Как и где искать открытые данные?

Ирина Радченко
кандидат технических наук, доцент
iradche@gmail.com

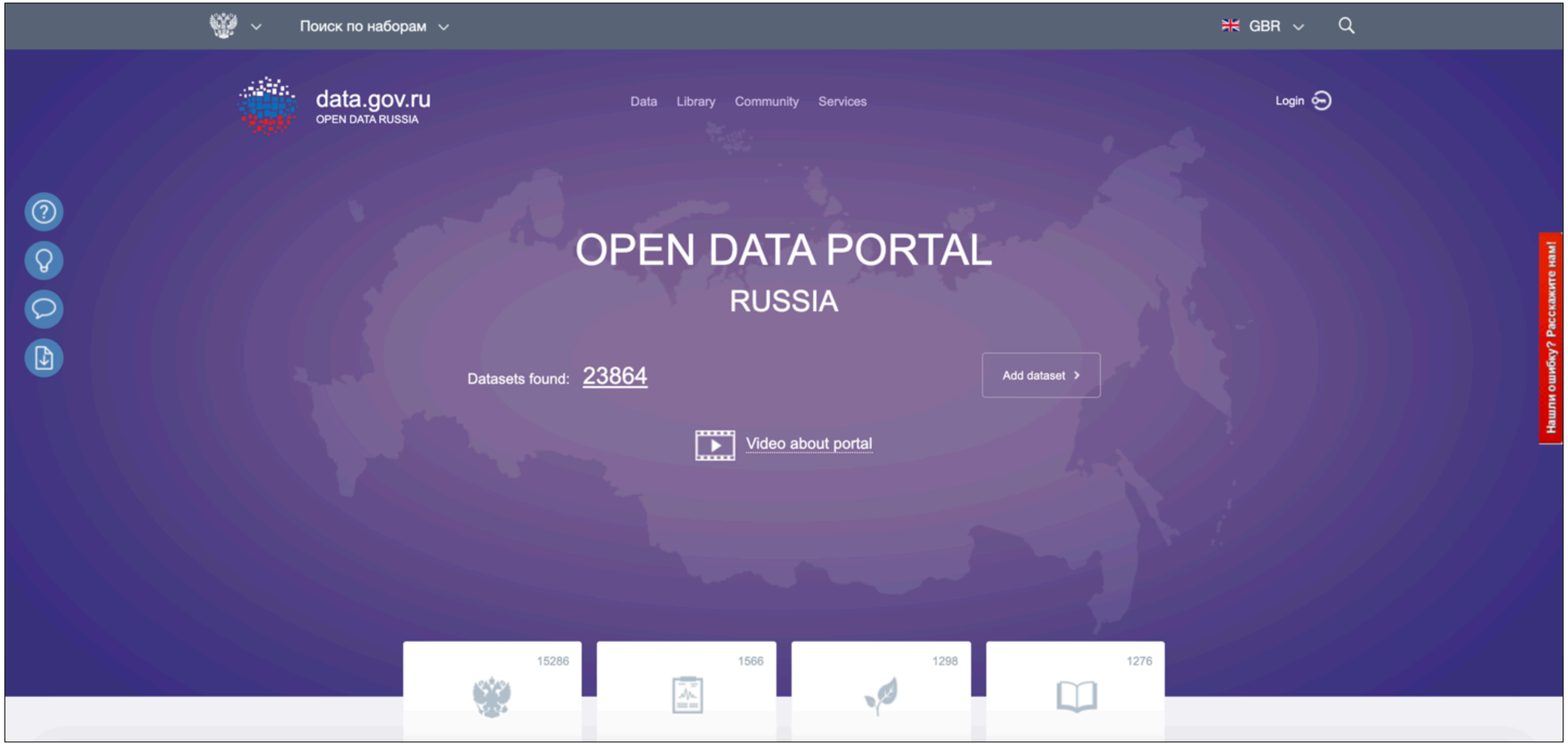
мастер-класс для Skills Lab
24 сентября 2020

Что такое открытые данные?

Открытые данные:

