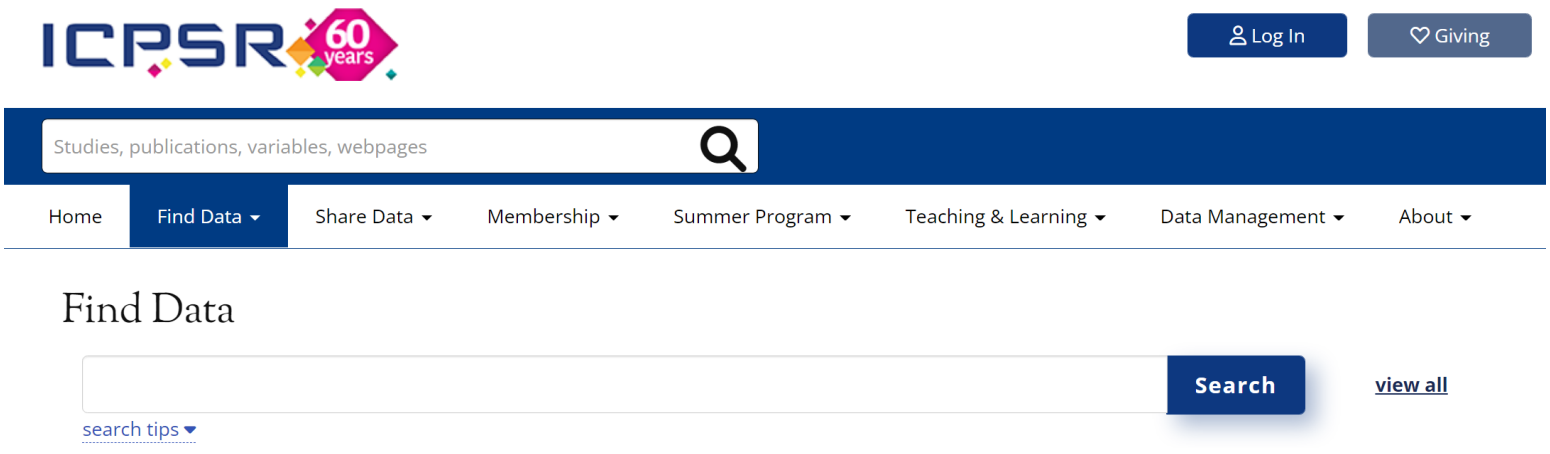
- Some repositories directly store data on a given linked website
- Other repositories simply retrieve data based on a search and provide url link to access the dataset you are looking for

Harvard Dataverse



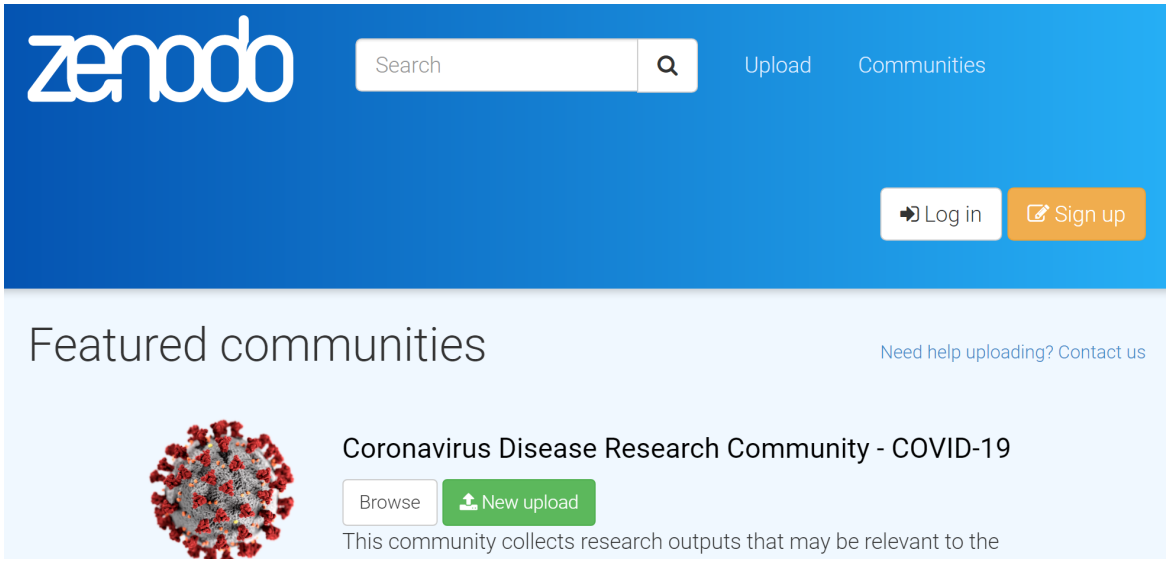
- Harvard Dataverse (https://dataverse.harvard.edu/) is a general purpose platform that enables students/scholars to upload and retrieve data across many fields
- Especially popular for depositing replication data
- Contains data in various formats, across various fields (not only social sciences)

