STUDY OF DATA SETS AVAILABLE TO RESEARCHERS

Dr. Sandeep Bhavsar*

* **Librarian,**
Welingkar Institute,
Mumbai, Maharashtra,
India

QR Code

Abstract - *Data Analytics is known as very important field in today's information technology era, all Librarians shall collect the information of available datasets for researchers. The analysis of these datasets will play the important role in knowing the concept/field indepth with respective research. This article will be useful to all researcher in knowing available datasets of their field. Researcher has studied the most of the available datasets on internet and their usefulness to future projects.*

A data set (or dataset) is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. The data set may comprise data for one or more members, corresponding to the number of rows. The term data set may also be used more loosely, to refer to the data in a collection of closely related tables, corresponding to a particular experiment or event. In the open data discipline, data set is the unit to measure the information released in a public open data repository.

Library Data Sets

International Data Set Web Resources

1. Trading Economics
(https://tradingeconomics.com)

TRADING ECONOMICS

Trading Economics provides its users with accurate information for 196 countries including historical data for more than 20 million economic indicators, exchange rates, stock market indexes, government bond yields and commodity prices. Our data is based on official sources, not third party data providers, and our facts are regularly checked for inconsistencies. TradingEconomics.com has received more than 380 million page views from more than 200 countries.

2. United Nations Statistics Division (https://unstats.un.org/home/nso_sites)

United Nations Statistics Division

The United Nations Statistics Division, in its mission to promote the development of national statistical systems, has developed a central repository of country profiles of statistical systems. The country profiles include, among others, a brief history of the cohttps://unstats.un.org/home/nso_sites/untry's statistical system, legal basis, the statistical programme and much more.

3. YouTube labeled Data Set https://research.google.com/youtube8m/index.html)

YouTube 8M

YouTube-8M is a large-scale labeled video dataset that consists of millions of YouTube video IDs, with high-quality machine-generated annotations from a diverse vocabulary of 3,800+ visual entities. It comes with precomputed audio-visual features from billions of frames and audio segments, designed to fit on a single hard disk. This makes it possible to train a strong baseline model on this dataset in less than a day on a single GPU! At the same time, the dataset's scale and diversity can enable deep exploration of complex audio-visual models that can take weeks to train even in a distributed fashion.

4. Reddit Data Sets (https://www.reddit.com/r/datasets/top/?sort=top&t=all)

reddit

Reddit, a popular community discussion site, has a section devoted to sharing interesting data sets. It's called the datasets subreddit, or /r/datasets. The scope of these data sets varies a lot, since they're all user-submitted, but they tend to be very interesting and nuanced.

5. The World Bank (https://data.worldbank.org)

THE WORLD BANK

The World Bank is a global development organization that offers loans and advice to developing countries. The World Bank regularly funds programs in developing countries, then gathers data to monitor the success of these programs.

You can browse World Bank data sets directly, without registering. The data sets have many missing values, and sometimes take several clicks to actually get to data.

Here are some examples:

World Development Indicators — contains country level information on development.

Educational Statistics — data on education by country.

World Bank project costs — data on World Bank projects and their corresponding costs.

The World Bank also provides data sets for rural population, rural poverty gap at national poverty lines etc. (https://data.worldbank.org/topic/agriculture-and-rural-development)

6. Quandl (https://www.quandl.com/search)

Quandl is a repository of economic and financial data. Some of this information is free, but many data sets require purchase. Quandl is useful for building models to predict economic indicators or stock prices. Due to the large amount of available data sets, it's possible to build a complex model that uses many data sets to predict values in another.

7. The UCI Machine Learning Repository (http://mlr.cs.umass.edu/ml/datasets.html)

The UCI Machine Learning Repository is one of the oldest sources of data sets on the web. Although the data sets are user-contributed, and thus have varying levels of documentation and cleanliness, the vast majority are clean and ready for machine learning to be applied. UCI is a great first stop when looking for interesting data sets.

You can download data directly from the UCI Machine Learning repository, without registration. These data sets tend to be fairly small, and don't have a lot of nuance, but are good for machine learning.