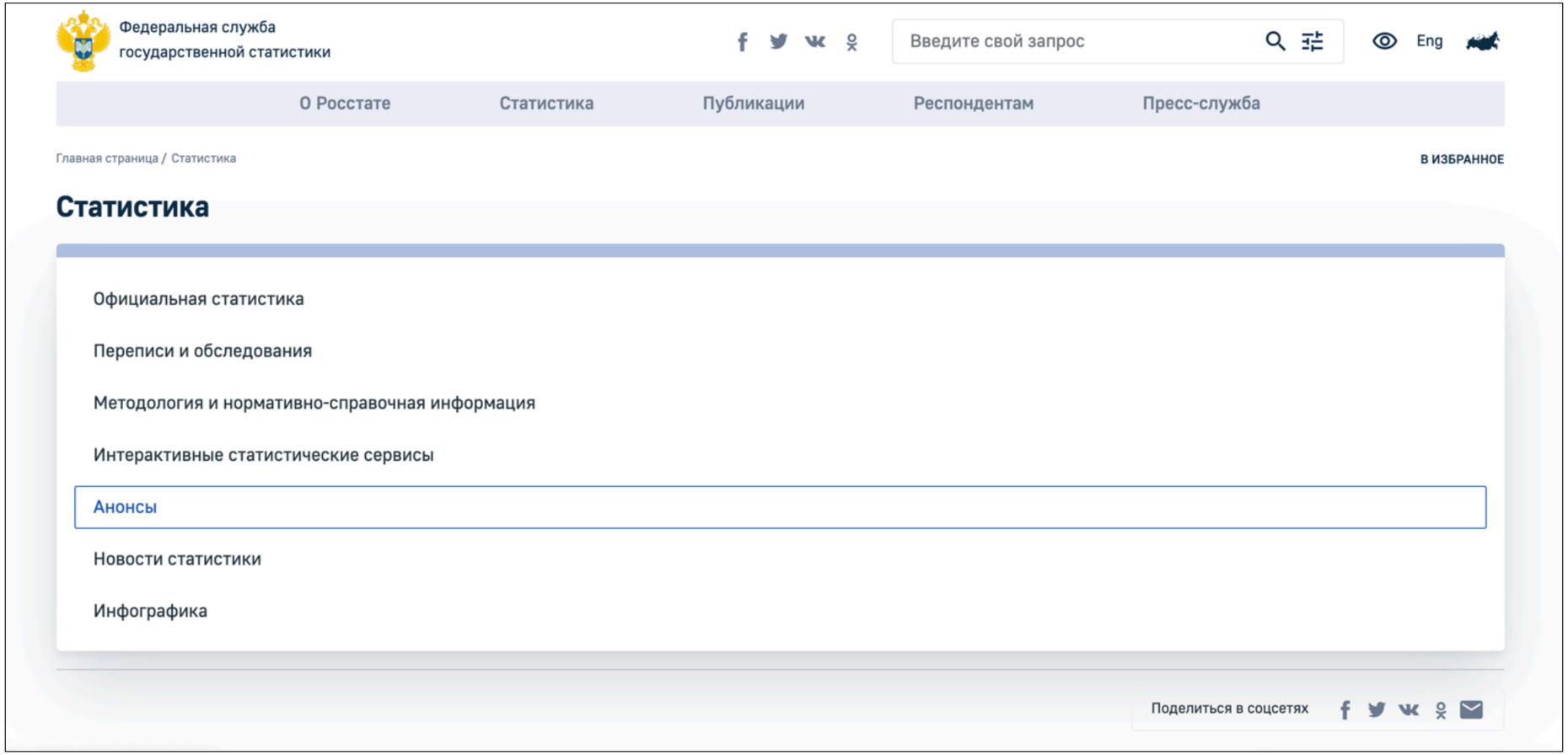
- машиночитаемость
- открытость

Портал российских открытых данных



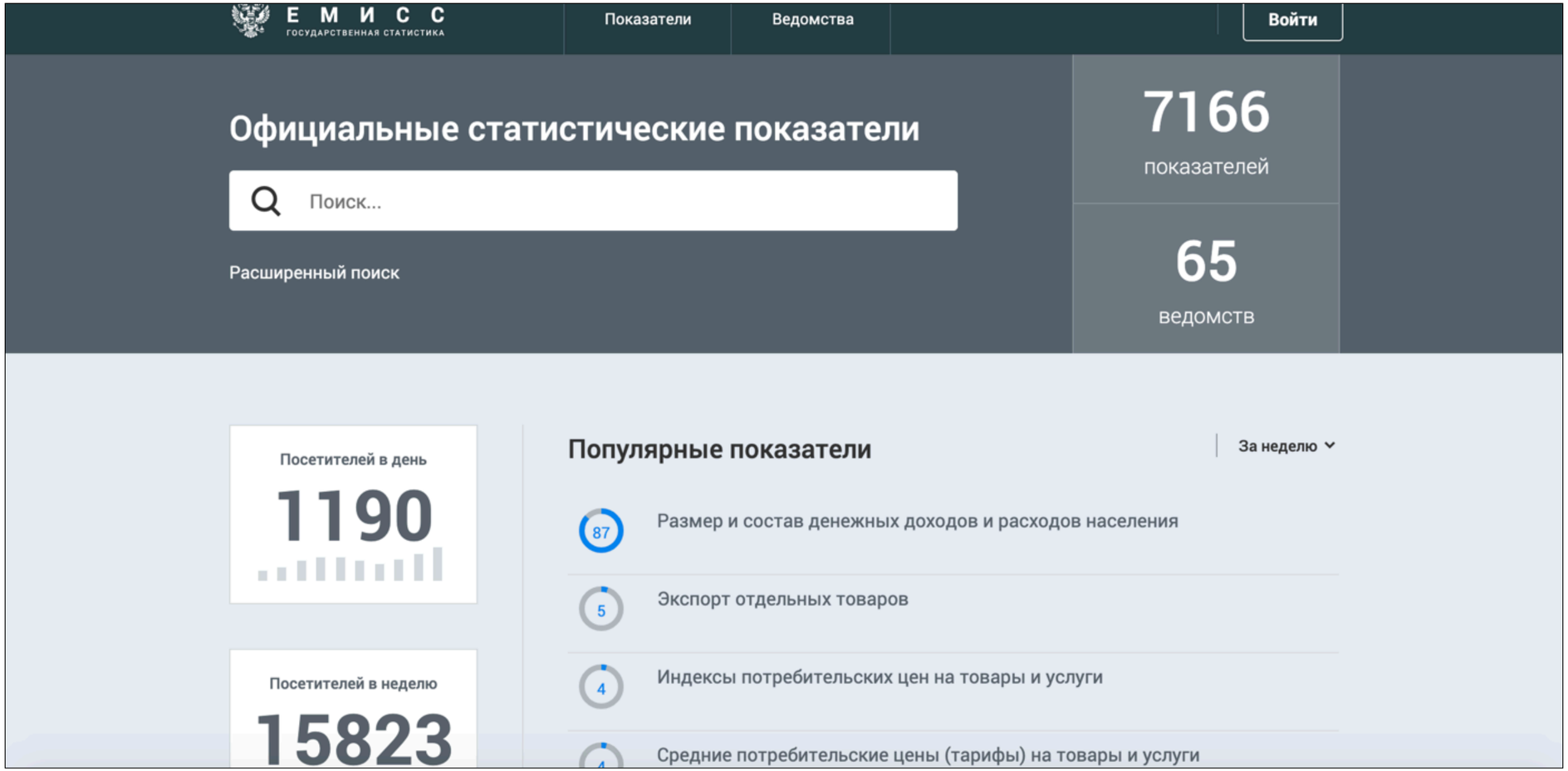
https://data.gov.ru/

Федеральная служба государственной статистики



https://rosstat.gov.ru/statistic

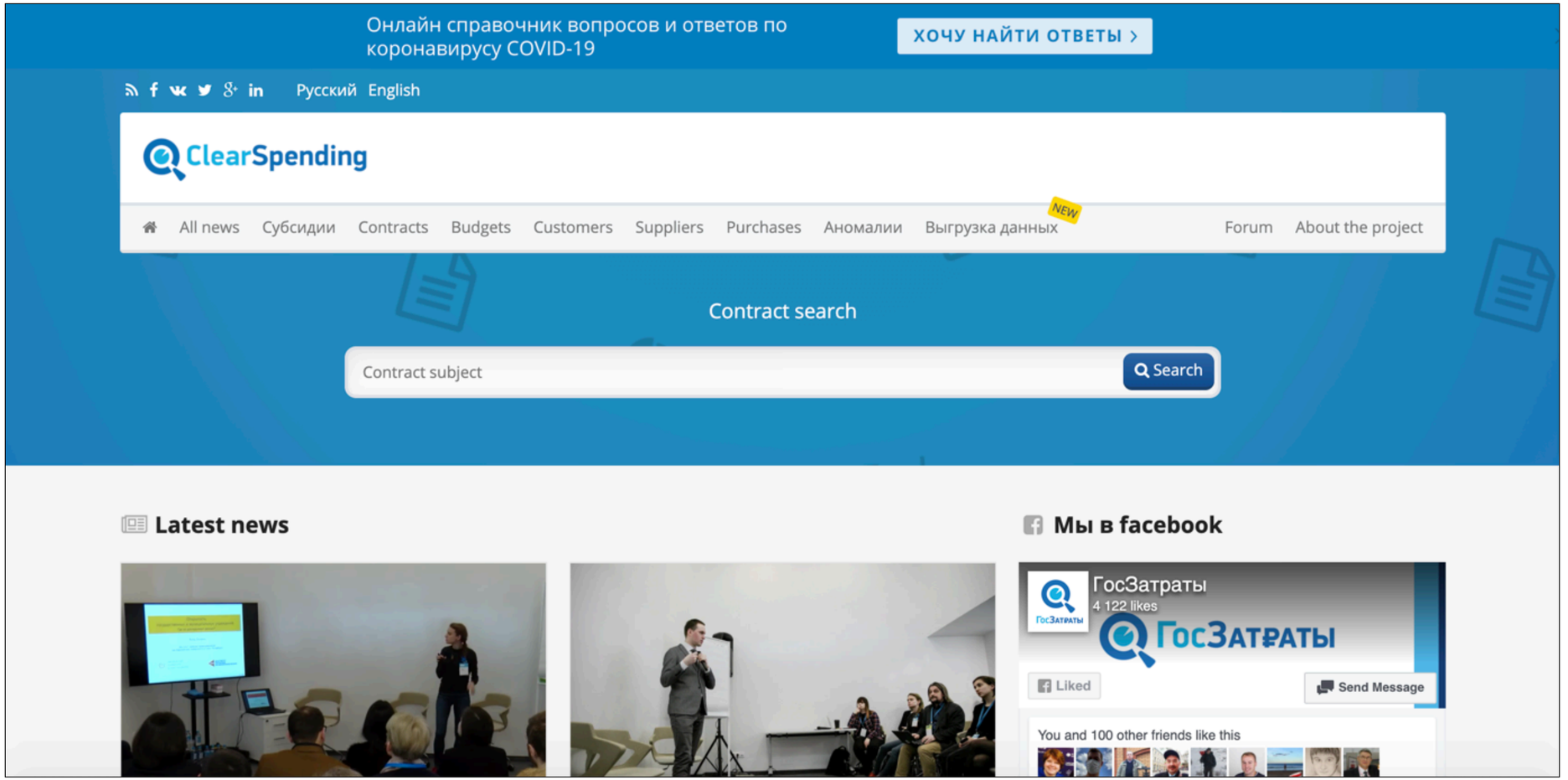
ЕМИСС государственная статистика



https://www.fedstat.ru/

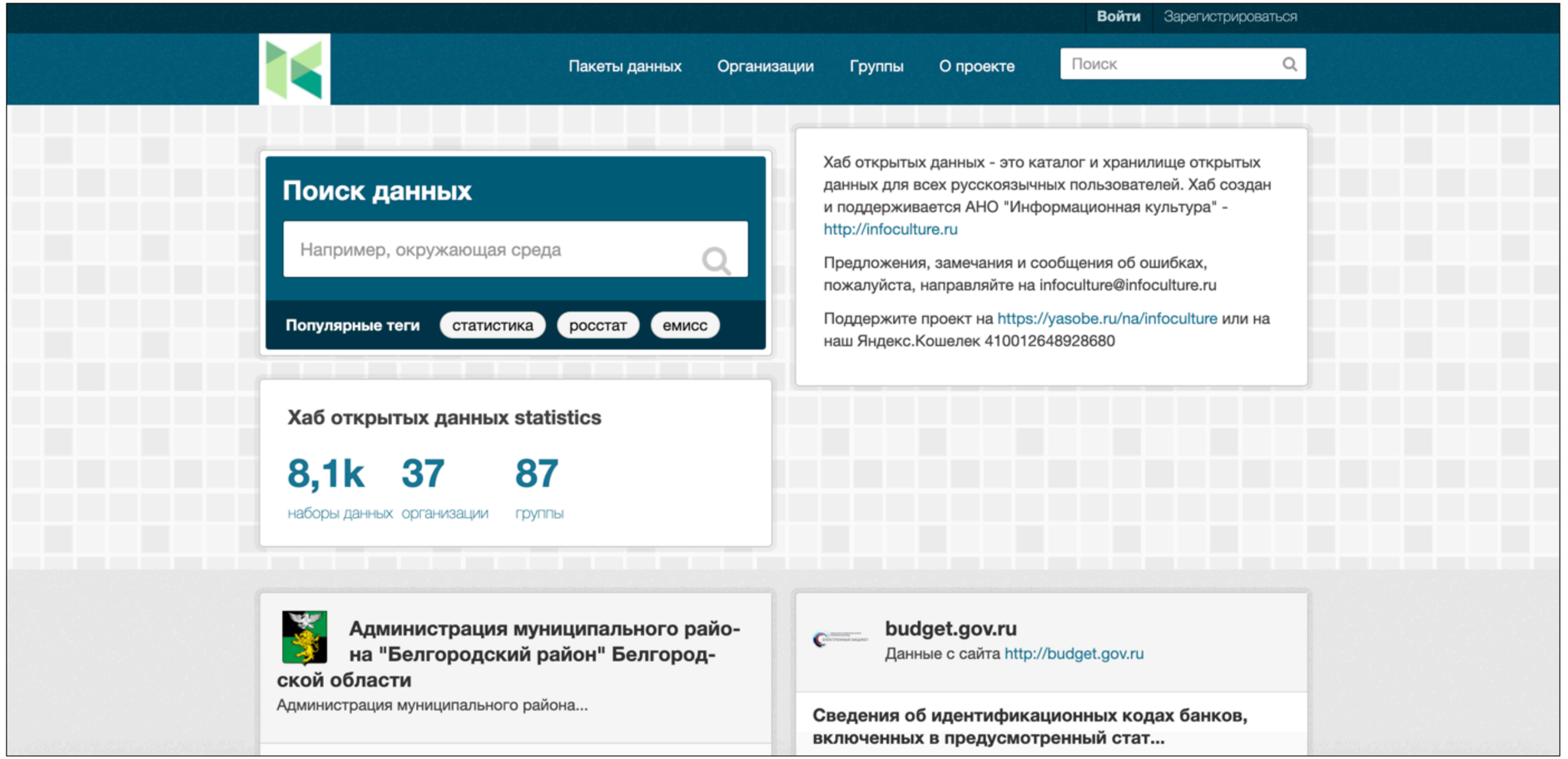
Проекты

Проект ГосЗатрат



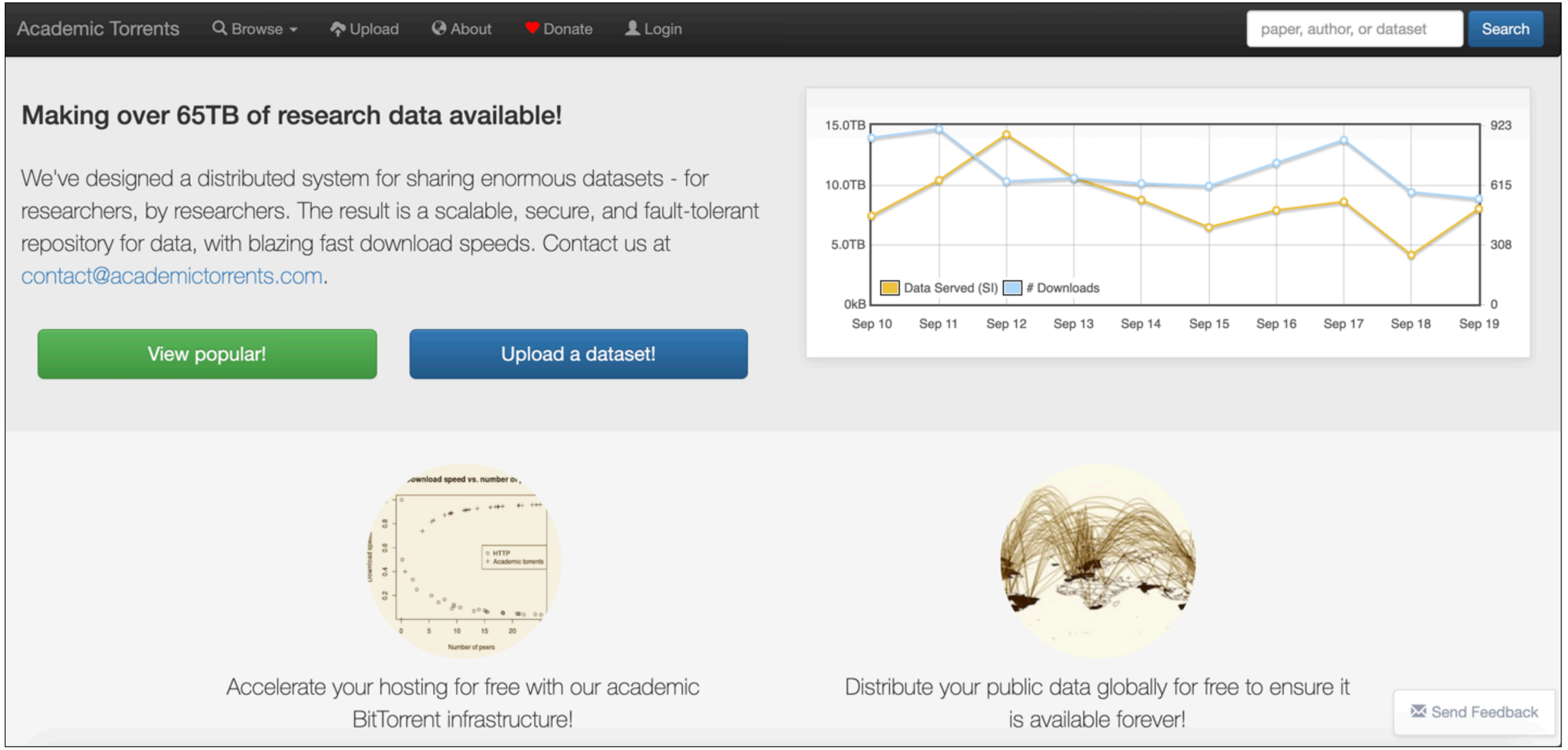
https://clearspending.ru/

Проект Хаб открытых данных



https://opengovdata.ru/

Академический торрент



https://academictorrents.com/

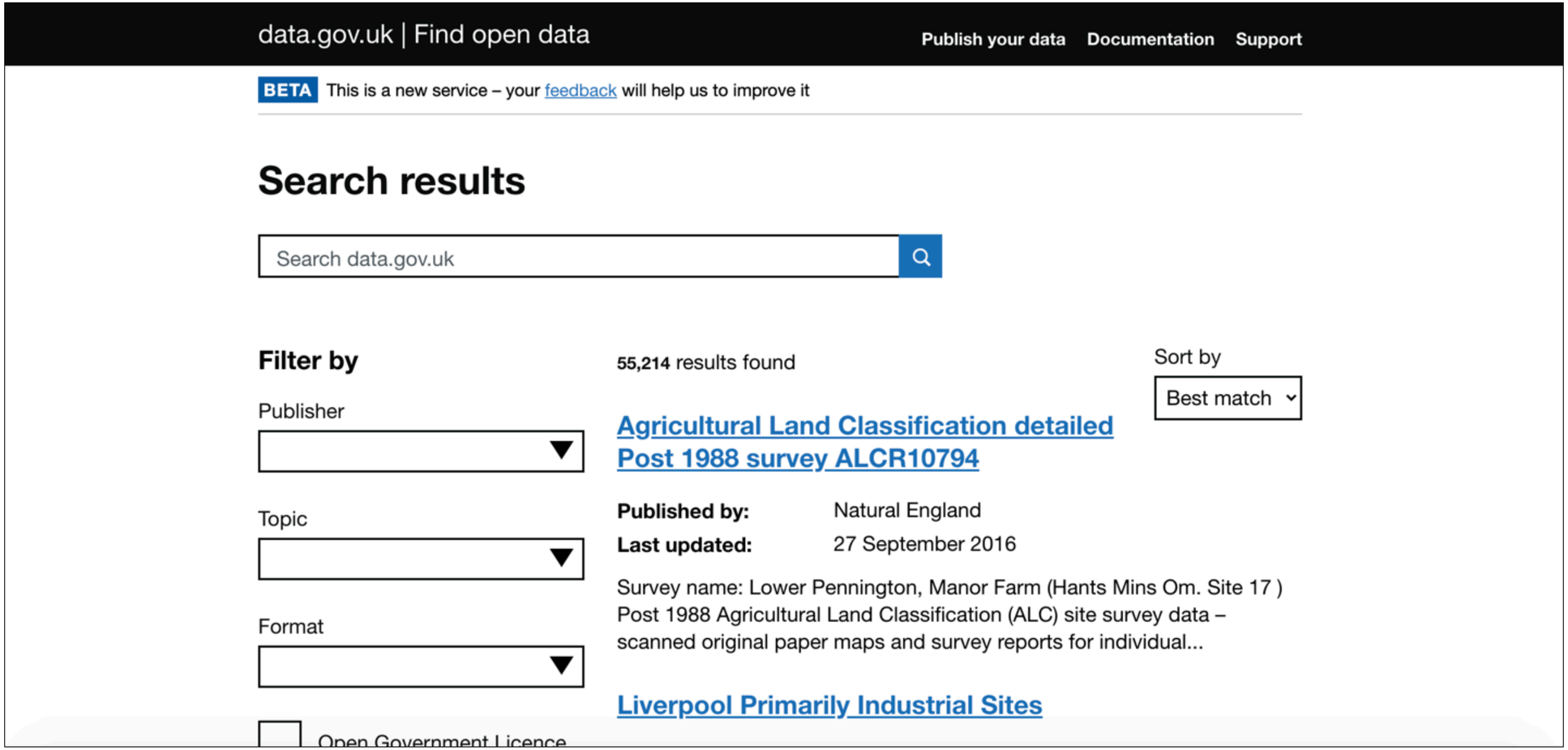
Платформы распространения данных

- Открытые государственные данные на платформах
- Международные организации
- Платформы для дата-ученых
- Транснациональные корпорации

US Government's open data

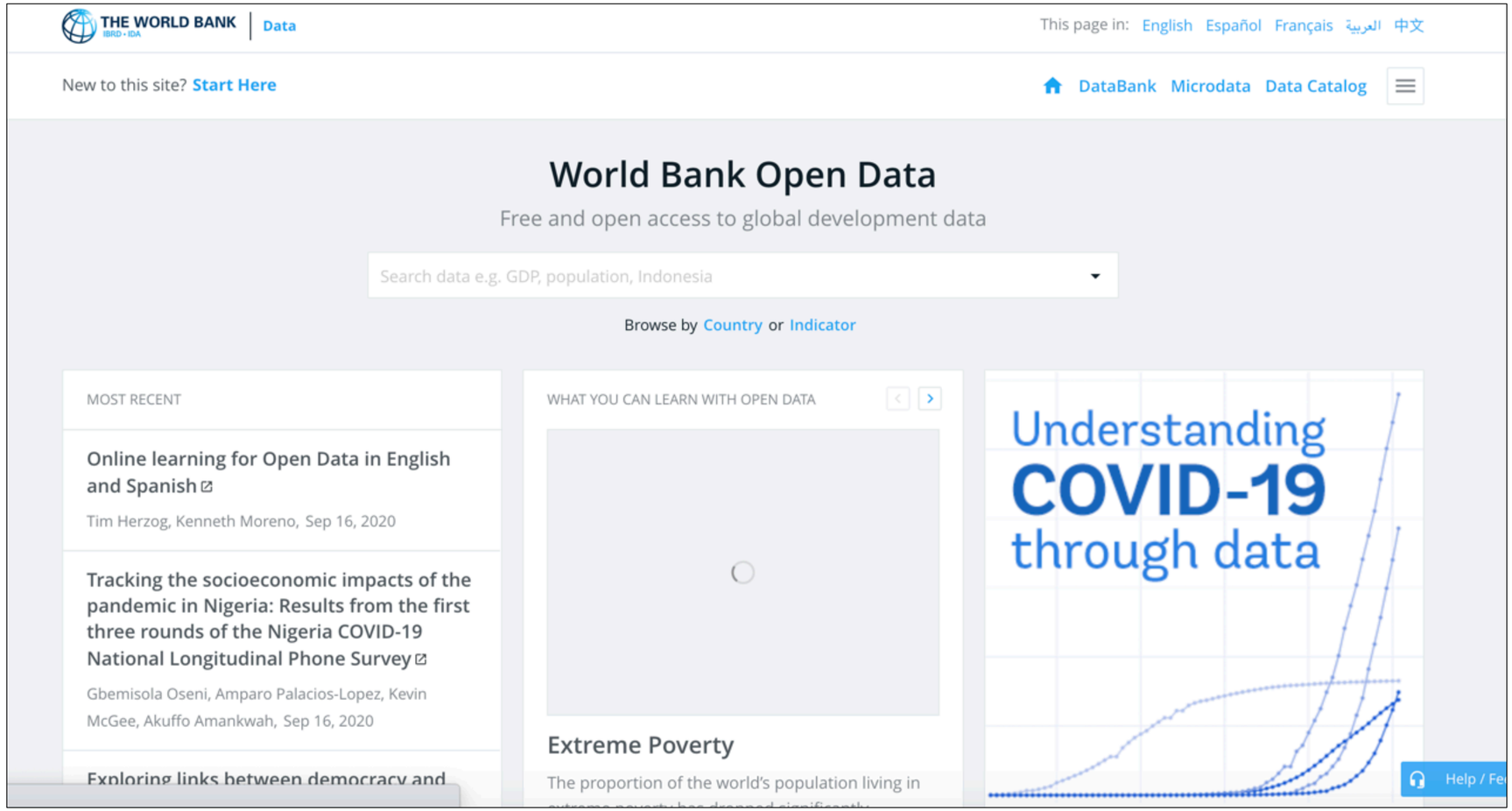
https://www.data.gov/

UK Government's open data



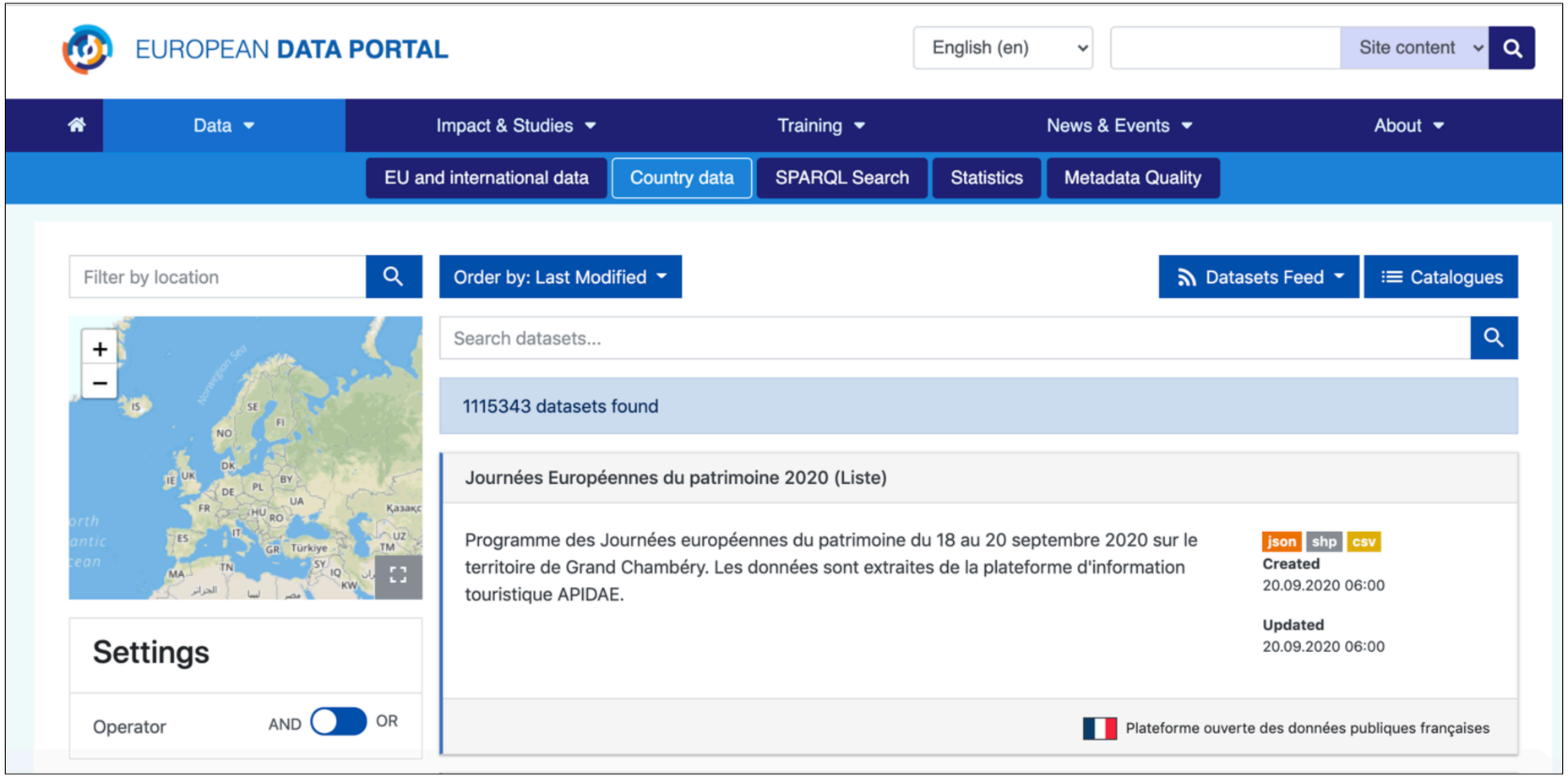
https://data.gov.uk/

World Bank Open Data



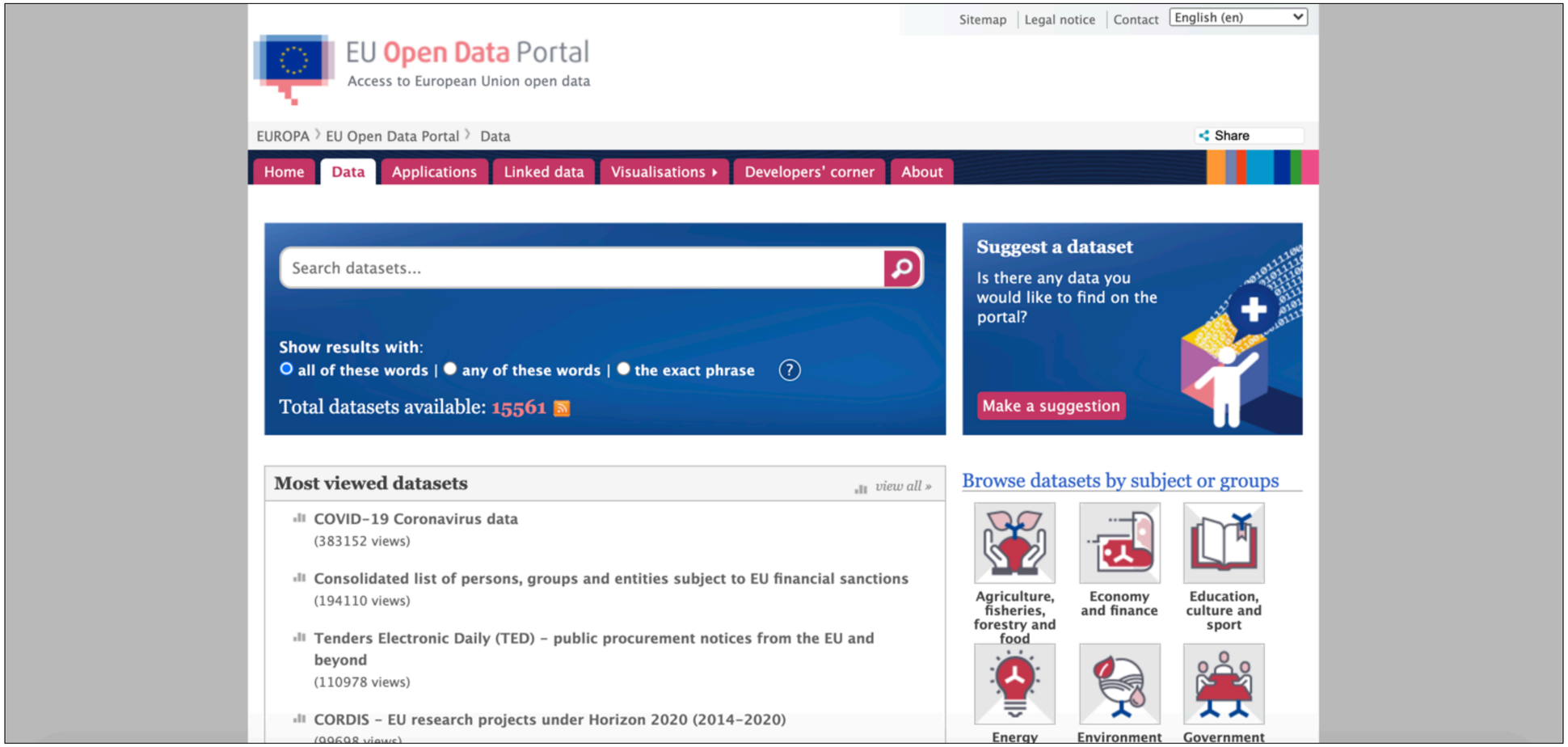
https://data.worldbank.org/

European Data portal: сборщик метаданных



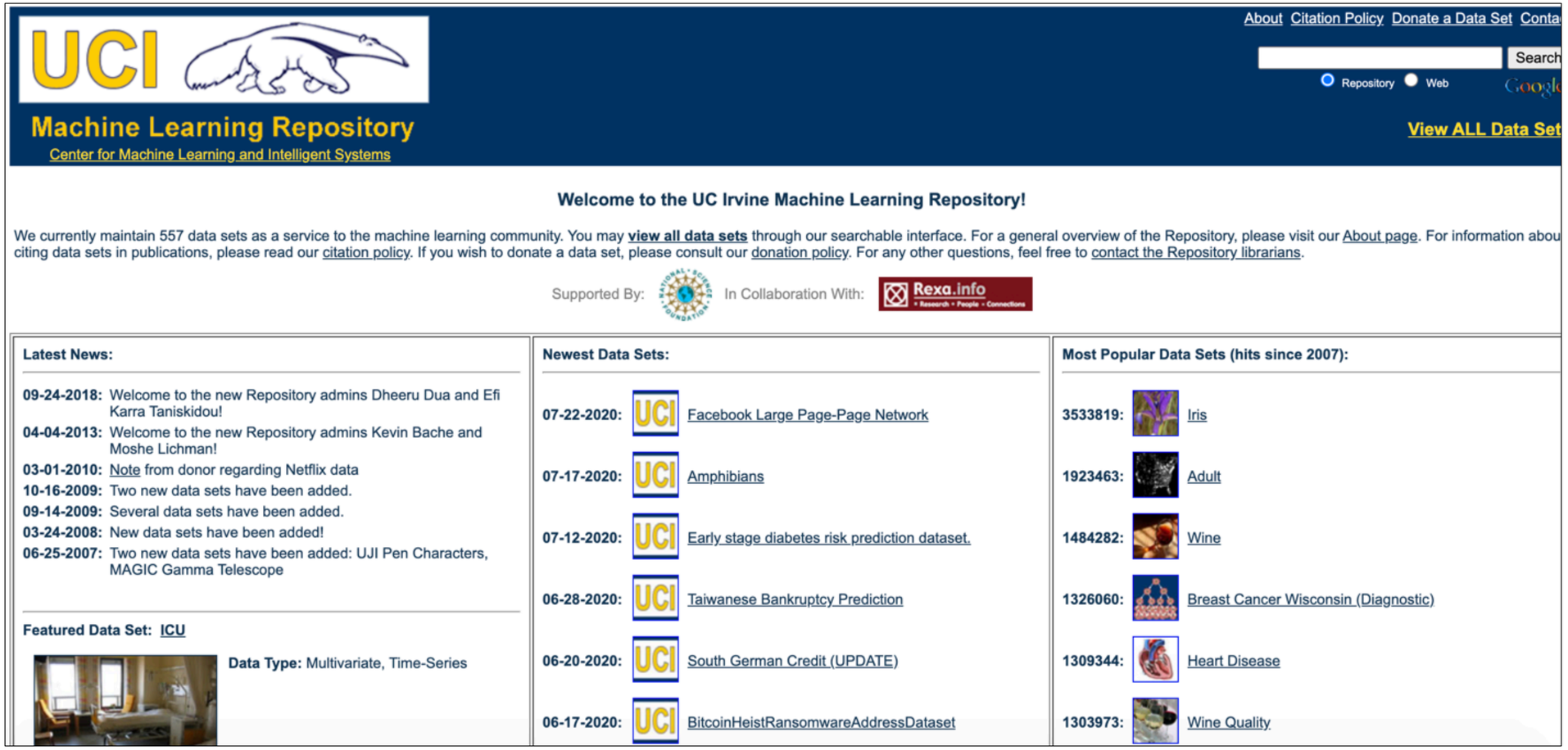
https://www.europeandataportal.eu/data/datasets?locale=en&minScoring=0