- The Inter-University Consortium for Political and Social Research platform (https://www.icpsr.umich.edu/web/pages/) is a hugely diffused platform specifically focusing on social and political sciences
- Allows various filtering procedures (also in terms of geography)
- Requires login (for free)



Zenodo

- Zenodo is a platform encouraging open science/reproducibility across fields
- Potentially less diffused in social sciences, overall very popular in last years
- Do not contain datasets only, plus some resources are restricted or closed
- Do not require login for downloading



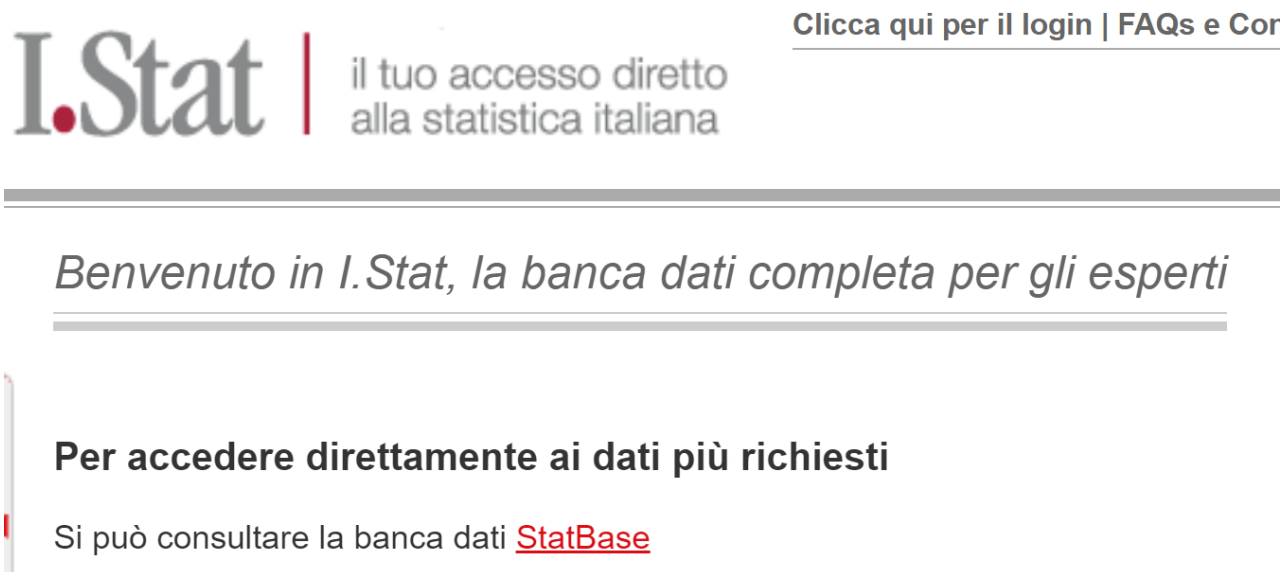
Google Dataset Search

- Google Dataset Search is an engine developed around 5 years ago by Google to facilitate data gathering. Link: https://datasetsearch.research.google.com/
- Works like a traditional Google engine: you run your query/search, it shows you results that are present on the web (not direct dataset storage)
- You can filter/order by date, format, license and user rights
- Risk (so beware): showing results that are not particularly reliable due to the source (don't assume that if it is found by Google it is a trustworthy dataset!)

Institutional Repositories

ISTAT

- The Italian National Institute of Statistics (ISTAT) has a portal containing datasets on various topics (http://dati.istat.it/)
- Topics include: demographics, job data, socio-economic indicators, justice data, public health
- My review: *tremendously* unfriendly website
- Often data are not disaggregated/are overly aggregated and not updated