8. Google Public Data Sets (https://cloud.google.com/bigquery/public-data)

Much like Amazon, Google also has a cloud hosting service, called Google Cloud Platform. With GCP, you can use a tool called BigQuery to explore large data sets. Google lists all of the data sets on a page. You'll need to sign up for a GCP account, but the first 1TB of queries you make are free.

9. AWS (Amazon Web Services) Public Data sets (https://registry.opendata.aws)

Amazon makes large data sets available on its Amazon Web Services platform. You can download the data and work with it on your own computer, or analyze the data in the cloud using EC2 and Hadoop via EMR. Amazon makes large data sets available on its Amazon Web Services platform. You can download the data and work with it on your own computer, or analyze the data in the cloud using EC2 and Hadoop via EMR.

10. Global Open Data Index (https://index.okfn.org)

The Global Open Data Index (GODI) is the annual global benchmark for publication of open

government data, run by the Open Knowledge Network. Our crowdsourced survey measures the openness of government data according to the Open Definition. By having a tool that is run by civil society, GODI creates valuable insights for government's data publishers to understand where they have data gaps. It also shows how to make data more useable and eventually more impactful. GODI therefore provides important feedback that governments are usually lacking.

11. USA Trade Online (https://usatrade.census.gov)

USA Trade Online is a dynamic data tool that gives users access to current and cumulative U.S. export and import data. With multiple data sets and capabilities, USA Trade Online can assist different types of customers from a wide range of industries and fields.

12. Eurostat (https://ec.europa.eu/eurostat/web/main/home)

Eurostat is the statistical office of the European Union situated in Luxembourg. Its mission is to provide high quality statistics for Europe. Providing the European Union with statistics at European level that enable comparisons between countries and regions is a key task. Democratic societies do not function properly without a solid basis of reliable and objective statistics. On one hand, decision-makers at EU level, in Member States, in local government and in business need statistics to make those decisions. On the other hand, the public and media need statistics for an accurate picture of contemporary society and to evaluate the performance of politicians and others. Of course, national statistics are still important for national purposes in Member States whereas EU statistics are essential for decisions and evaluation at European level.

13. Academic Torrents (http://academictorrents.com/browse.php?cat=6)

Academic Torrents is a new site that is geared around sharing the data sets from scientific papers. You can browse the data sets directly on the site. Since it's a torrent site, all of the data sets can be immediately downloaded, but you'll need a Bittorrent client. Deluge is a good free option.

14. data.world (https://data.world)

data.world describes itself at 'the social network for data people', but could be more correctly describe as 'GitHub for data'. It's a

place where you can search for, copy, analyze, and download data sets. In addition, you can upload your data to data.world and use it to collaborate with others.

In a relatively short time it has become one of the 'go to' places to acquire data, with lots of user contributed data sets as well as fantastic data sets through data.world's partnerships with various organizations including a large amount of data from the US Federal Government.

15. Kaggle (https://www.kaggle.com/datasets)

Kaggle is a data science community that hosts machine learning competitions. There are a variety of externally-contributed interesting data sets on the site. Kaggle has both live and historical competitions. You can download data for either, but you have to sign up for Kaggle and accept the terms of service for the competition.

You can download data from Kaggle by entering a competition. Each competition has its own associated data set. There are also user-contributed data sets found in the new Kaggle Data sets offering.

16. Socrata OpenData (https://opendata.socrata.com)

Socrata OpenData is a portal that contains multiple clean data sets that can be explored in the browser or downloaded to visualize. A significant portion of the data is from US government sources. You can explore and download data from OpenData without registering. You can also use visualization and exploration tools to explore the data in the browser.

17. FiveThirtyEight (https://data.fivethirtyeight.com)

FiveThirtyEight is an incredibly popular interactive news and sports site started by Nate Silver. They write interesting data-driven articles, like "How Popular is Donald Trump" and "2018 MLB Predictions". FiveThirtyEight makes the data sets used in its articles available online on Github.

18. KD Nuggets (https://www.kdnuggets.com/datasets/index.html)

KD stands for Knowledge Discovery. KDnuggets is a leading site on Business Analytics, Big Data, Data Mining, Data Science, and Machine Learning. One will find data sets for Data Mining and Data Science.