EU open data portal



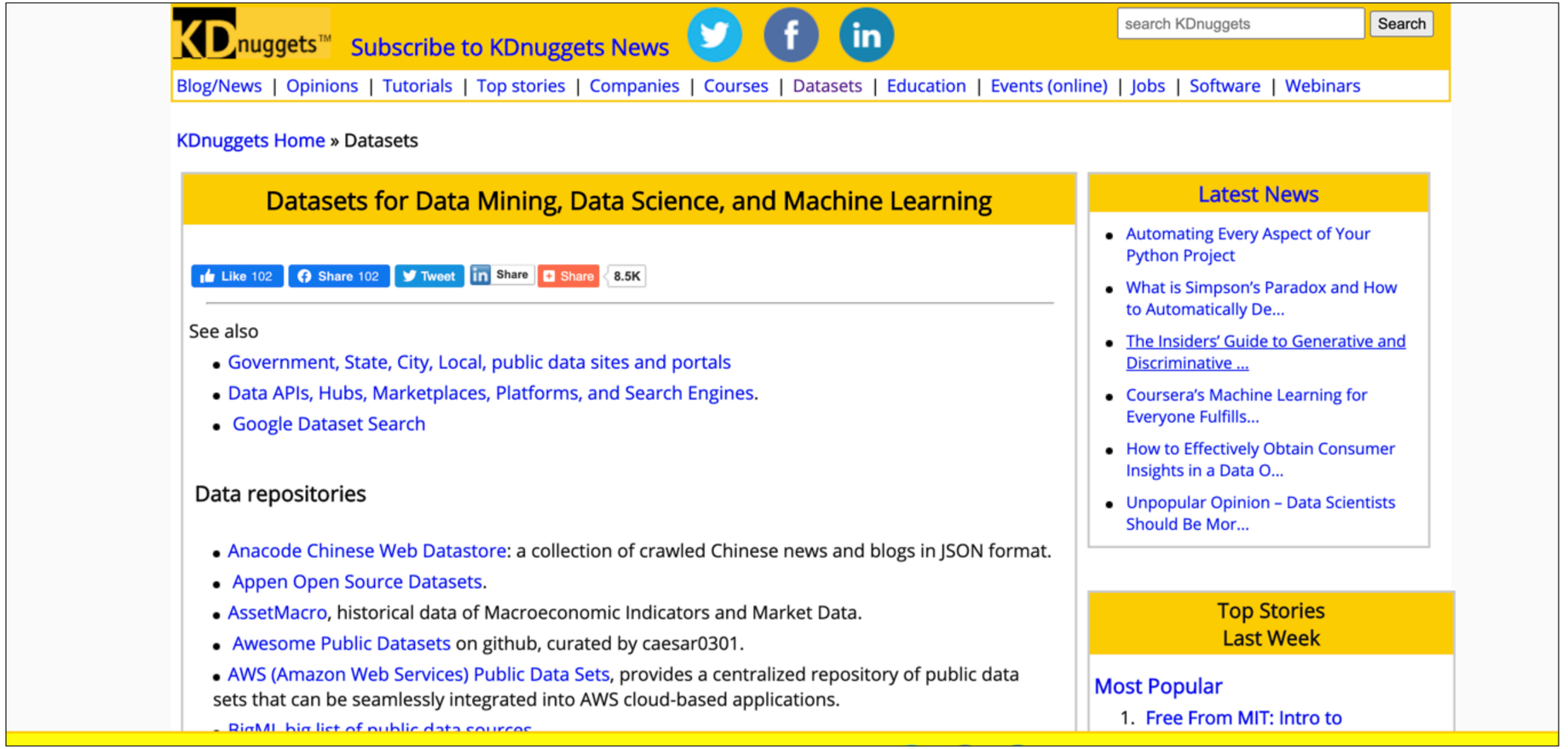
https://data.europa.eu/euodp/en/data

Machine Learning Repository



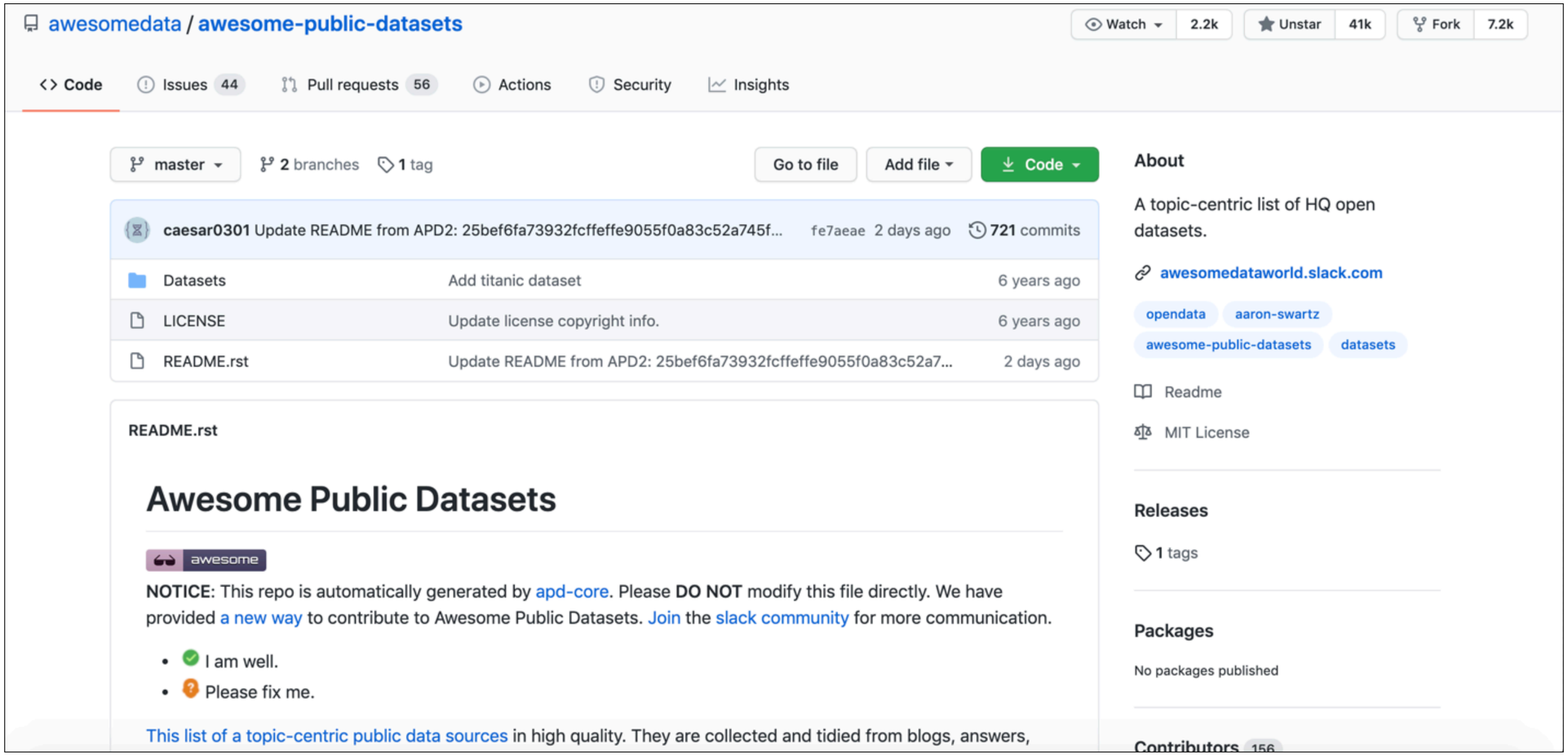
https://archive.ics.uci.edu/ml/index.php

Коллекция KDnuggets



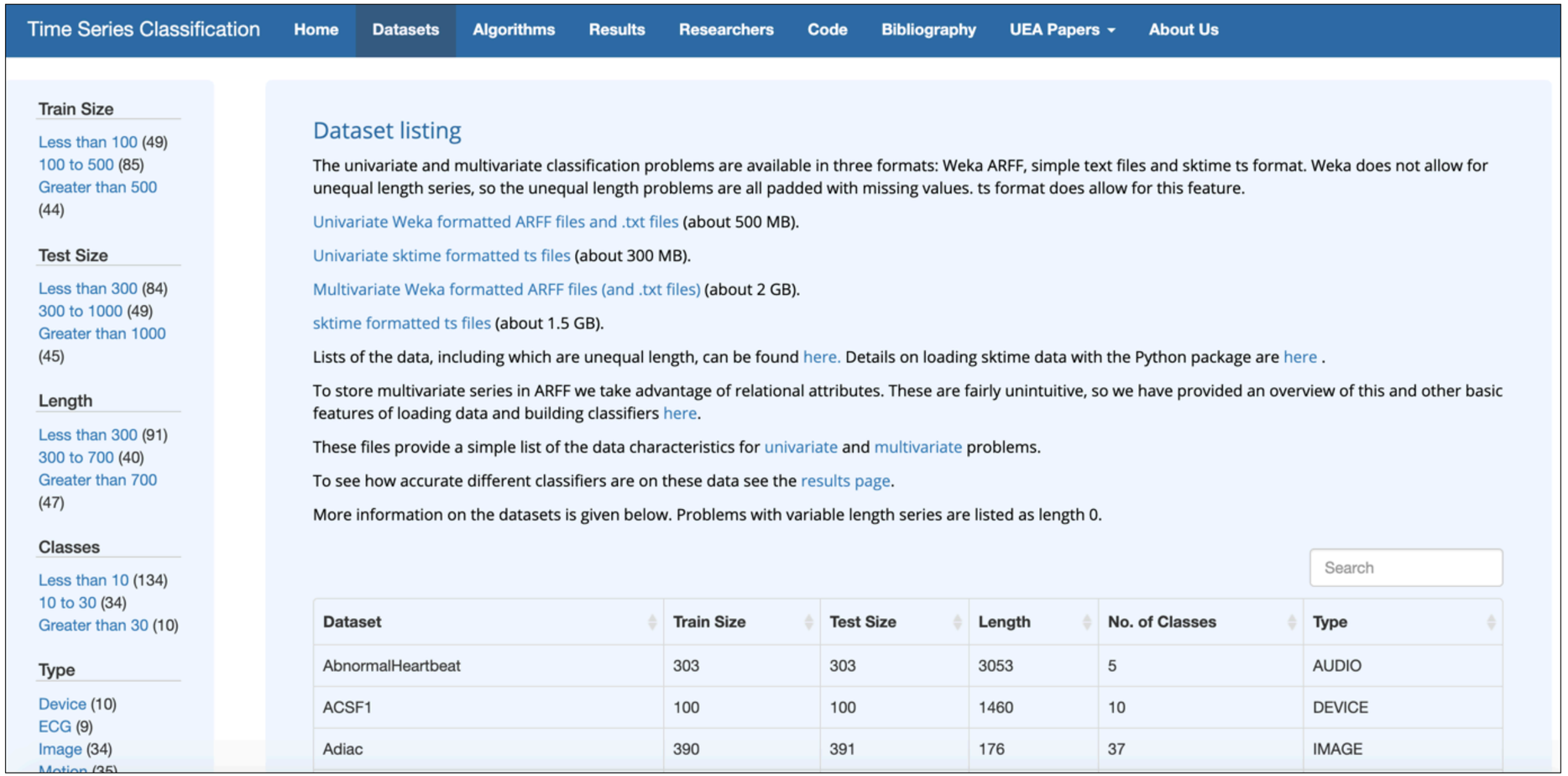
https://www.kdnuggets.com/datasets/index.html

Awesome Public Datasets



https://github.com/awesomedata/awesome-public-datasets

Time Series Classification

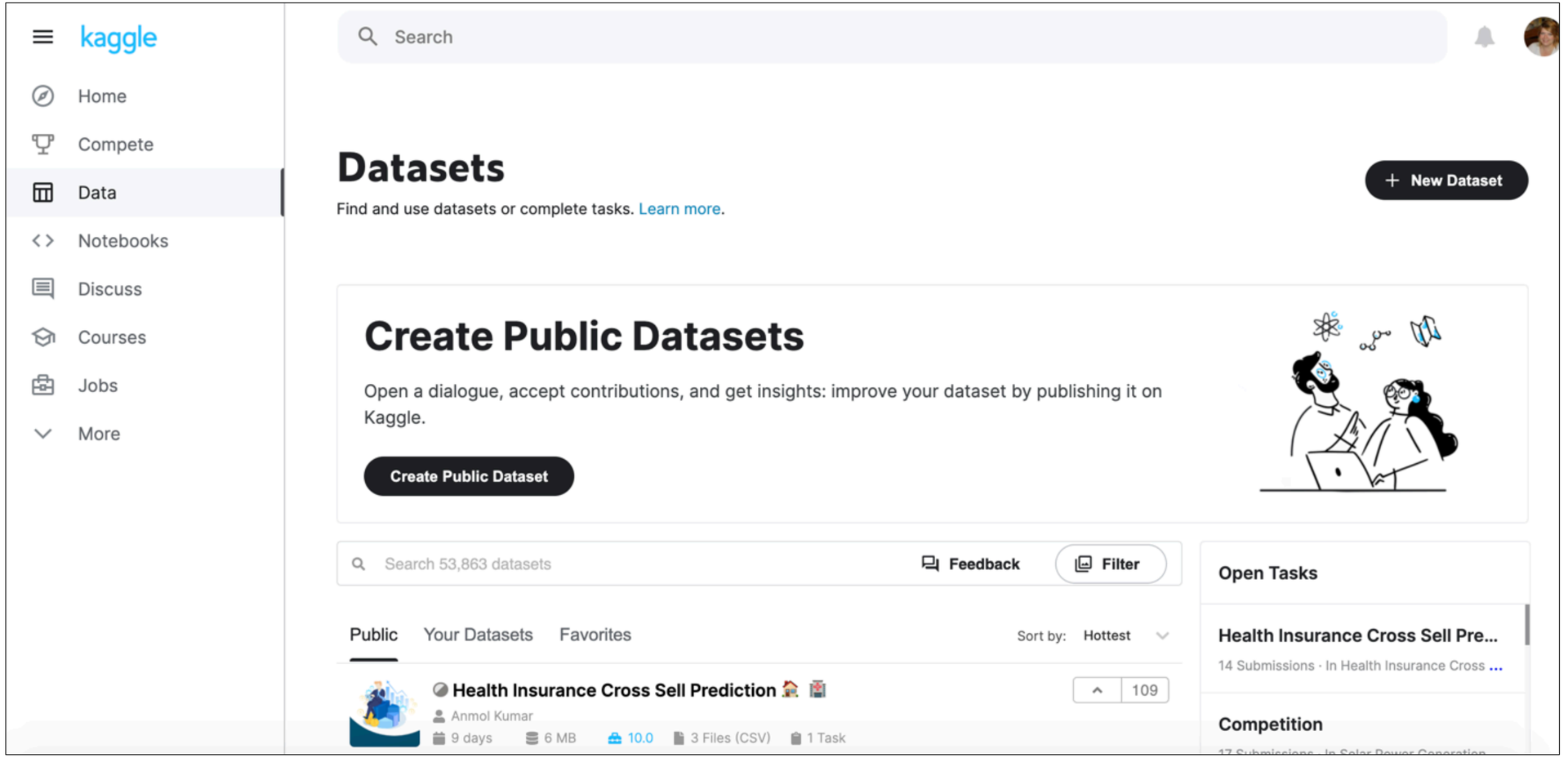


http://timeseriesclassification.com/aboutus.php

Платформы для дата-ученых

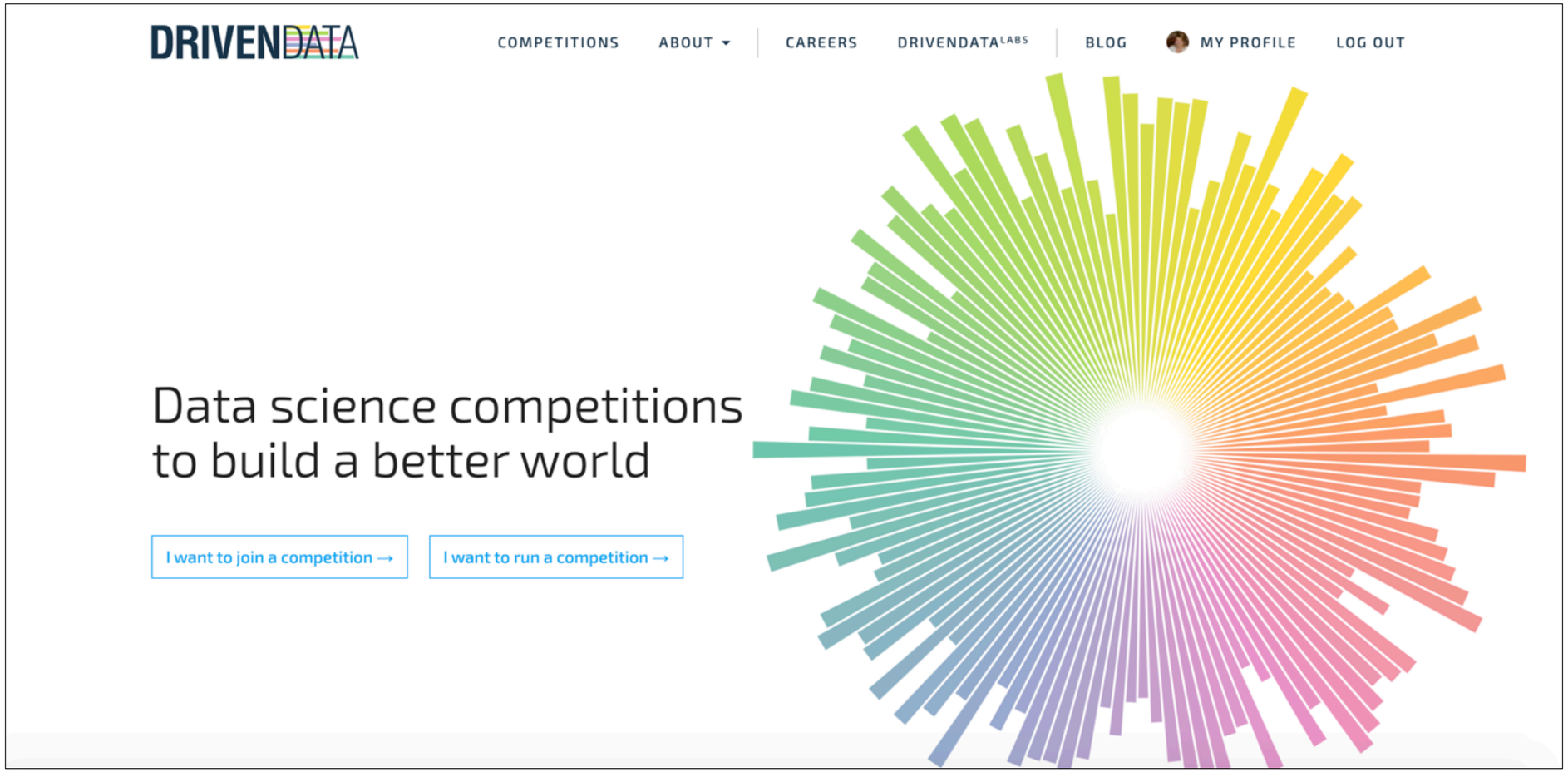
- Kaggle
- DrivenData

Платформа Kaggle



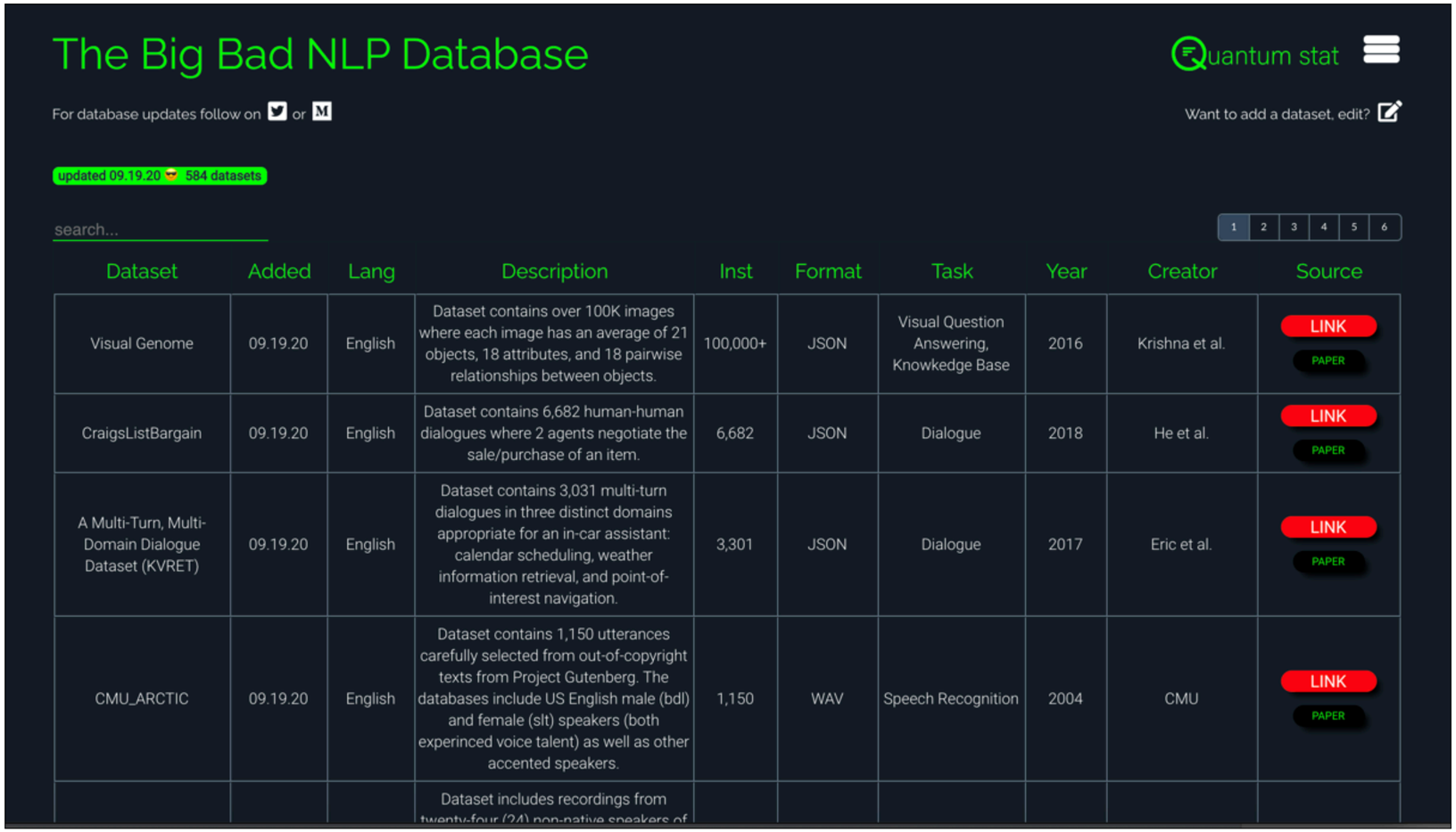
https://www.kaggle.com/datasets

Платформа DrivenData



https://www.drivendata.org/

NLP Database



The Big Bad NLP Database

Quantum stat

For database updates follow on or

Want to add a dataset, edit?