Homologous Portals for other countries (selected)

- **Germany:** National Data Portal (https://www.govdata.de/)
- **France:** Open Platform for French Public Data (https://www.data.gouv.fr/en/)
- **United Kingdom:** Find Open Data (https://data.gov.uk/)
- **Netherlands:** Open Data - Centraal Bureau voor de Statistiek (https://www.cbs.nl/en-gb/onze-diensten/open-data)
- **Spain:** Open Data initiative of the Government of Spain (https://datos.gob.es/en)
- **Mexico:** Datos Abiertos de Mexico (https://datos.gob.mx/)
- **Canada:** Open Data, Open Government (https://open.canada.ca/en/open-data)
- **Australia:** Open Data (https://data.gov.au/)

Eurostat and World Bank

Two other important resources are the World Bank Data Catalog and the Eurostat Database:

- **World Bank Data Catalog**
  - Link: https://datacatalog.worldbank.org/home. More than 5000 datasets
  - Contains data on a variety of economic-related topics, such as development, production, natural resources,
- **Eurostat Database**
  - Official data repository of the European Union. Link: https://ec.europa.eu/eurostat/data/database
  - Themes include regional stats, economics and finance, population, industry, international trade, agriculture, environment, science
  - Depending on dataset, various spatial/temporal disaggregation schemes exist

Other Institutional Sources

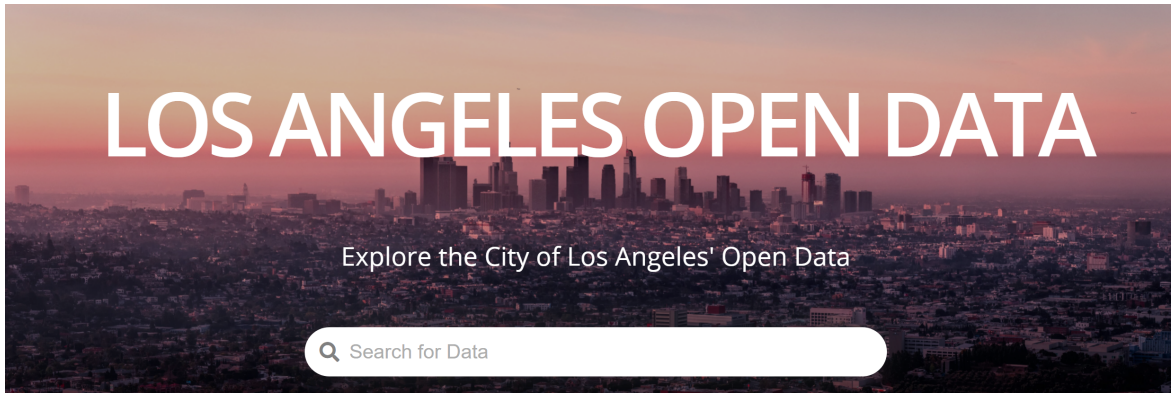
- **World Health Organization** https://www.who.int/data/collections
- **Unicef Data** https://data.unicef.org/open-data/
- **United Nations Data** https://data.un.org/
- **OECD Data** https://data.oecd.org/