19. BuzzFeed
(https://github.com/BuzzFeedNews)

BuzzFeed makes the data sets used in its articles available on Github.

Here are some examples:

- Federal Surveillance Planes — contains data on planes used for domestic surveillance.
- Zika Virus — data about the geography of the Zika virus outbreak.
- Firearm background checks — data on background checks of people attempting to buy firearms.

National Data Set Web Resources

1. Directorate General of Commercial Intelligence and Statistics (http://www.dgciskol.gov.in)

DIRECTORATE GENERAL OF COMMERCIAL INTELLIGENCE AND STATISTICS
MINISTRY OF COMMERCE AND INDUSTRY, GOVERNMENT OF INDIA

Foreign Trade Data Dissemination Portal of DGCIS

Foreign trade data are captured on the basis of flows/movements of goods across the custom frontiers of India. The value of import is based on CIF, i.e. Cost, Insurance and Freight whereas the value of export is on FOB (Free on Board).

The Foreign Trade Statistics are released in three phases :

(1) Monthly Quick Estimates are released by fifteenth of the following month , (2) Principal Commodity level data is released within 30 days, and (3) Detailed 8-digit commodity level data are released within 60 days.

All these data are provisional and are dynamically revised till it is finalised on the basis of late receipt data.

Certain types of foreign trade such as foreign trade in treasure, currency notes &coins in circulation, articles warehoused under bond, foreign trade of neighbouring countries passing in transit through India (e.g., Bhutan transit trade) etc. are out of coverage of Foreign Trade Statistics. Other types of trade namely, Transshipment Trade, Passengers Baggages, Defence Goods, Diplomatic Goods, etc. are also excluded.

2. Reserve Bank of India
(https://www.rbi.org.in/Scripts/Statistics.aspx)

भारतीय रिज़र्व बैंक
Reserve Bank of India
India's Central Bank

RBI provides data on various aspects of Indian economy, banking and finance. While the current data defined as data for the past one year is available at the links provided on the website. This includes several metrics on money market operations, balance of payments, use of banking and several products.

3. ICSSR Data Service
(http://www.icssrdataservice.in)

The "ICSSR Data Service" is culmination of signing of Memorandum of Understanding (MoU) between Indian Council of Social Science Research (ICSSR) and Ministry of Statistics and Programme Implementation (MoSPI).

The platform provides single point access to a wide range of primary datasets including datasets generated by large-scale government surveys i.e. ASI and NSS that provides unit level data as well as qualitative studies. The ICSSR Data Service as a policy, promotes data sharing to encourage the reuse of data and provide information on developing and generating social science research data & its management.

The ICSSR Data Service extracts and transforms the data from raw datasets before uploading it in the data repository with necessary documentation for the benefit of researcher. Besides hosting unit-level datasets, the ICSSR Data Service also provides access to secondary datasets derived from the unit-level datasets using selected parameters. Training materials and guidance to meet the needs of data users, owners and creators is also offered through this platform. Users can explore collection of datasets accompanied with user guides and supporting materials using search interface of the ICSSR Data Service. All datasets documentation and resources are freely available through this platform.

4. National Data Repository (NDR)
(https://www.ndrdgh.gov.in/NDR)

Government of India in envisaging on an ambitious project to set up a National Knowledge Hub (NKH) alternatively called as National Knowledge Centre (NKC) in E&P area in coming few years.

NDR is a turnkey project on Build, Populate and Operate basis. Below are the data classes of NDR:

- Seismic Data
- Well & Log Data
- Spatial Data
- Other G&G data like Drilling, Reservoir, Production, Geological, Gravity & Magnetic
- Reports and Documents

5. Department of Commerce Analytics
(http://dashboard-commerce.gov.in)

Dashboard of Ministry of Commerce and Industry allows you to see the complete picture of imports, exports, and balance of trade of India in a graphical form. The import view is the first tab

on the dashboard showing how India's imports have changed over a period of time. The top commodities and top ports section highlights the top 5 commodities and ports as well as the ones at the bottom. A clickable world map shows the country-wise India's imports from around the globe. In the similar way, exports data is available on the exports tab.