updated 09.19.20 584 datasets

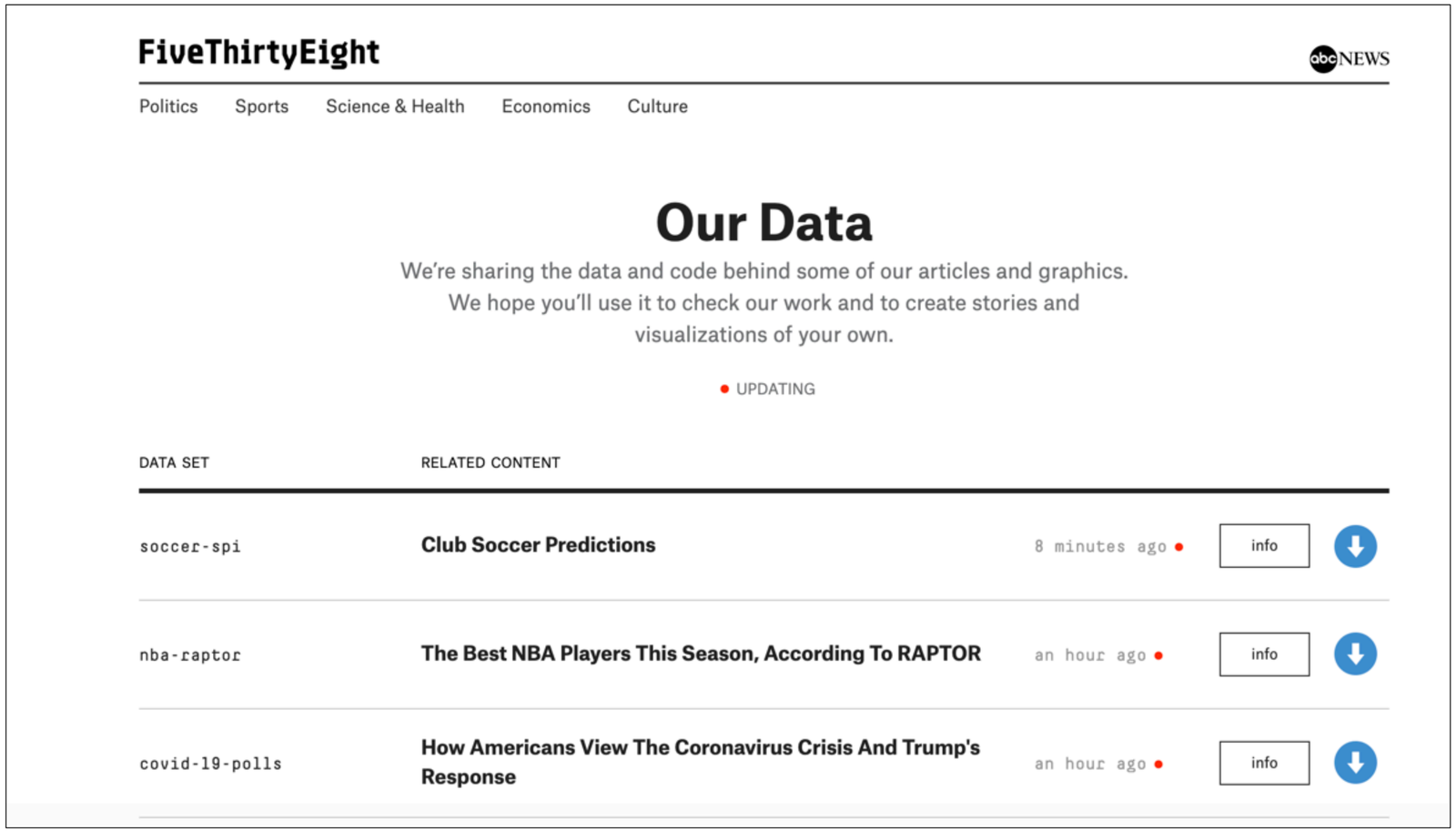
search...

1 2 3 4 5 6

Dataset	Added	Lang	Description	Inst	Format	Task	Year	Creator	Source
Visual Genome	09.19.20	English	Dataset contains over 100K images where each image has an average of 21 objects, 18 attributes, and 18 pairwise relationships between objects.	100,000+	JSON	Visual Question Answering, Knowkedge Base	2016	Krishna et al.	LINK PAPER
CraigsListBargain	09.19.20	English	Dataset contains 6,682 human-human dialogues where 2 agents negotiate the sale/purchase of an item.	6,682	JSON	Dialogue	2018	He et al.	LINK PAPER
A Multi-Turn, Multi-Domain Dialogue Dataset (KVRET)	09.19.20	English	Dataset contains 3,031 multi-turn dialogues in three distinct domains appropriate for an in-car assistant: calendar scheduling, weather information retrieval, and point-of-interest navigation.	3,301	JSON	Dialogue	2017	Eric et al.	LINK PAPER
CMU_ARCTIC	09.19.20	English	Dataset contains 1,150 utterances carefully selected from out-of-copyright texts from Project Gutenberg. The databases include US English male (bdl) and female (slt) speakers (both experinced voice talent) as well as other accented speakers.	1,150	WAV	Speech Recognition	2004	CMU	LINK PAPER
			Dataset includes recordings from twenty-four (24) non-native speakers of						

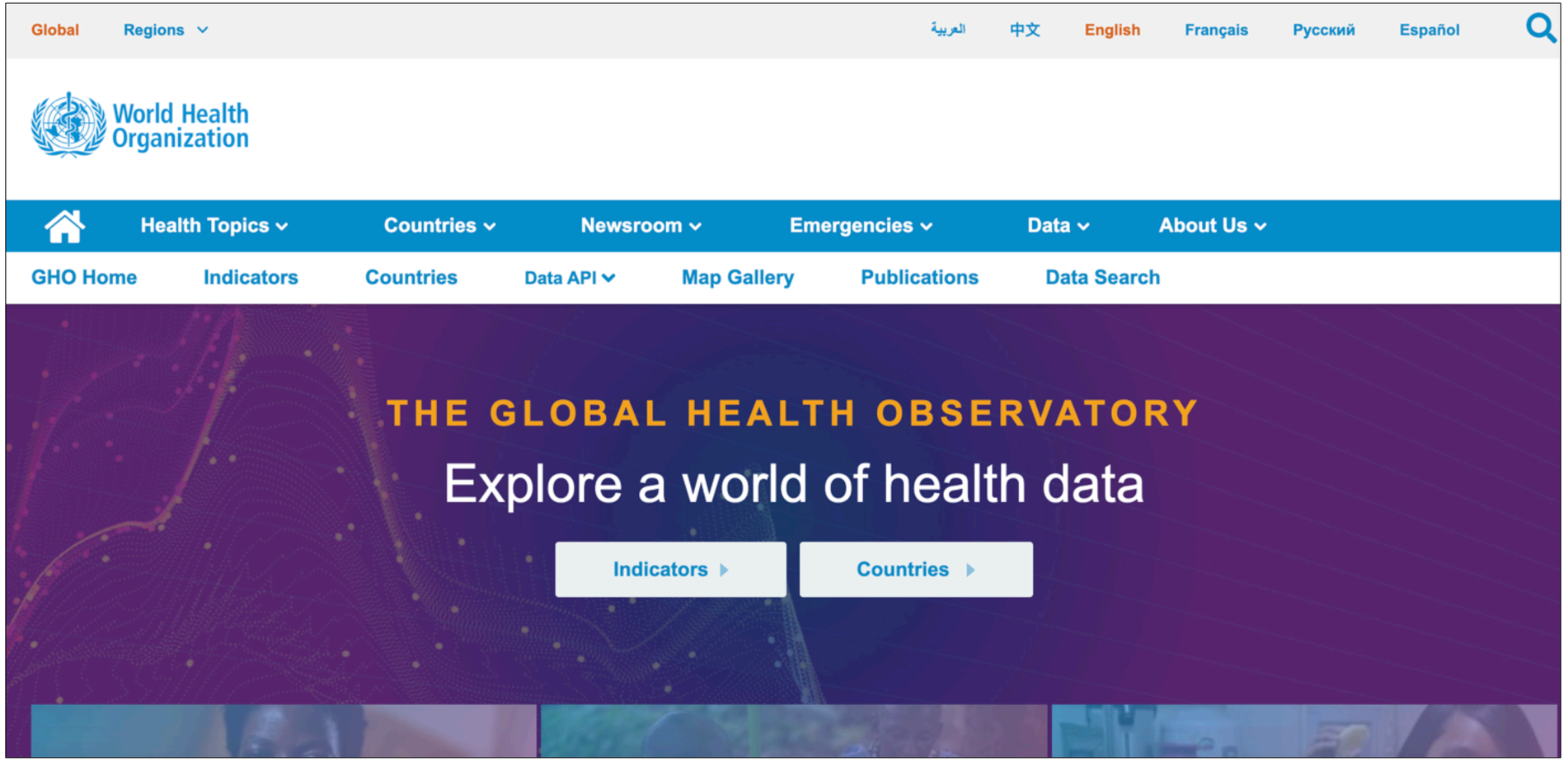
https://datasets.quantumstat.com/

Проект FiveThirtyEight



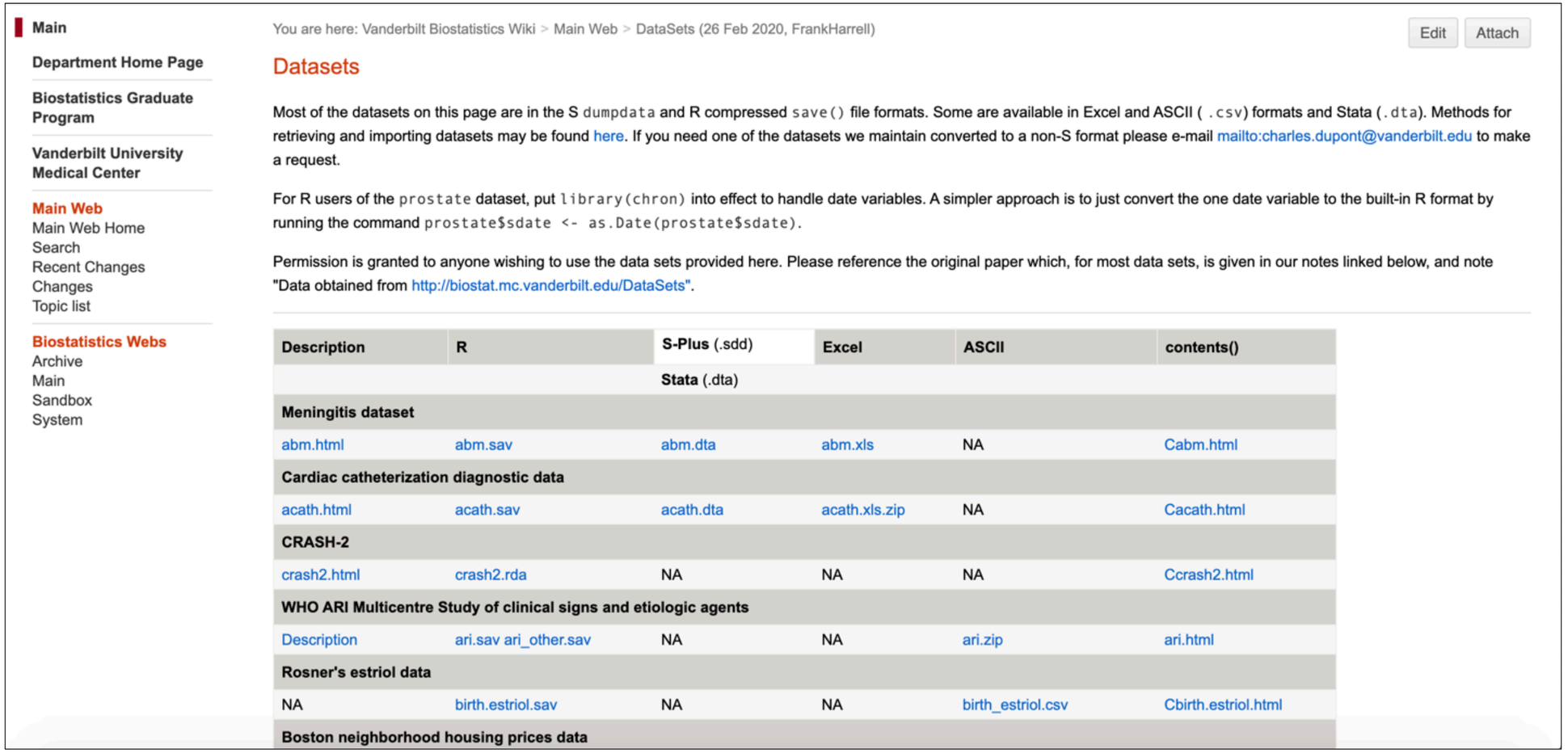
https://data.fivethirtyeight.com/

Медицинские данные: Global Health Data



https://www.who.int/data/gho

Vanderbilt Biostatistics



Main

Department Home Page

Biostatistics Graduate Program

Vanderbilt University Medical Center

Main Web
Main Web Home
Search
Recent Changes
Changes
Topic list

Biostatistics Webs
Archive
Main
Sandbox
System

You are here: Vanderbilt Biostatistics Wiki > Main Web > DataSets (26 Feb 2020, FrankHarrell)

Edit Attach

Datasets

Most of the datasets on this page are in the S `dumpdata` and R compressed `save()` file formats. Some are available in Excel and ASCII (`.csv`) formats and Stata (`.dta`). Methods for retrieving and importing datasets may be found here. If you need one of the datasets we maintain converted to a non-S format please e-mail mailto:charles.dupont@vanderbilt.edu to make a request.

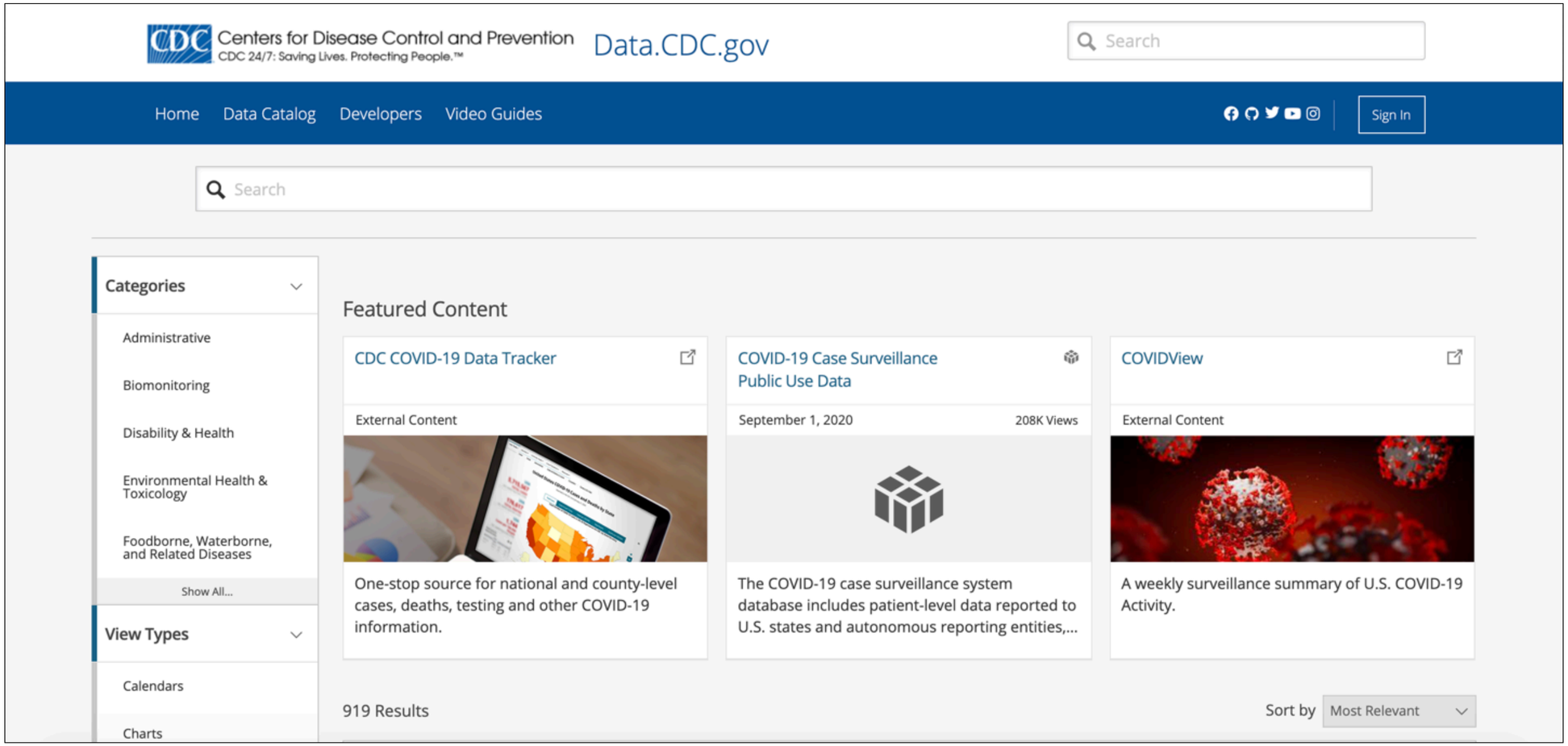
For R users of the `prostate` dataset, put `library(chron)` into effect to handle date variables. A simpler approach is to just convert the one date variable to the built-in R format by running the command `prostate$sdate <- as.Date(prostate$sdate)`.

Permission is granted to anyone wishing to use the data sets provided here. Please reference the original paper which, for most data sets, is given in our notes linked below, and note "Data obtained from http://biostat.mc.vanderbilt.edu/DataSets".

Description	R	S-Plus (.sdd)	Excel	ASCII	contents()
		Stata (.dta)			
Meningitis dataset					
abm.html	abm.sav	abm.dta	abm.xls	NA	Cabm.html
Cardiac catheterization diagnostic data					
acath.html	acath.sav	acath.dta	acath.xls.zip	NA	Cacath.html
CRASH-2					
crash2.html	crash2.rda	NA	NA	NA	Ccrash2.html
WHO ARI Multicentre Study of clinical signs and etiologic agents					
Description	ari.sav ari_other.sav	NA	NA	ari.zip	ari.html
Rosner's estriol data					
NA	birth.estriol.sav	NA	NA	birth_estriol.csv	Cbirth.estriol.html
Boston neighborhood housing prices data					

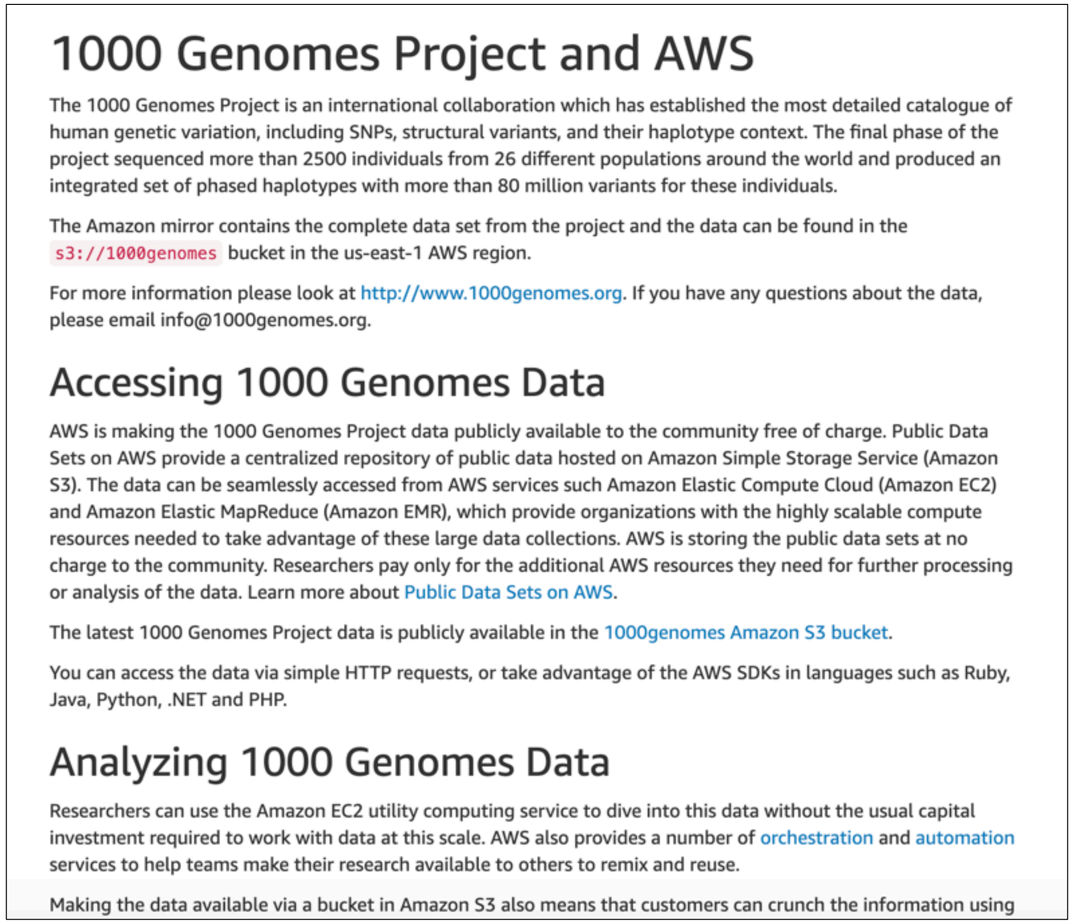
http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets

Centers for Disease Control and Prevention



https://data.cdc.gov/browse

Биологические данные



1000 Genomes Project and AWS

The 1000 Genomes Project is an international collaboration which has established the most detailed catalogue of human genetic variation, including SNPs, structural variants, and their haplotype context. The final phase of the project sequenced more than 2500 individuals from 26 different populations around the world and produced an integrated set of phased haplotypes with more than 80 million variants for these individuals.