US Cities Open Data Portals

Many US cities have rich open data portals where you can easily access data on a number of topics. Among others:

- *Crime*
- *Demographics*
- *Urban development*
- *Economic indicators*

Depending on the city, you can either export data via traditional formats (e.g., csv) or connect to dedicated API that can be integrated in your programming suite, like R (see for instance https://www.chicago.gov/city/en/narr/foia/sample_code0.html for the city of Chicago)



Some Examples of Open Data Portals

- Los Angeles (https://data.lacity.org/)
- Chicago (https://data.cityofchicago.org/)
- Philadelphia (https://www.opendataphilly.org/)
- New York (https://opendata.cityofnewyork.us/)
- Seattle (https://data.seattle.gov/)

And also less "famous" cities:

- Tuscaloosa, AL (https://data.tuscaloosa.com/)
- Wichita, KS (https://openwichita.org/data)
- Tucson, AZ (https://gisdata.tucsonaz.gov/)
- Omaha, NE (https://data-dogis.opendata.arcgis.com/)
- ...

Specific Databases/Datasets

Global Terrorism Database

- The Global Terrorism Database (GTD) is the richest open access dataset on terrorism events occurred worldwide from 1970 to 2019 (200,000+ events)
- Access is free for students/researchers at: https://www.start.umd.edu/gtd/access/
- Codebook available: https://www.start.umd.edu/gtd/downloads/Codebook.pdf
- Event-based, dozens of variables associated with each attack:
  - *Perpetrator (up to three terrorist actors)*
  - *Time (daily level) and Location (disaggregated per different geographical scale + LAT/LON)*
  - *Weapons used, tactics employed, targets hit (more than one column per each)*
  - *Connected events*
  - etc...

ACLED

- The Armed Conflict Location & Event Data project (ACLED) collects data on political violence, riots, protests around the world
- Access is free pending (free) registration and token acquisition at https://acleddata.com/data-export-tool/
- Includes special collections on data dedicated to specific topics such as:
  - *COVID-19 disorders*
  - *Violence specifically targeting women*
  - *Violence against civilians*
- Event-based, less variables than GTD but still pretty informative

Network Repositories

UCI Network Data Repository

- The University of California, Irvine Network Data Repository (https://networkdata.ics.uci.edu/) contains datasets in network format to encourage/facilitate the study of networks.
- Divided into various subfamilies:
  - *Collaboration nets*
  - *Ecology nets*
  - *Communication nets*
  - *Friendship nets*
  - etc.

UCIrvine | University of California

Network Data Repository

Stanford Large Network Dataset Collection

- Maintained by the SNAP group at Stanford, this dataset collection (https://snap.stanford.edu/data/) includes network datasets across various domains
- Generally: large datasets (also in range of millions nodes/edges), so make sure your machine has enough computational power to sustain them
- Categories:
  - *Social networks (including social media ones from Facebook and Twitter)*
  - *Citation/Collaboration networks (to study scientific collaboration patterns)*
  - *Wikipedia network (to study knowledge production patterns)*
  - *Product co-purchasing (to study consumers behavior)*
  - etc...

UCINET Repo

- UCINET (one of the most common - and outdated - software for social network analysis) has a list of datasets that are very common in the literature (https://sites.google.com/site/ucinetsoftware/datasets?authuser=0)
- They cover various disciplines/topics and are mostly available in UCINET and .csv formats
- Particular focus on covert networks (https://sites.google.com/site/ucinetsoftware/datasets/covert-networks?authuser=0), like:
  - *Montreal Street Gangs*
  - *Big Allied and Dangerous Data (on alliances among terror groups)*
  - *Al Qaeda Network 1993-2003*
  - *Bali Bombings network (2005)*