6. Export Import Data Bank (http://commerce-app.gov.in/eidb/Default.asp)

Export Import Data Bank

India's Imports/Exports include re-imports/re-exports also.

Imports/Exports from unspecified country includes -

(a) Trade transactions where country of origin/consignment/destination is not specified or invalid country codes have been assigned in the customs declaration.

(b) All re-imported/re-exported transactions which fulfill condition (a) above.

7. Open Government Data (OGD) Platform India (https://data.gov.in)

Open Government Data (OGD) Platform India - data.gov.in - is a platform for supporting Open Data initiative of Government of India. The portal is intended to be used by Government of India Ministries/ Departments their organizations to publish datasets, documents, services, tools and applications collected by them for public use. It intends to increase transparency in the functioning of Government and also open avenues for many more innovative uses of Government Data to give different perspective.

This portal also gives

- *Open Data on Agriculture.* (https://data.gov.in/sectors/Agriculture-9212)
- *Demographic Details* (https://data.gov.in)
- *Data on Education* (https://data.gov.in/sectors/Education-9264)
- *Data on Infrastructure*(https://data.gov.in/catalogs/sector/Infrastructure-9345)
- *Energy & Power* (https://data.gov.in/sectors/Power%20and%20Energy-9361)
- *Rural Development Dataset* (https://data.gov.in/sectors/Rural-9364)
- *Telecommunication* (https://data.gov.in)
- *Water and Sanitation* (https://data.gov.in/sectors/Water%20and%20Sanitation-9249)
- *Healthcare* (https://data.gov.in/sectors/Health%20and%20Family%20welfare-9312)
- *Tourism & Hospitality* (https://data.gov.in/sectors/Travel%20and%20Tourism-9389)

8. National Data Bank
(http://www.mospi.gov.in/national-data-bank)

Government of India
Ministry of Statistics and
Programme Implementation

National Data Bank : The National Data Bank of Socio-Religious categories is developed with a view to provide users access to all data, pertaining to various aspects of socio-economic life of population falling in different social/religious categories, from a single window.

9. Chars74K
(http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k)

Character recognition is a classic pattern recognition problem for which researchers have worked since the early days of computer vision. With today's omnipresence of cameras, the applications of automatic character recognition are broader than ever. For Latin script, this is largely considered a solved problem in constrained situations, such as images of scanned documents containing common character fonts and uniform background. However, images obtained with popular cameras and hand held devices still pose a formidable challenge for character recognition. The challenging aspects of this problem are evident in this dataset.

In this dataset, symbols used in both English and Kannada are available.

In the English language, Latin script (excluding accents) and Hindu-Arabic numerals are used. For simplicity we call this the "English" characters set.

This gives a total of over 74K images (which explains the name of the dataset).

Data Sets - Agriculture

Government of India Data Sets on Agriculture

1. Variety-wise Daily Market Prices Data of Lemon, Apple, Carrot, Pine appleetc. of various fruits and vegetables
(https://data.gov.in/catalog/variety-wise-daily-market-prices-data-apple?filters%5Bfield_catalog_reference%5D=92284&format=json&offset=0&limit=6&sort%5Bcreated%5D=desc)

The data refers to State-wise, market-wise, variety-wise prices of fruits and vegetables. It has the daily wholesale maximum price, minimum price and modal price. These dataset is generated through the AGMARKNET Portal (http://agmarknet.nic.in) which disseminates daily market information of various commodities.

2. Knoema (https://knoema.com/atlas/India/topics/Agriculture

knoema

Knoema is the most comprehensive source of global decision-making data in the world. Their tools allow individuals and organizations to discover, visualize, model, and present their data and the world's data to facilitate better decisions and better outcomes.

Knoema is a search engine for data seamlessly connecting public and private sources and making data discoverable and accessible to information workers. Knoema does for data what Google did for websites and the Internet overall. It makes it trivial to find data when you need it and make a story out of it. Their public data library features more than 2.4B time series from thousands of sources.