The Amazon mirror contains the complete data set from the project and the data can be found in the `s3://1000genomes` bucket in the us-east-1 AWS region.

For more information please look at http://www.1000genomes.org. If you have any questions about the data, please email info@1000genomes.org.

Accessing 1000 Genomes Data

AWS is making the 1000 Genomes Project data publicly available to the community free of charge. Public Data Sets on AWS provide a centralized repository of public data hosted on Amazon Simple Storage Service (Amazon S3). The data can be seamlessly accessed from AWS services such Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Elastic MapReduce (Amazon EMR), which provide organizations with the highly scalable compute resources needed to take advantage of these large data collections. AWS is storing the public data sets at no charge to the community. Researchers pay only for the additional AWS resources they need for further processing or analysis of the data. Learn more about Public Data Sets on AWS.

The latest 1000 Genomes Project data is publicly available in the 1000genomes Amazon S3 bucket.

You can access the data via simple HTTP requests, or take advantage of the AWS SDKs in languages such as Ruby, Java, Python, .NET and PHP.

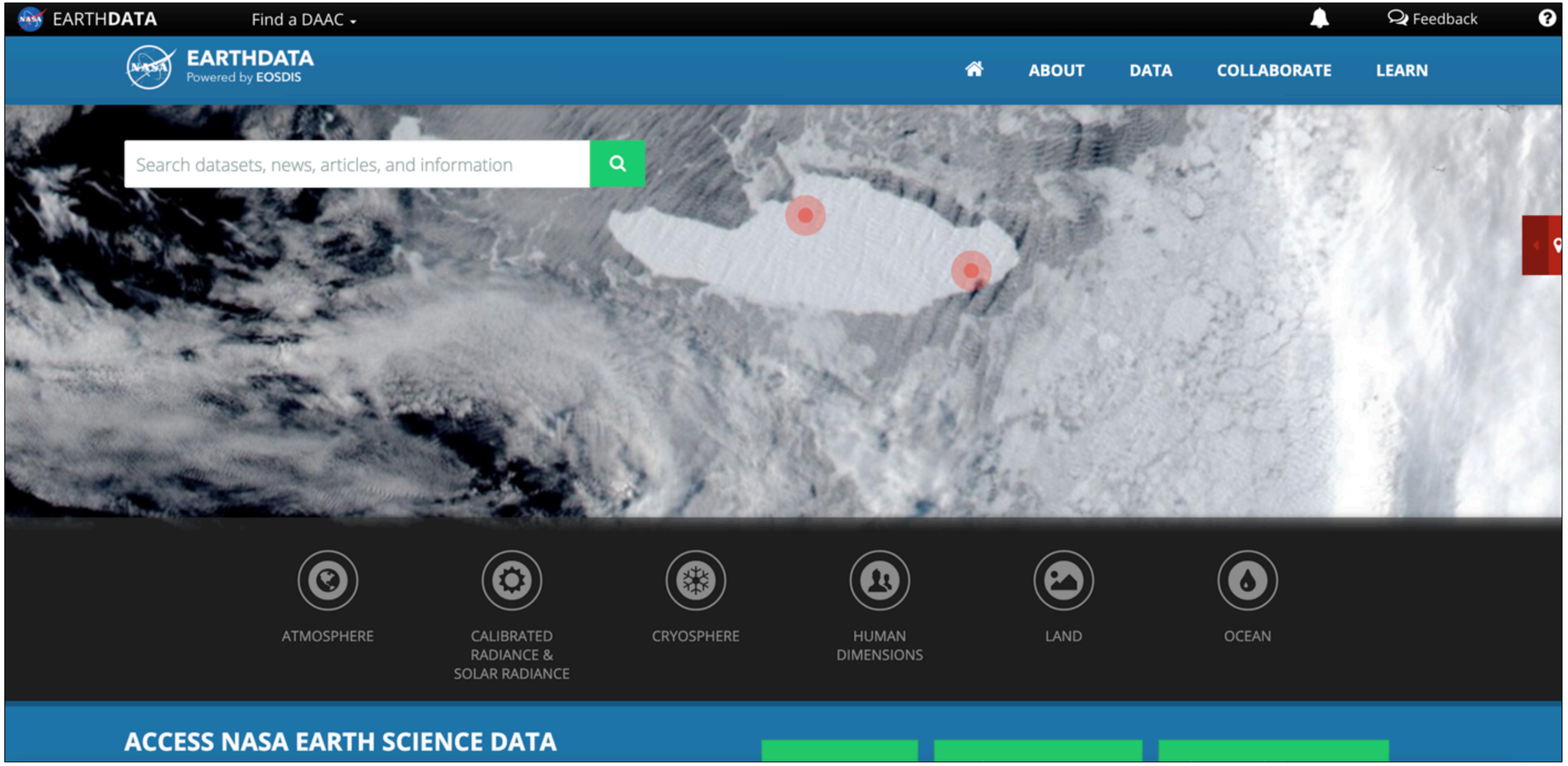
Analyzing 1000 Genomes Data

Researchers can use the Amazon EC2 utility computing service to dive into this data without the usual capital investment required to work with data at this scale. AWS also provides a number of orchestration and automation services to help teams make their research available to others to remix and reuse.

Making the data available via a bucket in Amazon S3 also means that customers can crunch the information using

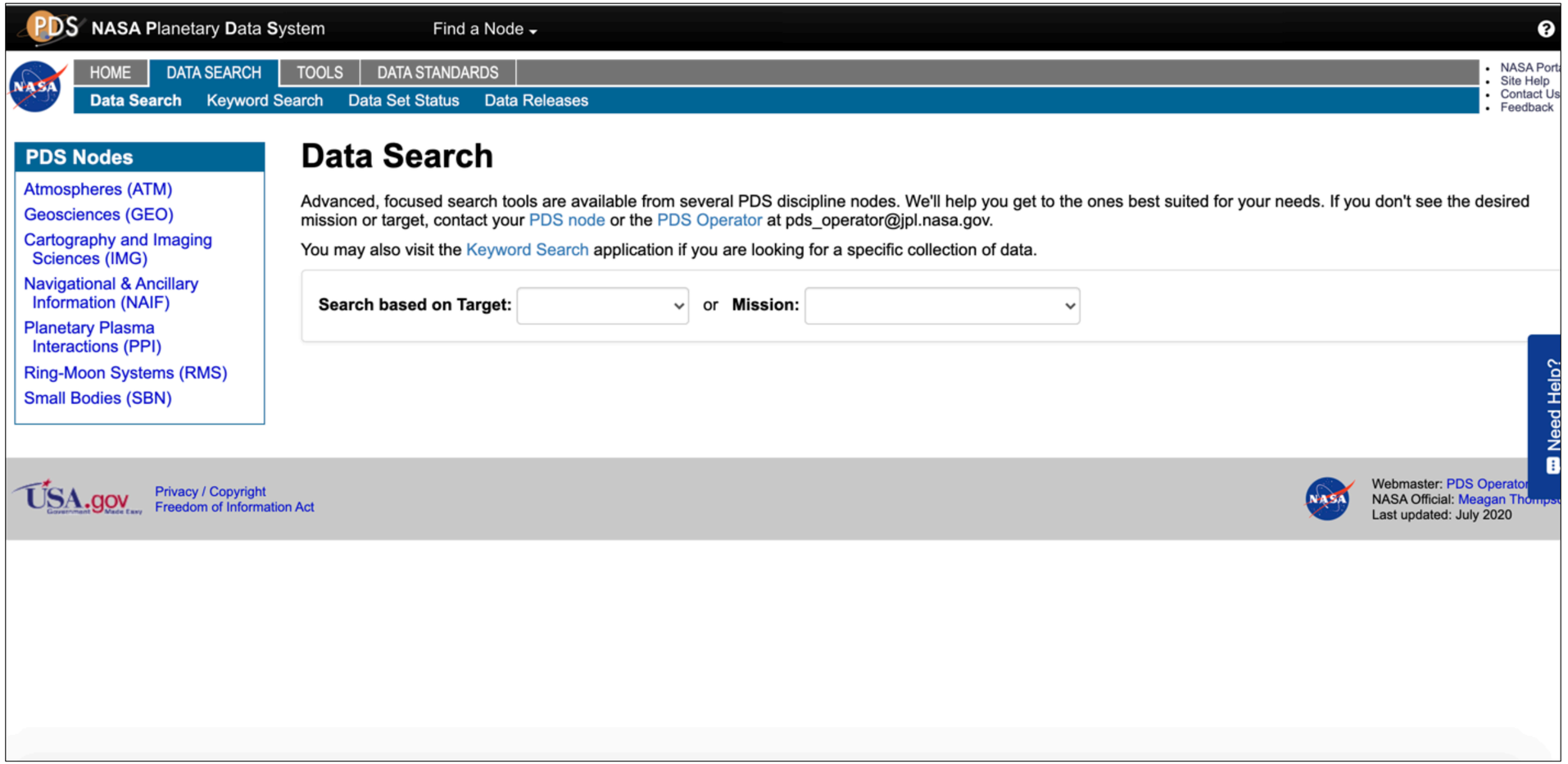
https://docs.opendata.aws/1000genomes/readme.html

NASA: Earth data



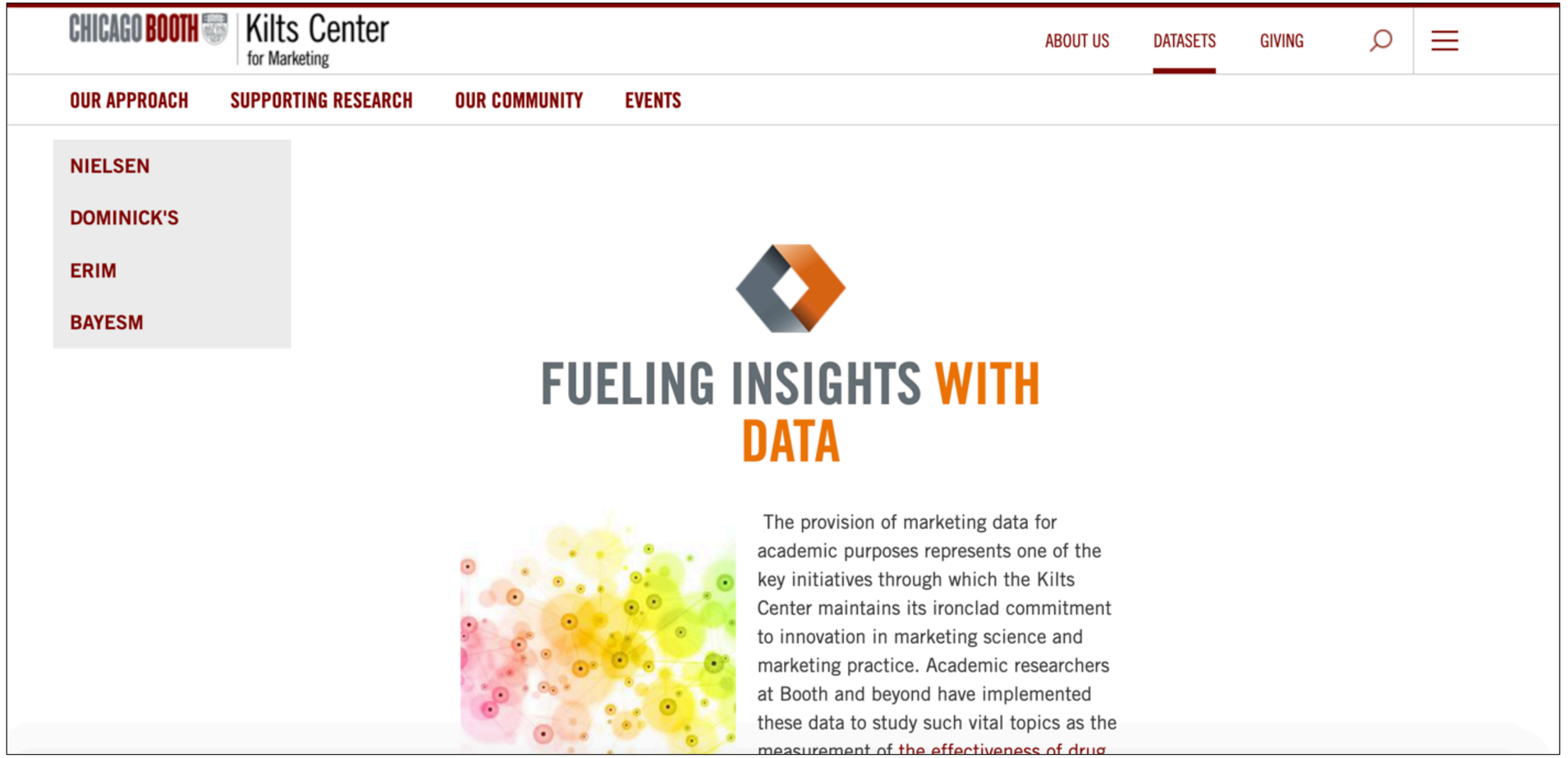
https://earthdata.nasa.gov/

NASA: Space data



https://pds.nasa.gov/datasearch/data-search/

Маркетинговые данные

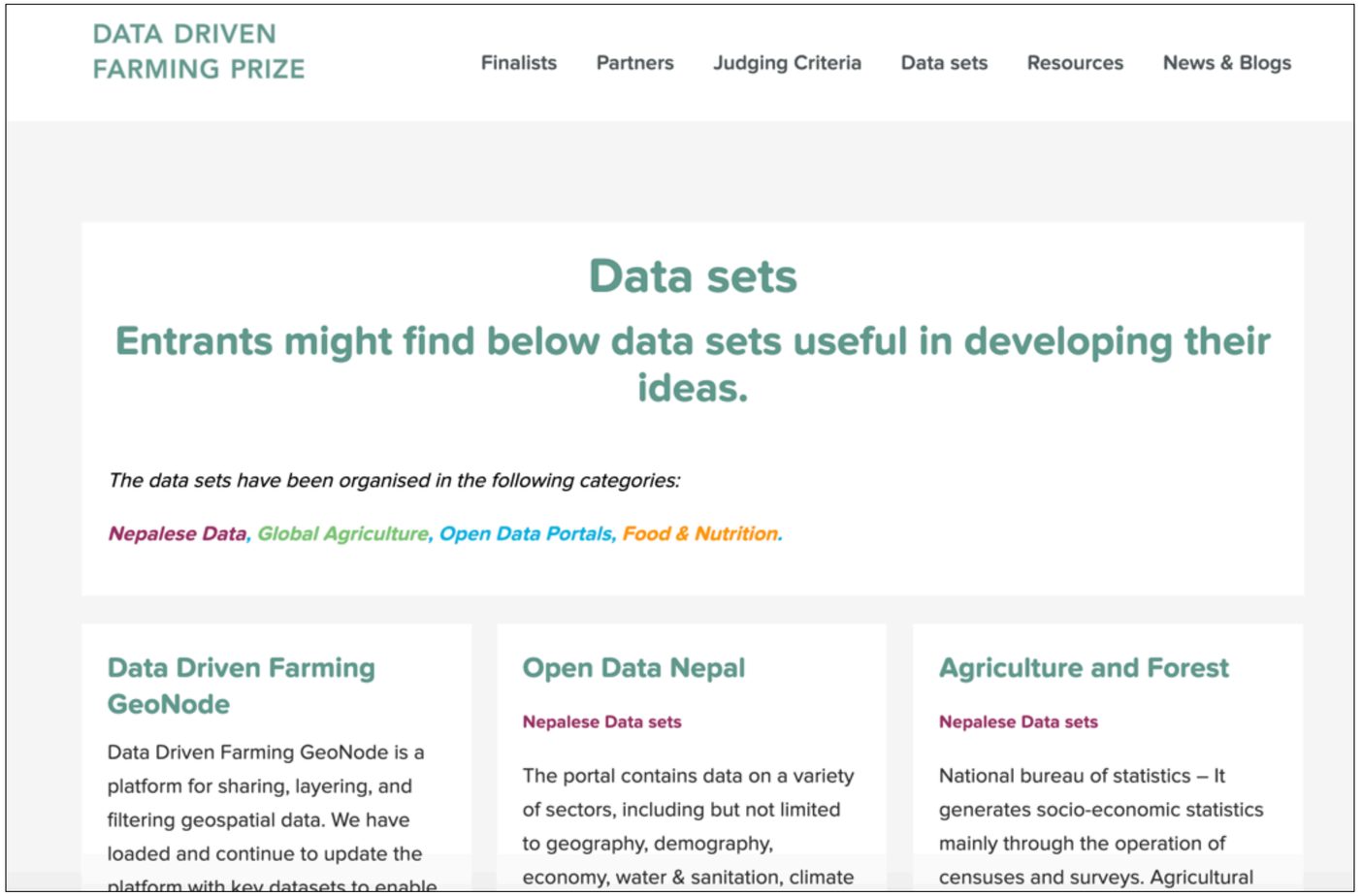


https://www.chicagobooth.edu/research/kilts/datasets

Экономические данные

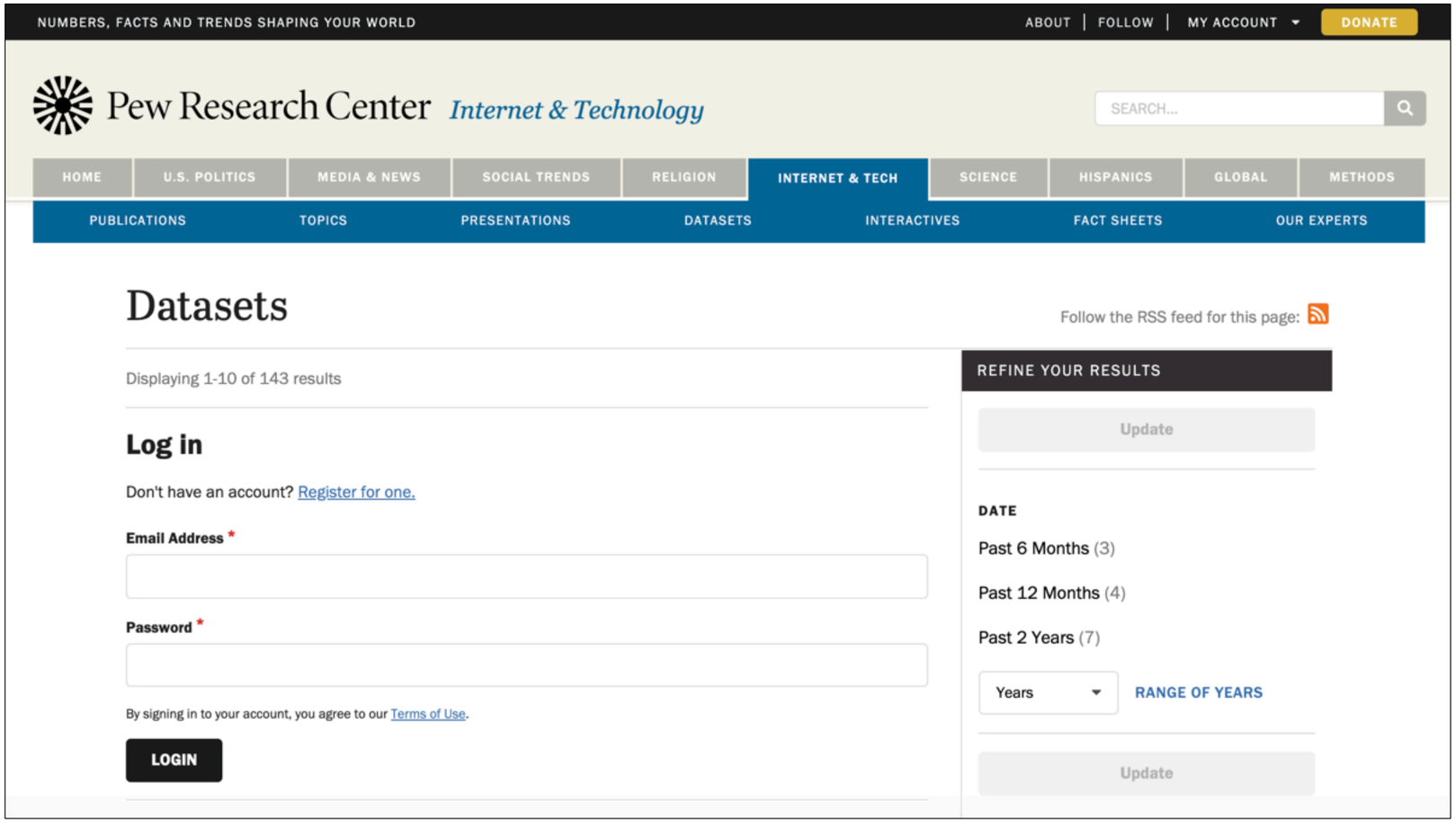
https://data.nber.org/data/

Сельскохозяйственные данные



https://datadrivenfarming.challenges.org/data-sets-2/

Pew Research Centre



https://www.pewresearch.org/internet/datasets/

MNIST database

THE MNIST DATABASE

of handwritten digits

Yann LeCun, Courant Institute, NYU
Corinna Cortes, Google Labs, New York
Christopher J.C. Burges, Microsoft Research, Redmond

The MNIST database of handwritten digits, available from this page, has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a fixed-size image.