This portal also gives

- *Country wise Demographic Details*
- *Country wise data on Education*
- *Country wise data on Infrastructure*
- *Country wide data on Energy & Power*
- *Telecommunication*
- *Healthcare & Pharma*

Data Sets Demographics

1. Census Digital Library(http://www.censusindia.gov.in/DigitalLibrary/Archive_home.aspx)

Census Digital Library makes available all Census Tables published from 1991 to 2011 Censuses, Census Reports and other digital files for free download in soft copy format.

- 2184 Books and Reports (in Pdf).
- 54016 Census Tables from 1991 Census to 2011 Census (in Excel and Csv).
- 908 Photographs of Census Publicity (in Jpg and Bmp).
- 44 Census Maps (in Pdf).
- 68 Publicity Audio Clips (in Mp3 and Wav).
- 80 Power Point Presentations (in Pptx).

Data Sets - Machine Learning and AI

Image Data Sets

1. ImageNet (http://www.image-net.org)

IMAGENET ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Currently we have an average of over five hundred images per node. We hope ImageNet will become a useful resource for

researchers, educators, students and all of you who share our passion for pictures.

2. Open Images Dataset (https://storage.googleapis.com/openimages/web/index.html)

Open Images V5

Open Images is a dataset of ~9M images annotated with image-level labels, object bounding boxes, object segmentation masks, and visual relationships. It contains a total of 16M bounding boxes for 600 object classes on 1.9M images, making it the largest existing dataset with object location annotations. The boxes have been largely manually drawn by professional annotators to ensure accuracy and consistency. The images are very diverse and often contain complex scenes with several objects (8.3 per image on average). Open Images also offers visual relationship annotations, indicating pairs of objects in particular relations (e.g. "woman playing guitar", "beer on table"). In total it has 329 relationship triplets with 391,073 samples. In V5 we added segmentation masks for 2.8M object instances in 350 classes. Segmentation masks mark the outline of objects, which characterizes their spatial extent to a much higher level of detail. Finally, the dataset is annotated with 36.5M image-level labels spanning 19,969 classes.

3. MNIST (http://yann.lecun.com/exdb/mnist)

MNIST is one of the most popular deep learning datasets out there. It's a dataset of handwritten digits and contains a training set of 60,000 examples and a test set of 10,000 examples. It's a good database for trying learning techniques and deep recognition patterns on real-world data while spending minimum time and effort in data preprocessing.

4. Fashion-MNIST (https://github.com/zalandoresearch/fashion-mnist)

Fashion-MNIST consists of 60,000 training images and 10,000 test images. It is a MNIST-like fashion product database. The developers believe MNIST has been overused so they created this as a direct replacement for that dataset. Each image is in greyscale and associated with a label from 10 classes.

5. CIFAR-10 (http://www.cs.toronto.edu/~kriz/cifar.html)

The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.

The dataset is divided into five training batches and one test batch, each with 10000 images. The test batch contains exactly 1000 randomly-selected images from each class. The training batches contain the remaining images in random order, but some training batches may contain more images from one class than another.

Between them, the training batches contain exactly 5000 images from each class.

6. VisualQA (https:/visualqa.org)

VQA is a new dataset containing open-ended questions about images. These questions require an understanding of vision, language and commonsense knowledge to answer.

- 265,016 images (COCO and abstract scenes)
- At least 3 questions (5.4 questions on average) per image
- 10 ground truth answers per question
- 3 plausible (but likely incorrect) answers per question
- Automatic evaluation metric

7. MS-COCO (http://cocodataset.org/#home)

COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

Object segmentation

- Recognition in context
- Superpixel stuff segmentation
- 330K images (>200K labeled)
- 1.5 million object instances
- 80 object categories
- 91 stuff categories
- 5 captions per image
- 250,000 people with keypoints

Natural Language Processing Data Sets

1. The Wikipedia Corpus (https://nlp.cs.nyu.edu/wikipedia-data)

The Wikipedia Corpus dataset is a collection of the full text on Wikipedia. It contains almost 1.9 billion words from more than 4 million articles. What makes this a powerful NLP dataset is that you search by word, phrase or part of a paragraph itself.