It is a good database for people who want to try learning techniques and pattern recognition methods on real-world data while spending minimal efforts on preprocessing and formatting.

Four files are available on this site:

```
train-images-idx3-ubyte.gz:  training set images (9912422 bytes)
train-labels-idx1-ubyte.gz:  training set labels (28881 bytes)
t10k-images-idx3-ubyte.gz:   test set images (1648877 bytes)
t10k-labels-idx1-ubyte.gz:   test set labels (4542 bytes)
```

please note that your browser may uncompress these files without telling you. If the files you downloaded have a larger size than the above, they have been uncompressed by your browser. Simply rename them to remove the .gz extension. Some people have asked me "my application can't open your image files". These files are not in any standard image format. You have to write your own (very simple) program to read them. The file format is described at the bottom of this page.

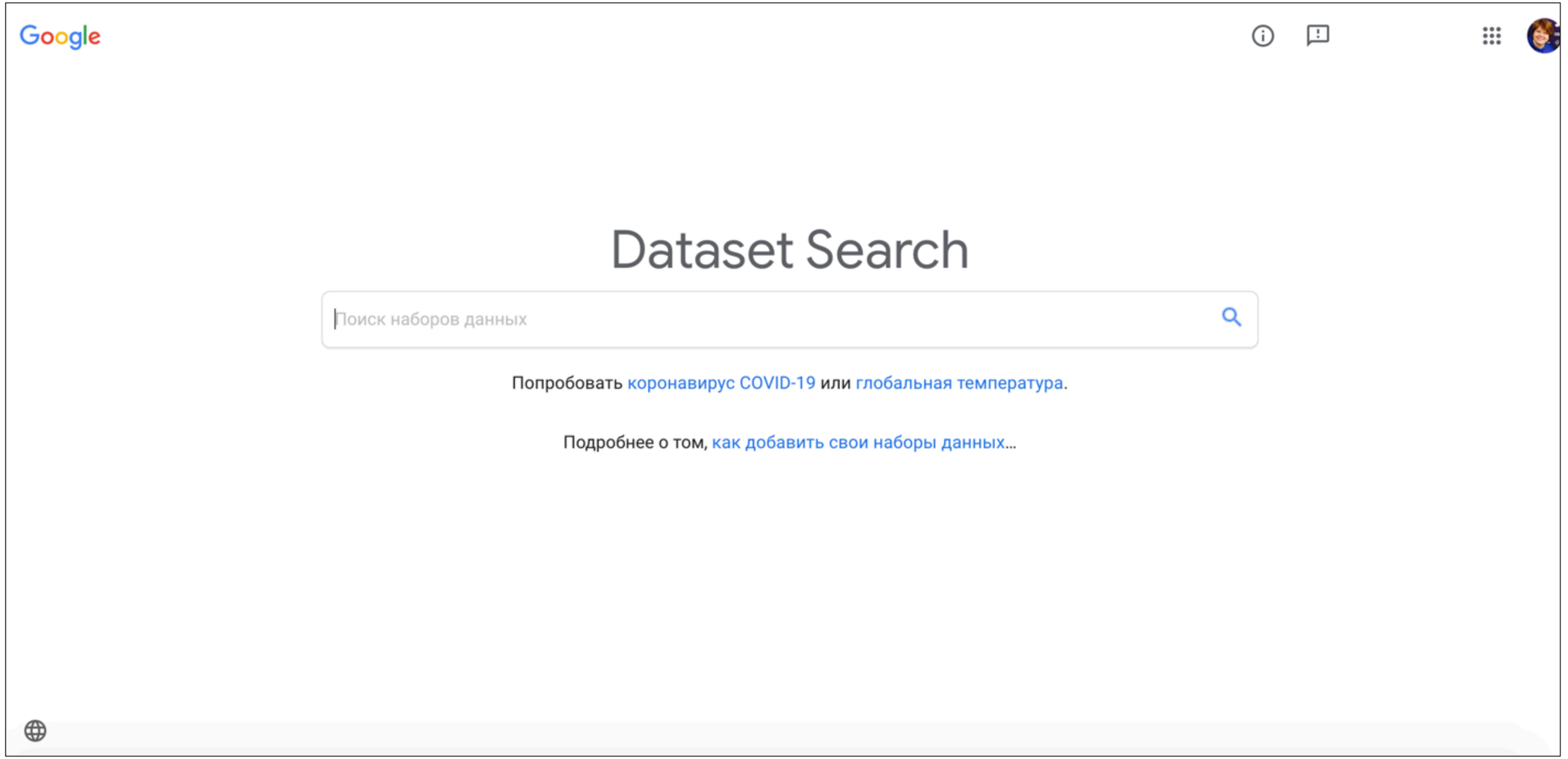
The original black and white (bilevel) images from NIST were size normalized to fit in a 20x20 pixel box while preserving their aspect ratio. The resulting images contain grey levels as a result of the anti-aliasing techniq used by the normalization algorithm. the images were centered in a 28x28 image by computing the center of mass of the pixels, and translating the image so as to position this point at the center of the 28x28 field.

With some classification methods (particuarly template-based methods, such as SVM and K-nearest neighbors), the error rate improves when the digits are centered by bounding box rather than center of mass. If you do this kind of pre-processing, you should report it in your publications.

The MNIST database was constructed from NIST's Special Database 3 and Special Database 1 which contain binary images of handwritten digits. NIST originally designated SD-3 as their training set and SD-1 as their te set. However, SD-3 is much cleaner and easier to recognize than SD-1. The reason for this can be found on the fact that SD-3 was collected among Census Bureau employees, while SD-1 was collected among high-schoo students. Drawing sensible conclusions from learning experiments requires that the result be independent of the choice of training set and test among the complete set of samples. Therefore it was necessary to build a new database by mixing NIST's datasets.

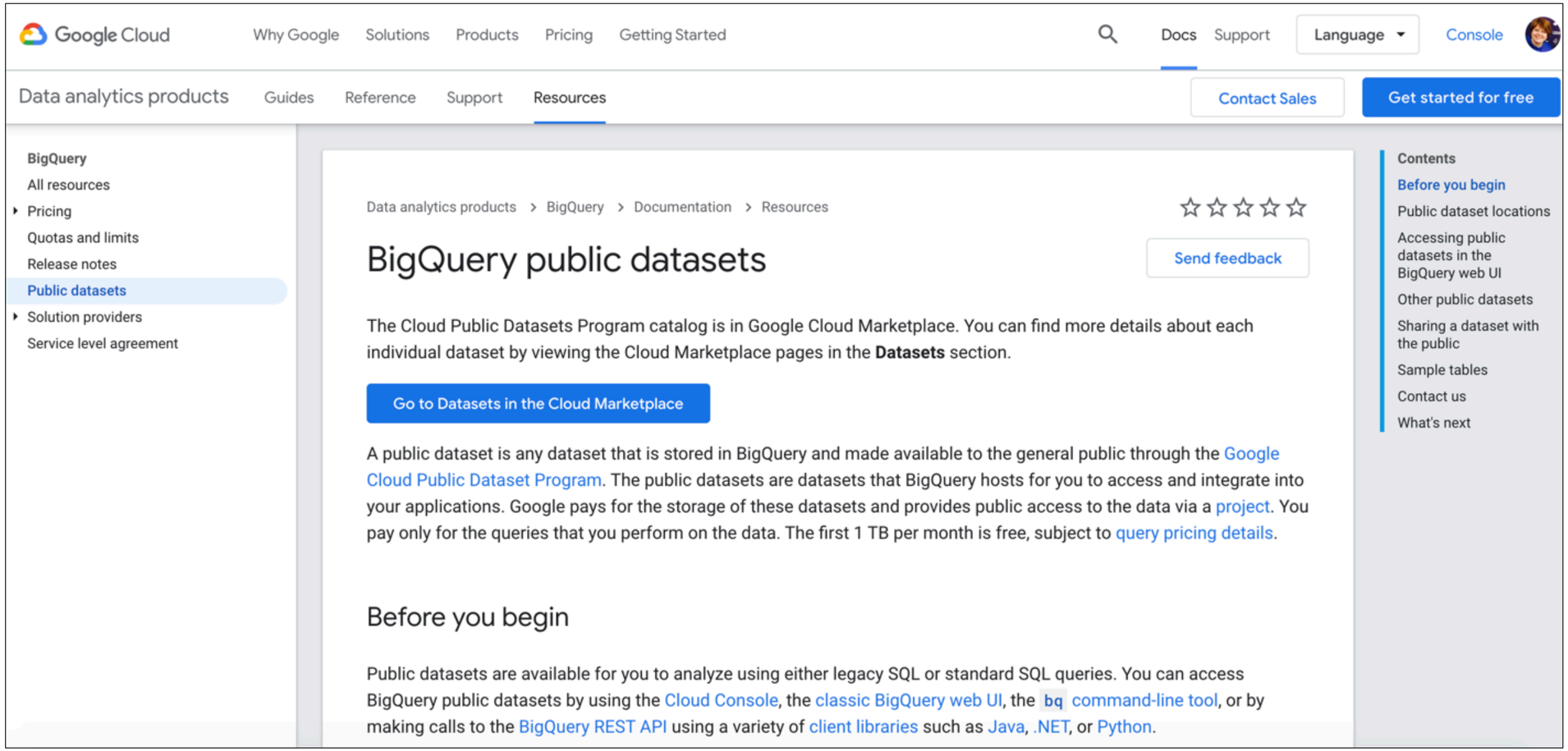
http://yann.lecun.com/exdb/mnist/

Google: Проект Dataset Search



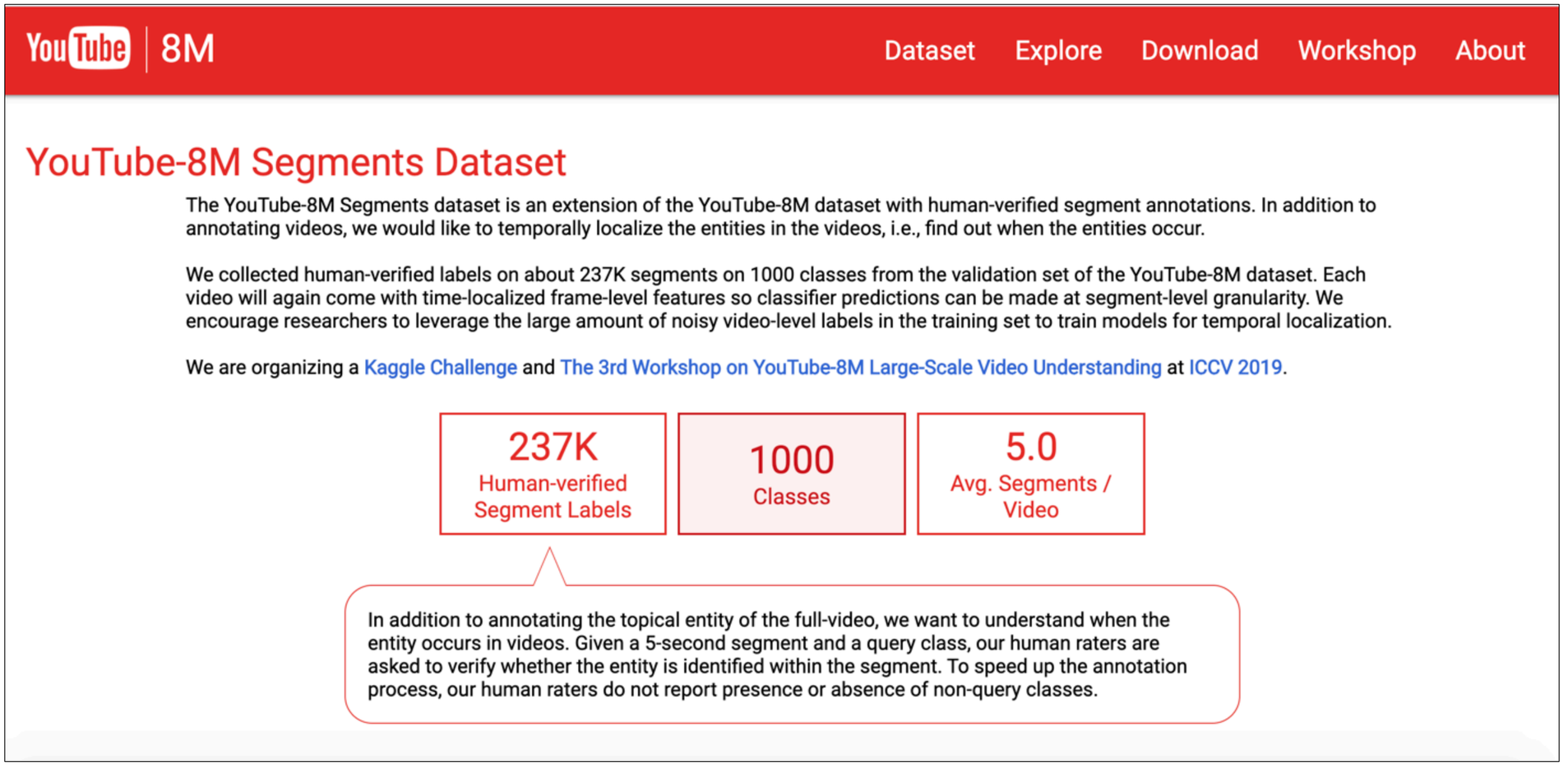
https://datasetsearch.research.google.com/

Google: BigQuery public datasets



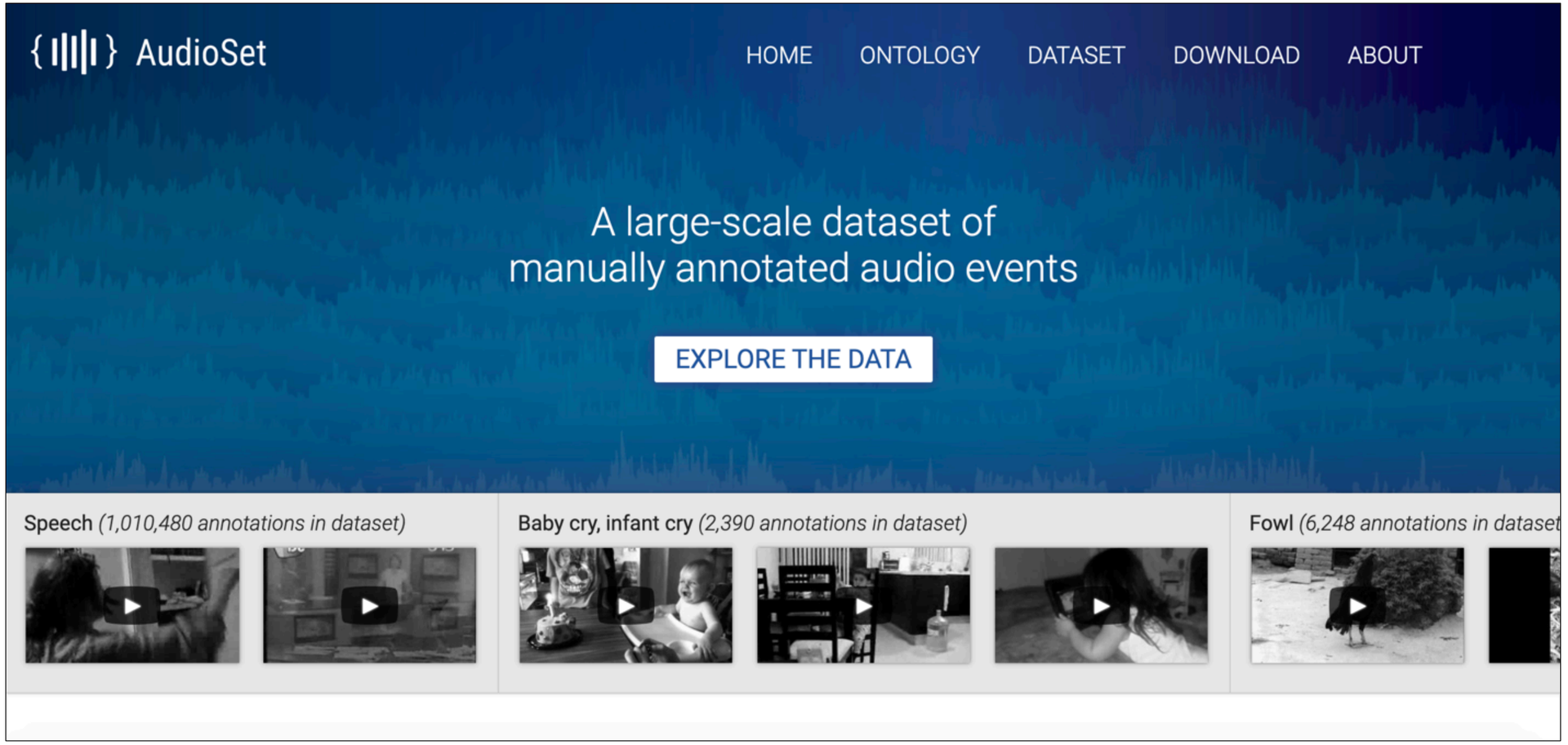
https://cloud.google.com/bigquery/public-data/

Google: YouTube-8M Segments Dataset



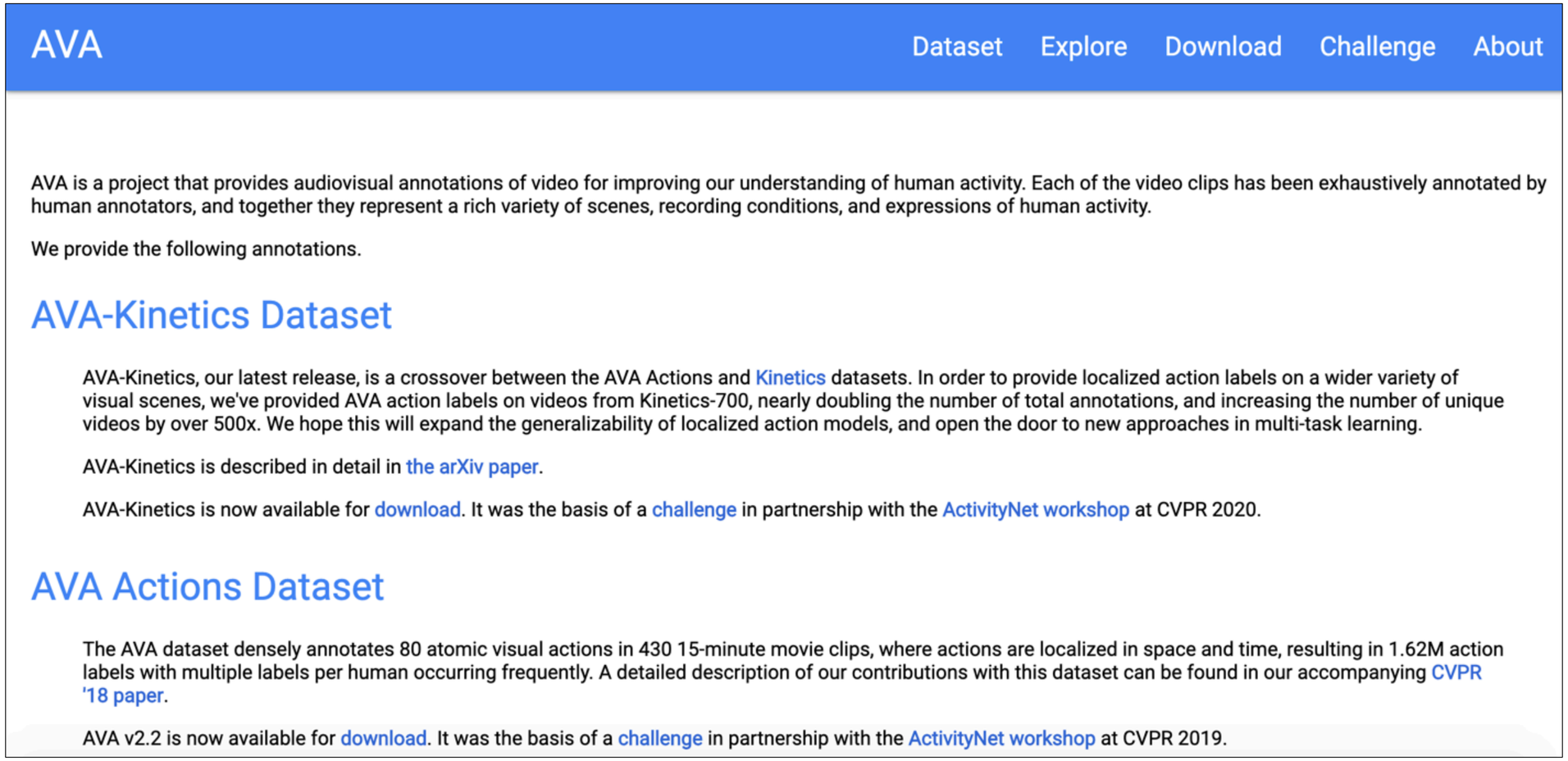
https://research.google.com/youtube8m/

Google: AudioSet



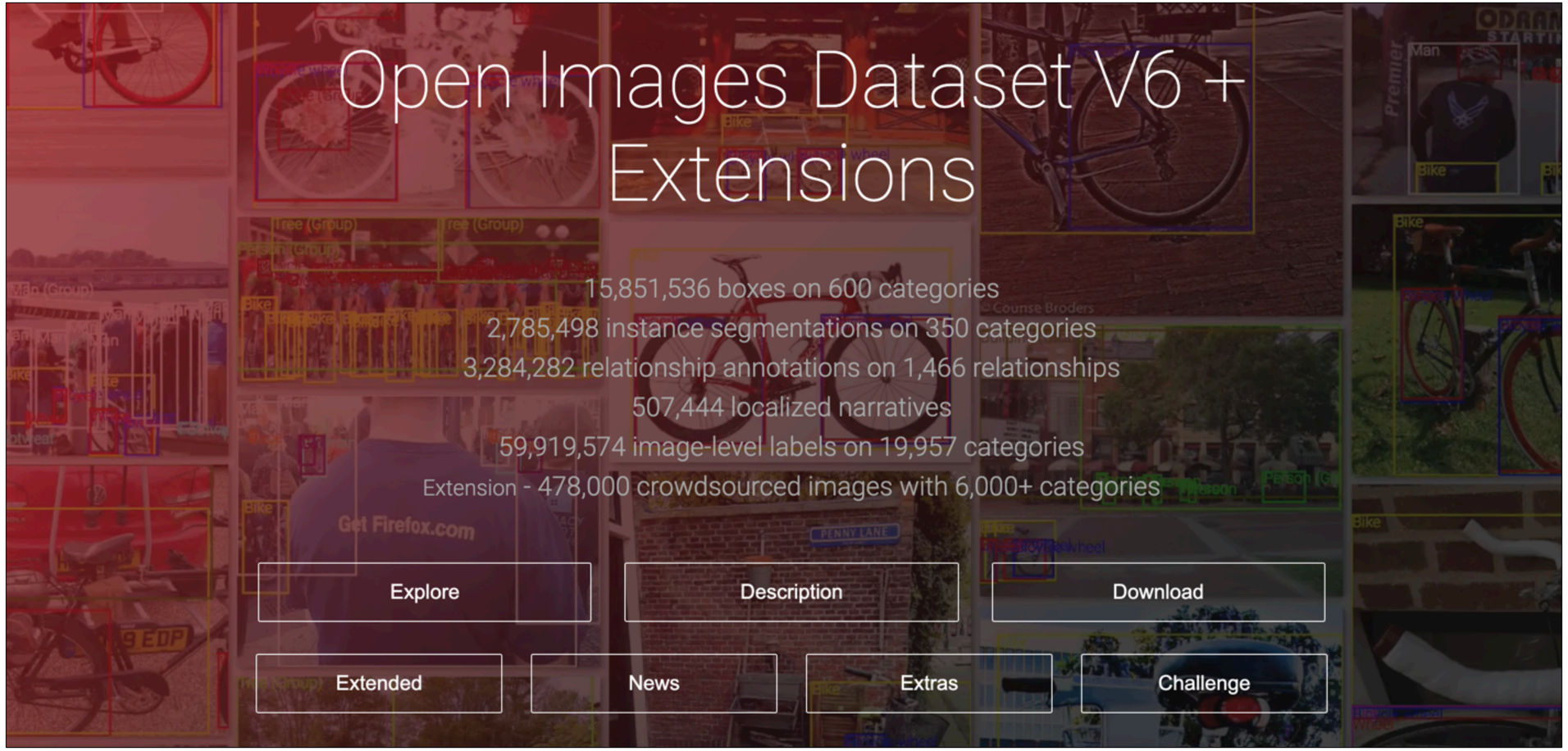
https://research.google.com/audioset/

Google: AVA, audiovisual annotations of video



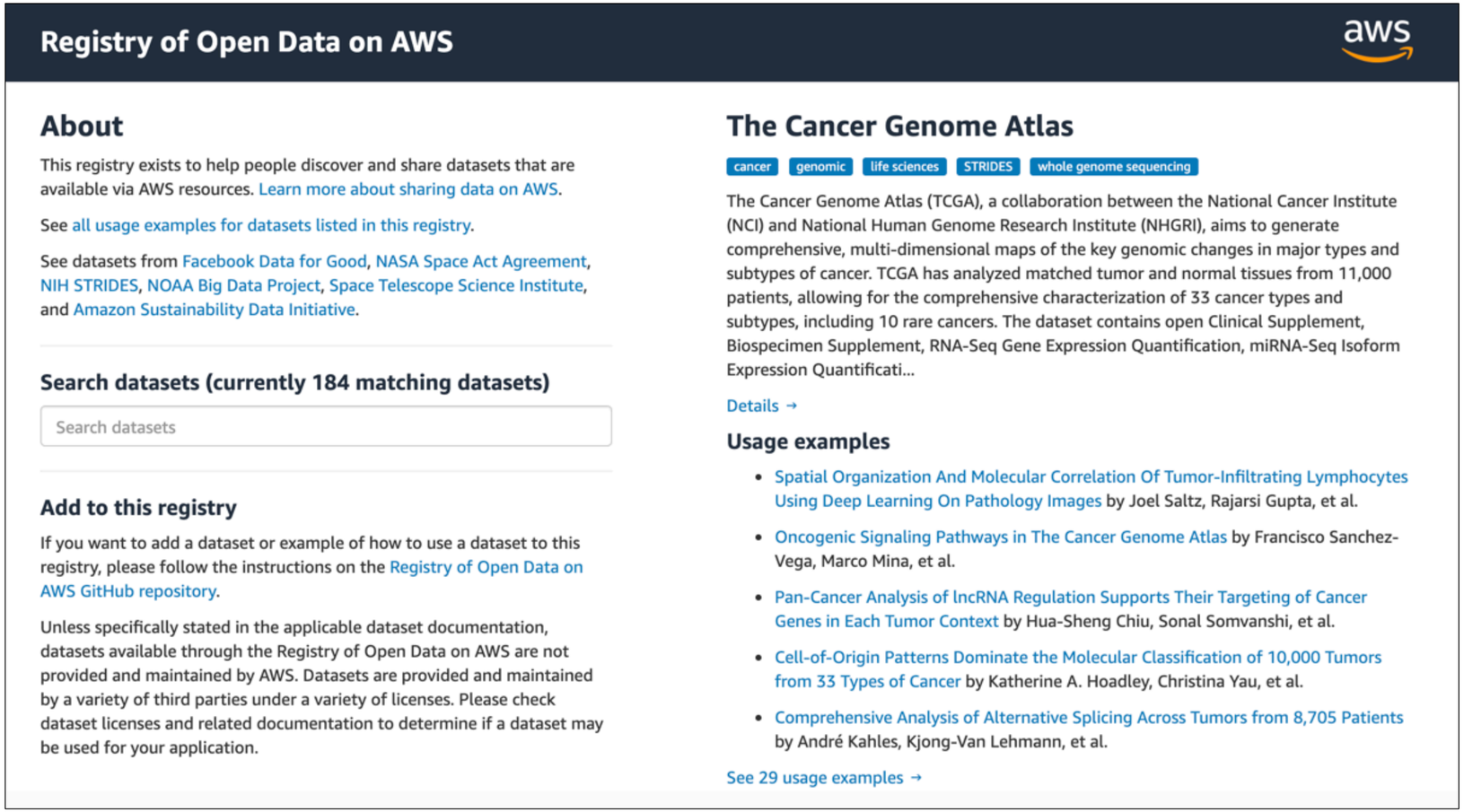
https://research.google.com/ava/

Google: Open Image



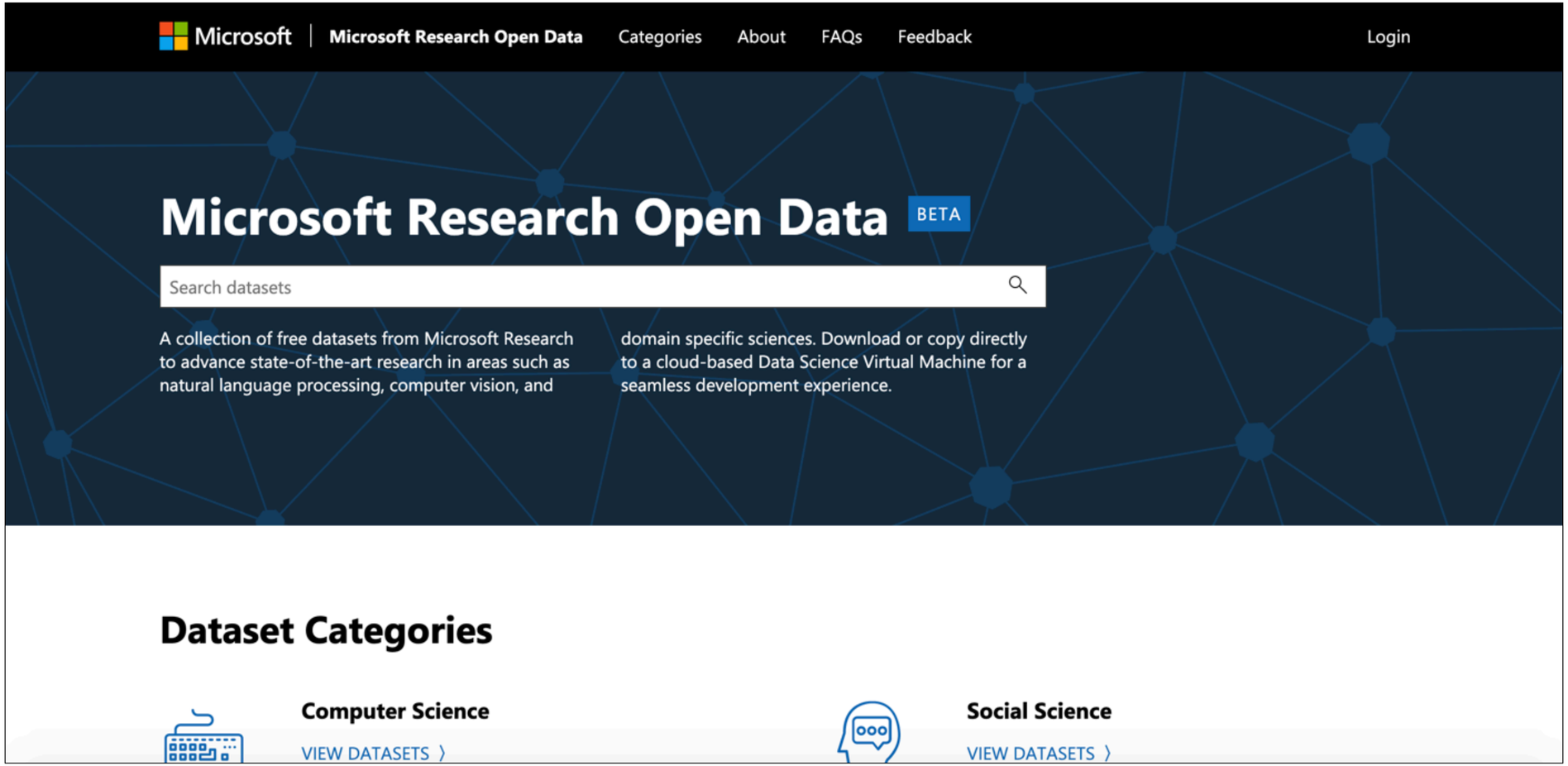
https://storage.googleapis.com/openimages/web/index.html

Open Data on Amazon



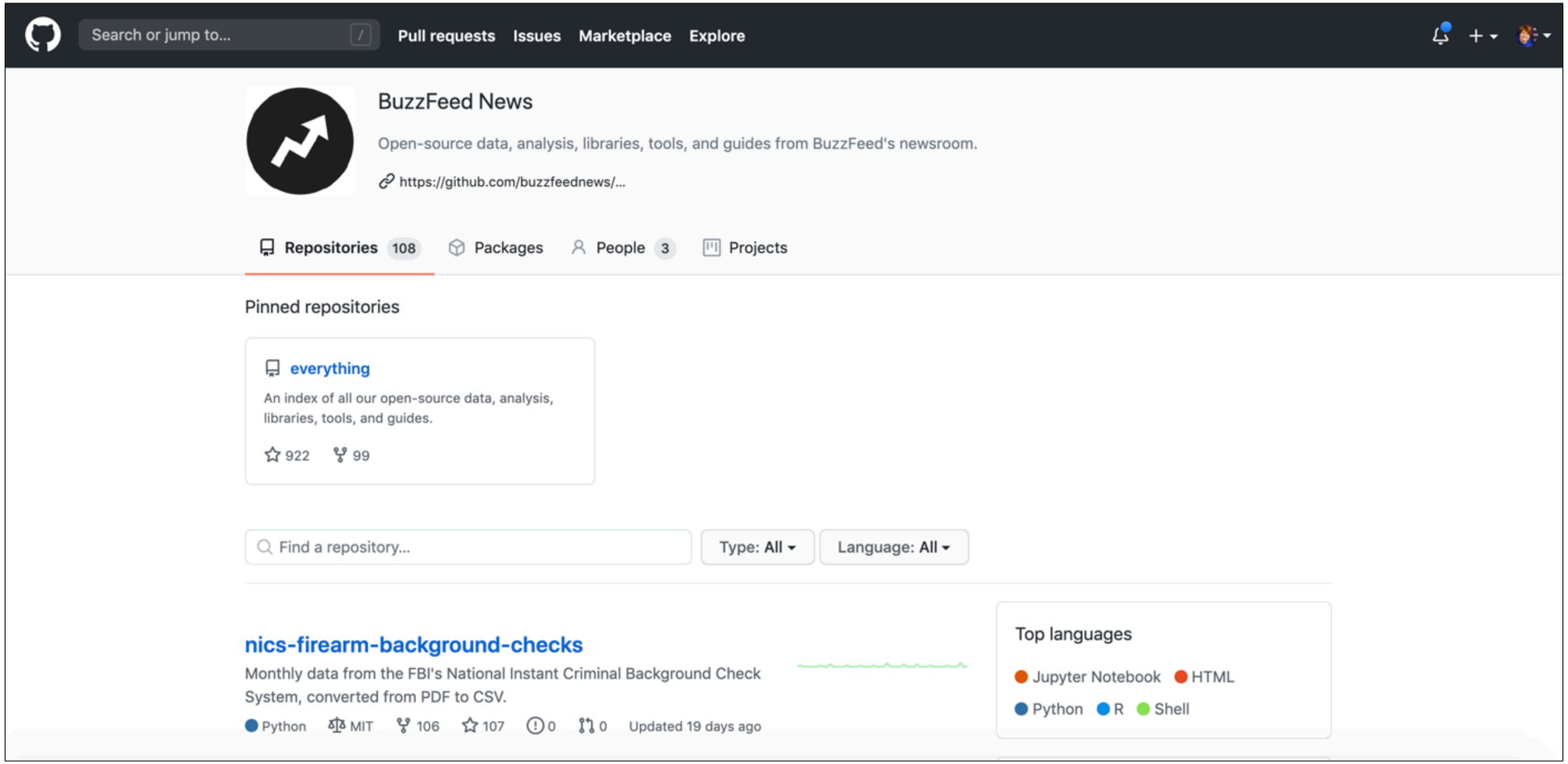
https://registry.opendata.aws/

Microsoft Research Open Data



https://msropendata.com/

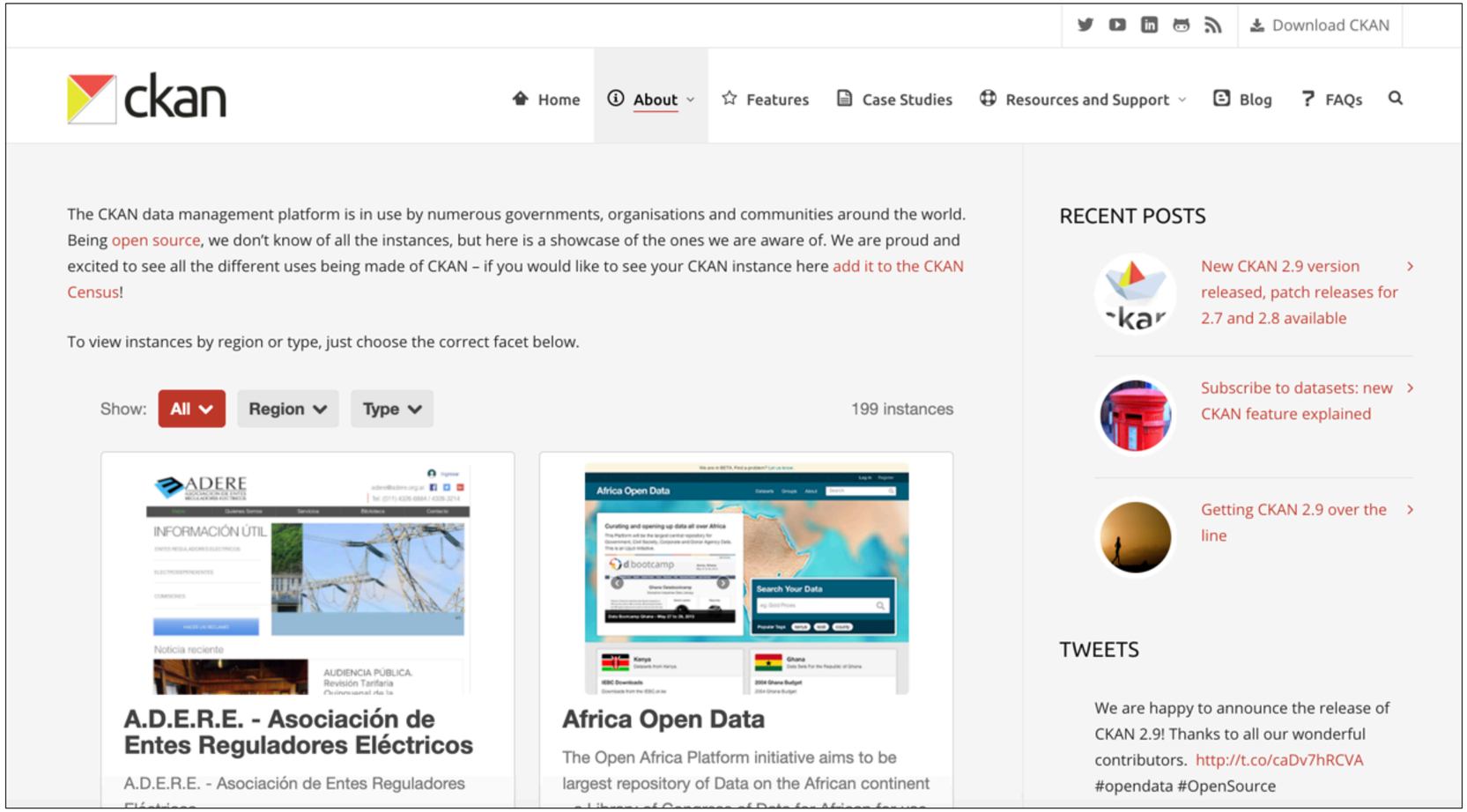
BuzzFeed



https://github.com/BuzzFeedNews

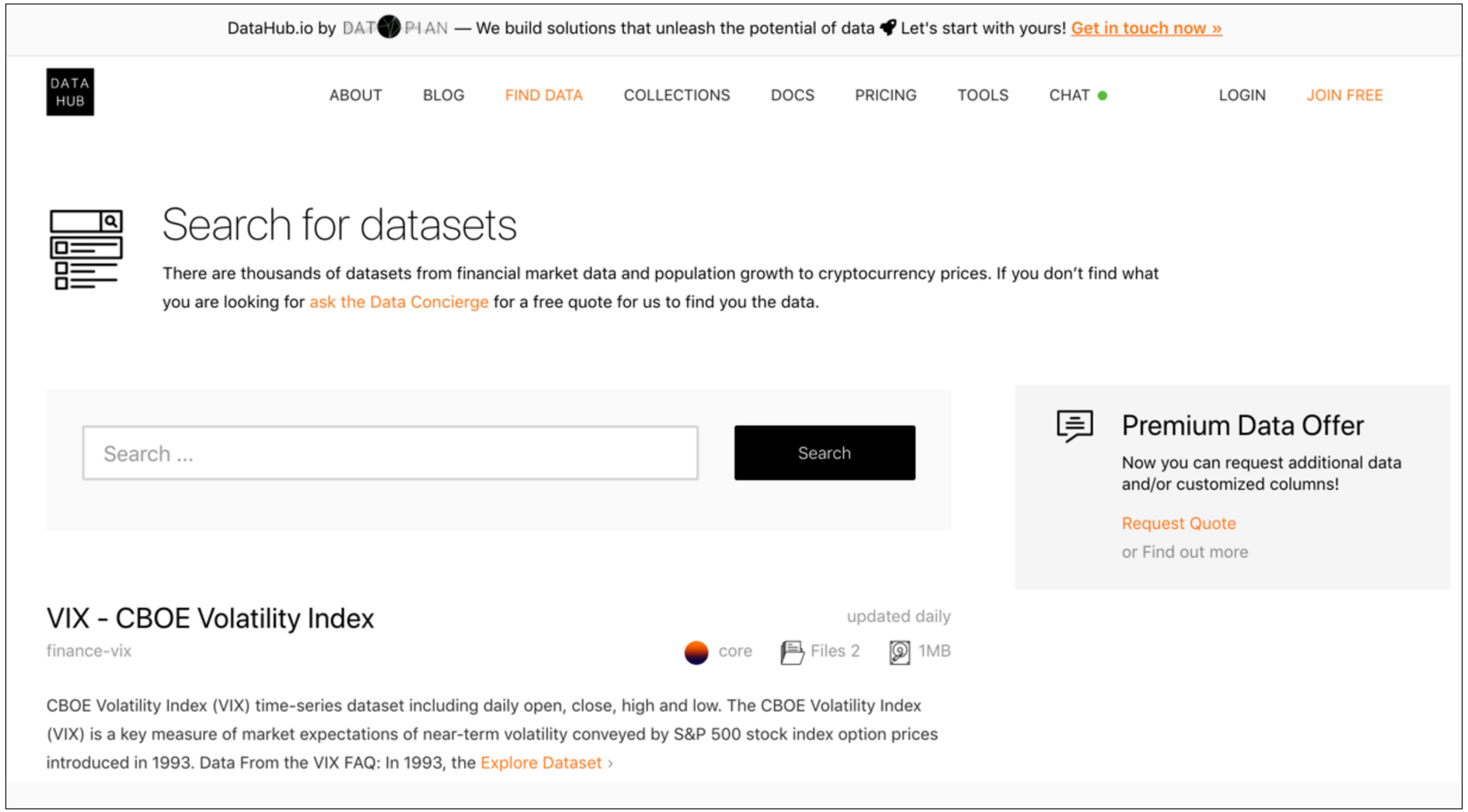
Data Sharing platforms

CKAN



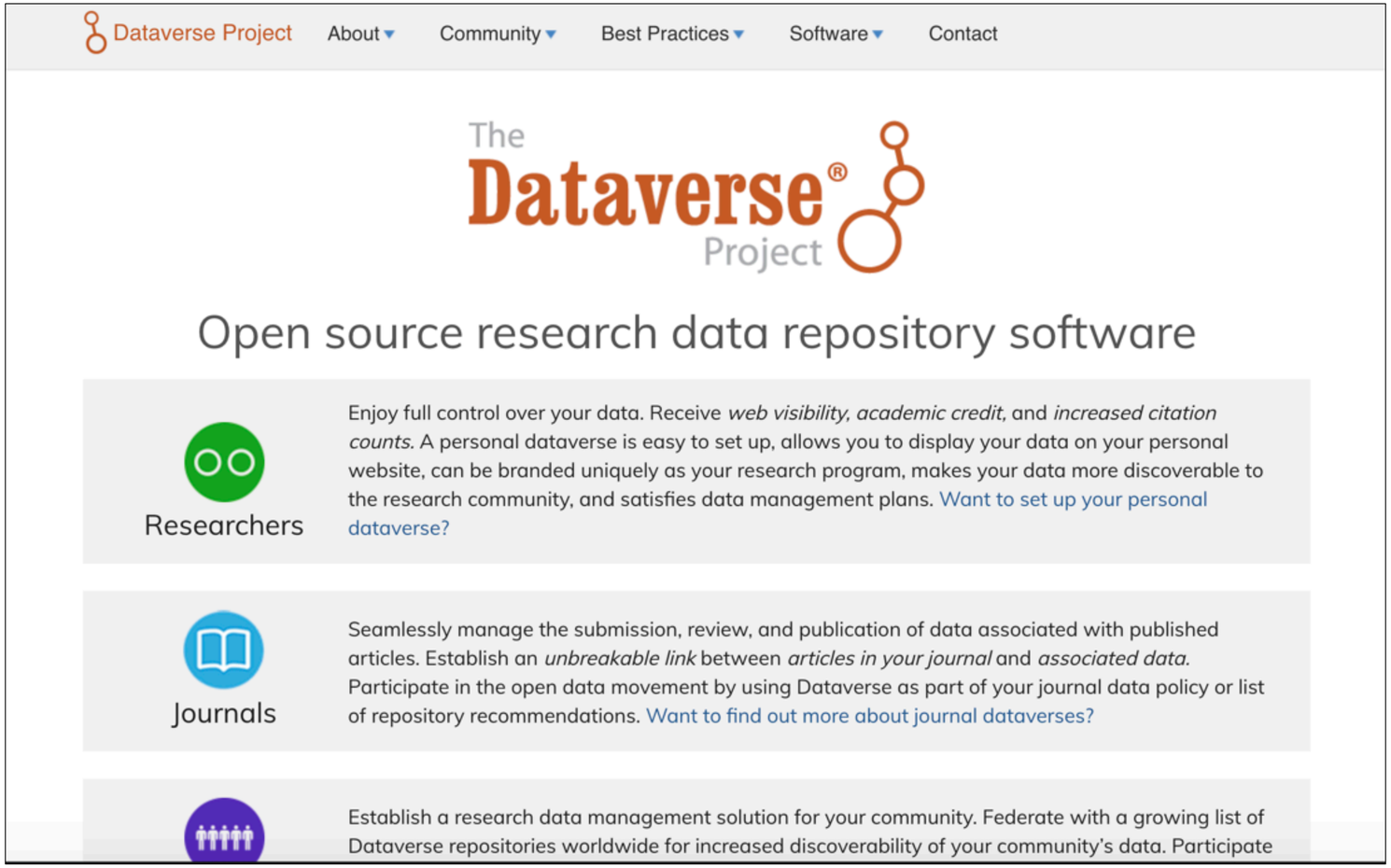
https://ckan.org/about/instances/

DataHub



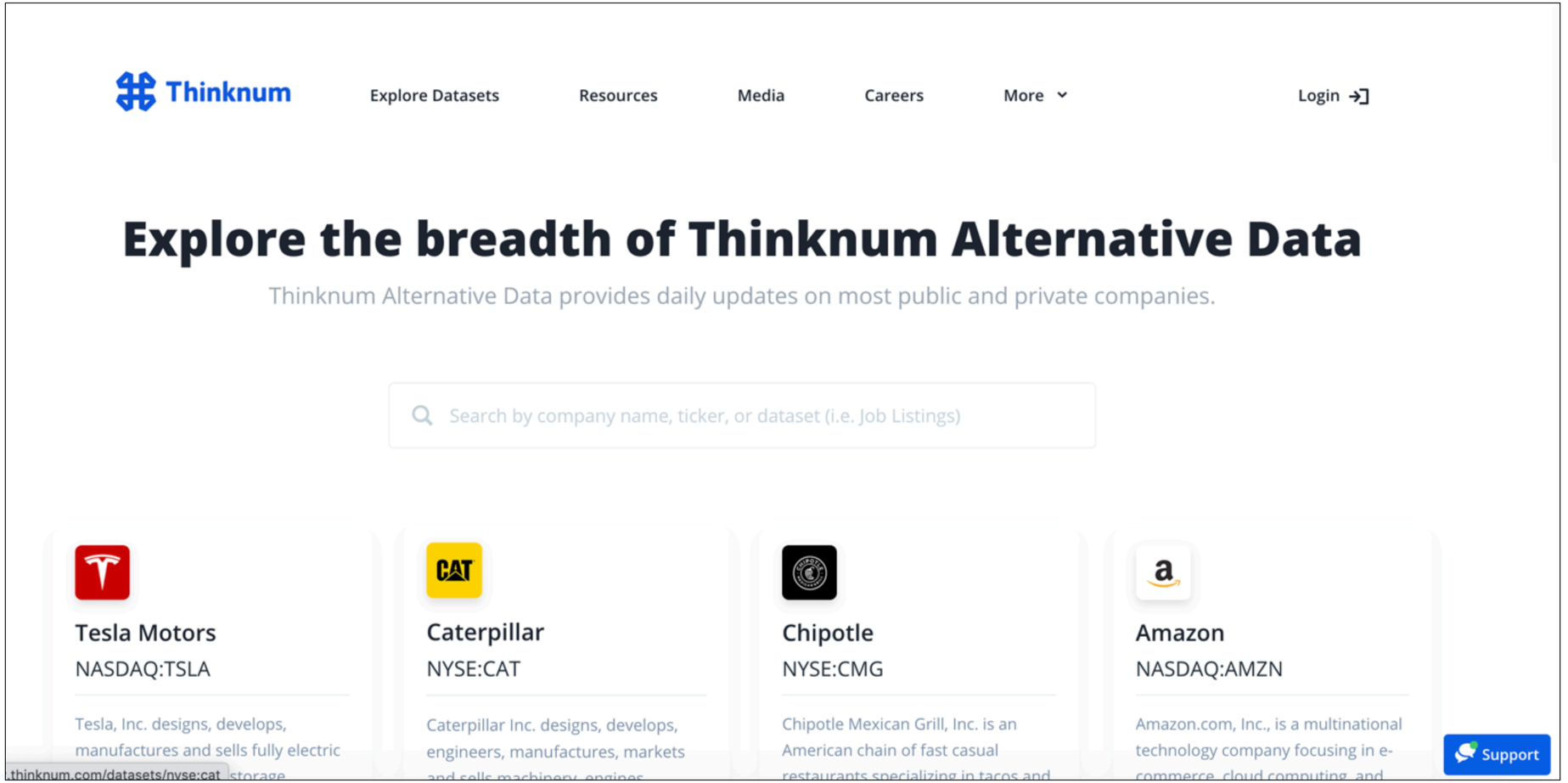
https://datahub.io/search

Dataverse



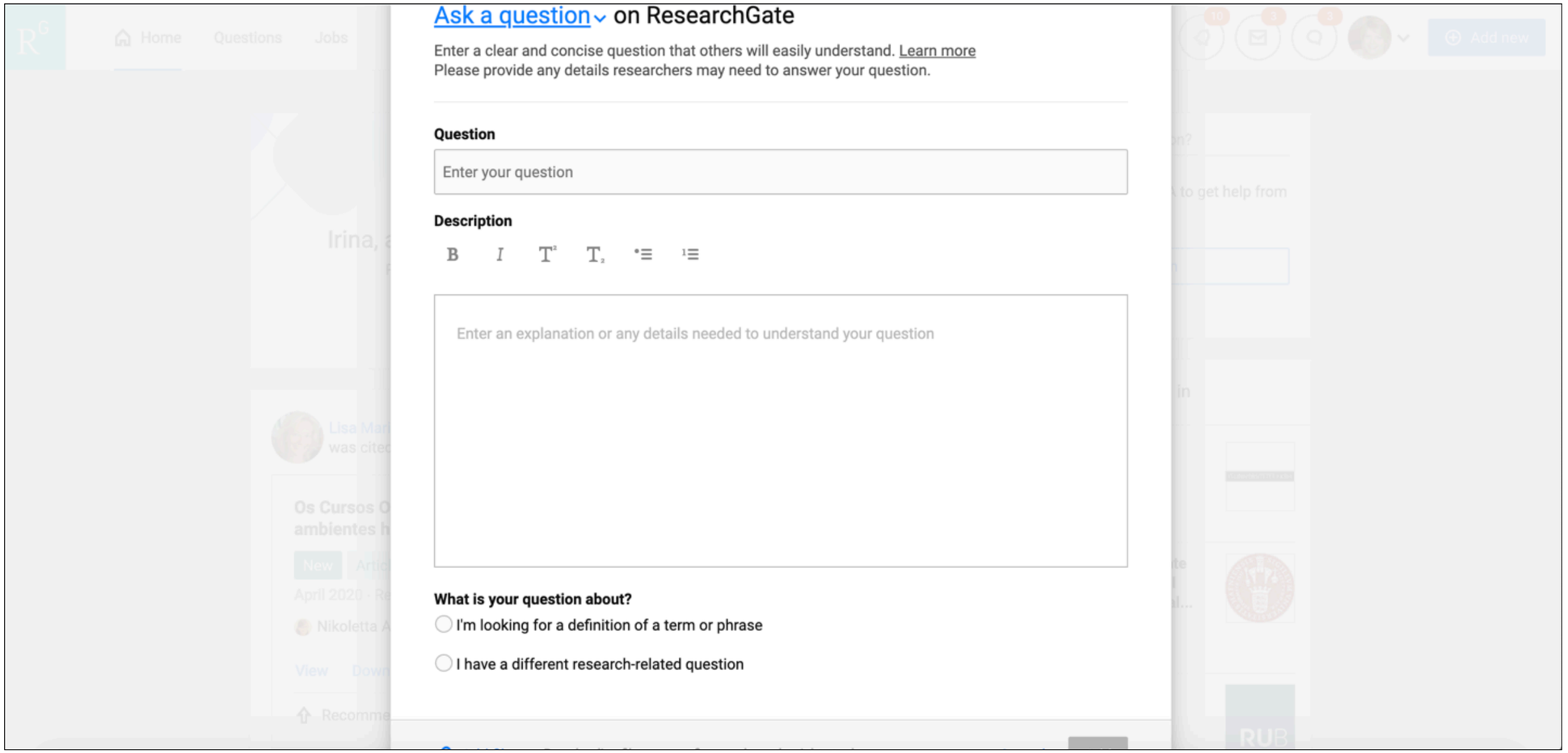
https://dataverse.org/

Thinknum



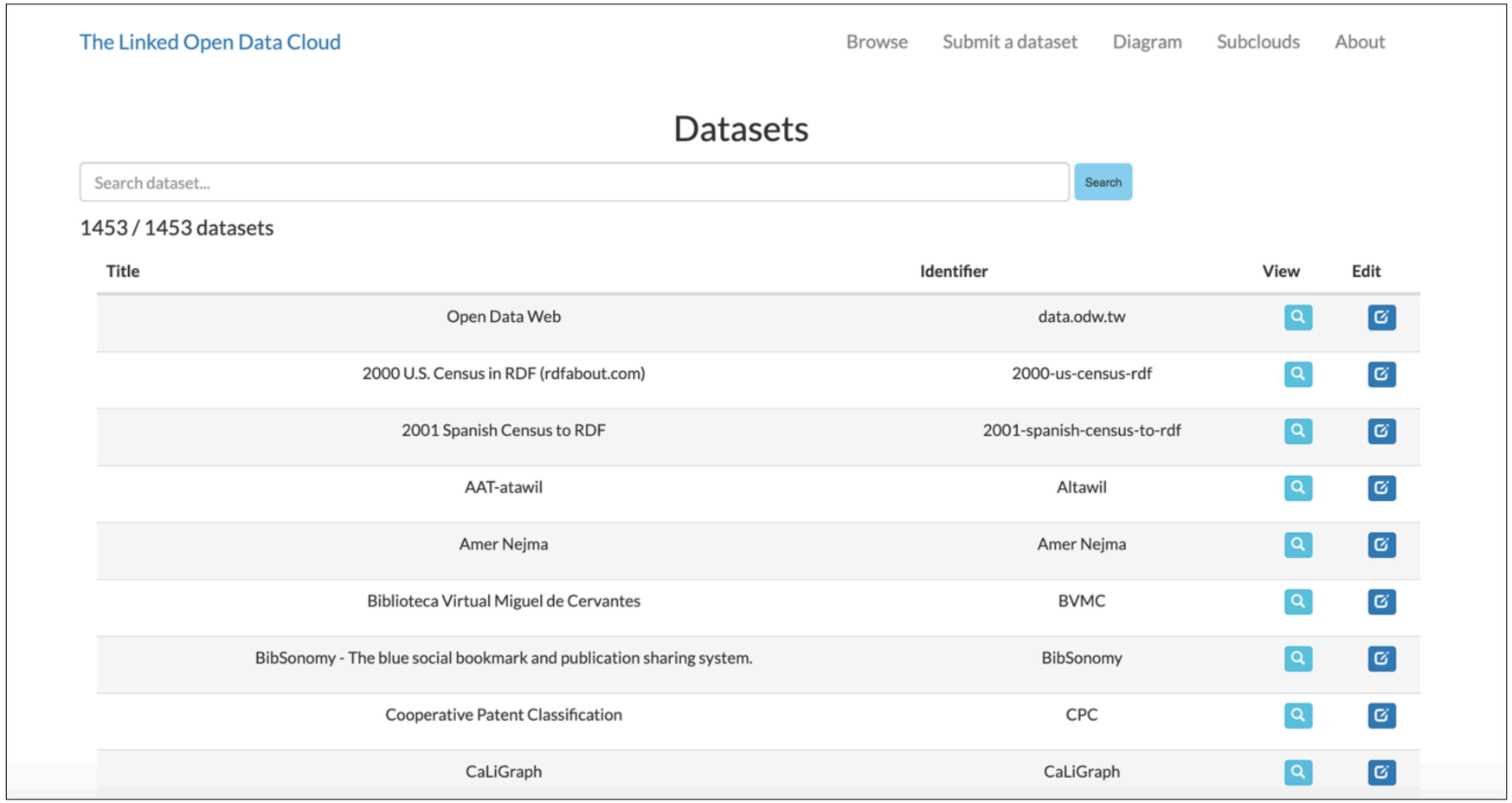
https://www.thinknum.com/datasets

Вопросы в ResearchGate



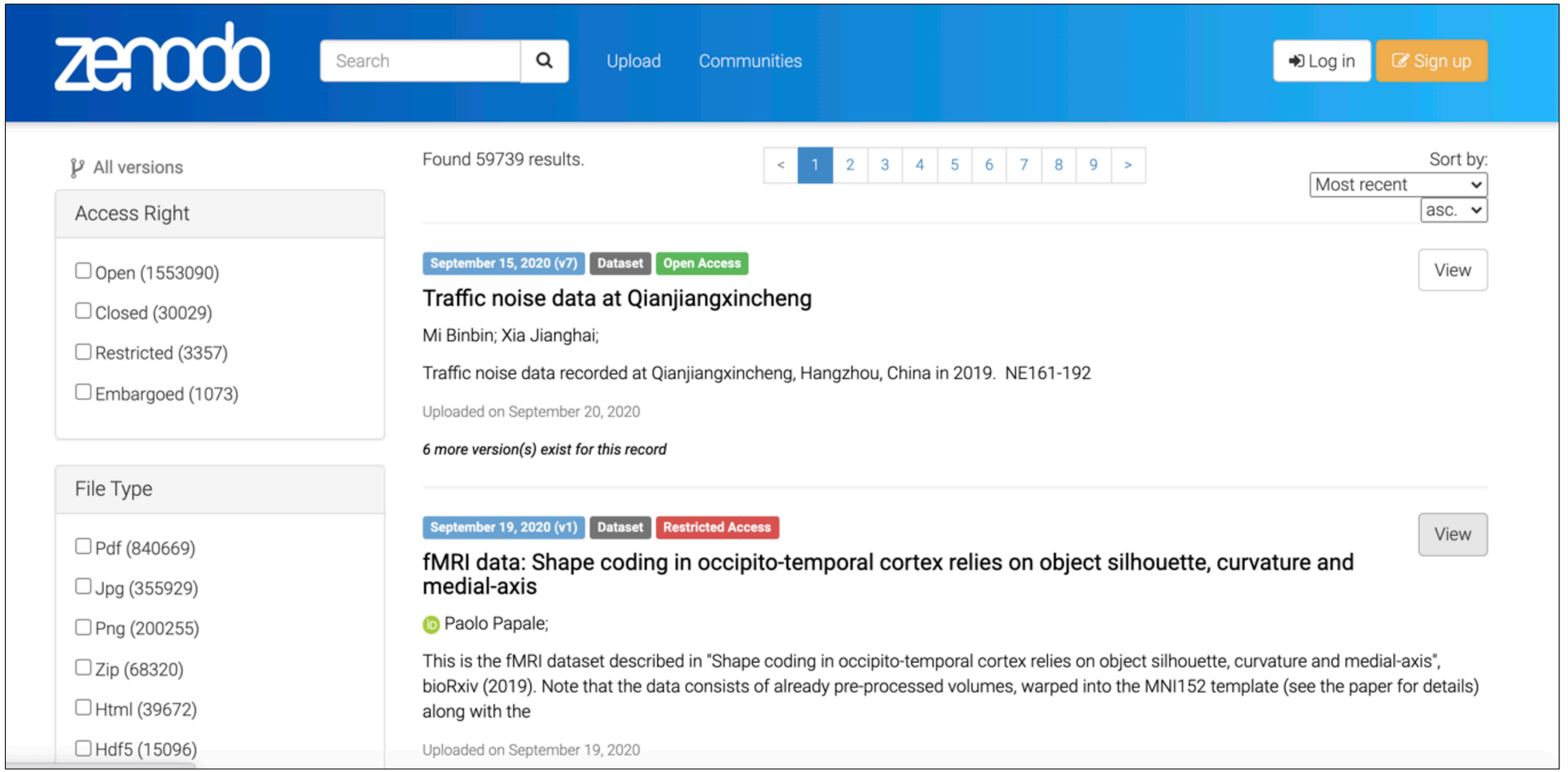
https://www.researchgate.net/

LOD cloud



https://lod-cloud.net/datasets

Платформа Zenodo



https://datadrivenfarming.challenges.org/data-sets-2/

Спасибо за внимание!

Сайт: http://iRadche.ru
Телеграм-канал: @DataPlace

@iRadche

@dadaistka

https://www.facebook.com/**iRadche**

http://www.slideshare.net/**iRadche**