Size: 20 MB

Number of Records: 4,400,000 articles containing 1.9 billion words

2. Sentiment140 (http://help.sentiment140.com/for-students/)

Sentiment140 is a dataset that can be used for sentiment analysis. A popular dataset, it is perfect to start off your NLP journey. Emotions have been pre-removed from the data. The final dataset has the below 6 features:

- polarity of the tweet
- id of the tweet
- date of the tweet
- the query
- username of the tweeter
- text of the tweet

Size: 80 MB (Compressed)

Number of Records: 1,60,000 tweets

3. Yelp Reviews (https://www.yelp.com/dataset)

Yelp Reviews is an open dataset released by Yelp for learning purposes. It consists of millions of user reviews, businesses attributes and over 200,000 pictures from multiple metropolitan areas. This is a very commonly used dataset for NLP challenges globally.

Size: 2.66 GB JSON, 2.9 GB SQL and 7.5 GB Photos (all compressed)

Number of Records: 5,200,000 reviews, 174,000 business attributes, 200,000 pictures and 11 metropolitan areas

4. Machine Translation of Various Languages (http://statmt.org/wmt18/index.html)

Machine Translation of Various Languages dataset consists of training data for four European languages. The task here is to improve the current translation methods. You can participate in any of the following language pairs:

- English-Chinese and Chinese-English
- English-Czech and Czech-English
- English-Estonian and Estonian-English
- English-Finnish and Finnish-English
- English-German and German-English
- English-Kazakh and Kazakh-English
- English-Russian and Russian-English
- English-Turkish and Turkish-English

Size: ~15 GB

Number of Records: ~30,000,000 sentences and their translations

5. IMDB Reviews (https://www.imdb.com/interfaces/)

IMDB Reviews is a dream dataset for movie lovers. It is meant for binary sentiment classification and has far more data than any previous datasets in this field. Apart from the training and test review examples, there is further unlabeled data for use as well. Raw text and preprocessed bag of words formats have also been included.

Size: 80 MB

Number of Records: 25,000 highly polar movie reviews for training, and 25,000 for testing

6. Twenty Newsgroups (https://www.imdb.com/interfaces/)

Twenty Newsgroups dataset, as the name suggests, contains information about newsgroups. To curate this dataset, 1000 Usenet articles were taken from 20 different newsgroups. The articles have typical features like subject lines, signatures, and quotes.

Size: 20 MB

Number of Records: 20,000 messages taken from 20 newsgroups

Audio/Speech Data Sets

1. Age Detection of Indian Actors (https://datahack.analyticsvidhya.com/contest/practice-problem-age-detection/)

This is a fascinating challenge for any deep learning enthusiast. The dataset contains thousands of images of Indian actors and your task is to identify their age. All the images are manually selected and cropped from the video frames resulting in a high degree of variability in terms of scale, pose, expression, illumination, age, resolution, occlusion, and makeup.

Size: 48 MB (Compressed)

Number of Records: 19,906 images in the training set and 6636 in the test set

2. LibriSpeech (http://www.openslr.org/12/)

OpenSLR LibriSpeech dataset is a large-scale corpus of around 1000 hours of English speech. The data has been sourced from audiobooks from the LibriVox project. It has been segmented and aligned properly. If you're looking for a starting point, check out already prepared Acoustic models that are trained on this data set at kaldi-asr.org and language models, suitable for evaluation.

Size: ~60 GB

Number of Records: 1000 hours of speech

3. Twitter Sentiment Analysis (https://datahack.analyticsvidhya.com/contest/practice-problem-twitter-sentiment-analysis/)

Hate Speech in the form of racism and sexism has become a nuisance on twitter and it is important to segregate these sort of tweets from the rest. In this Practice problem, we provide Twitter data that has both normal and hate tweets. Your task as a Data Scientist is to identify the tweets which are hate tweets and which are not.

Size: 3 MB

Number of Records: 31,962 tweets

4. Million Song Dataset (http://millionsongdataset.com/)

The Million Song Dataset is a freely-available collection of audio features and metadata for a million contemporary popular music tracks. Its purposes are:

- To encourage research on algorithms that scale to commercial sizes
- To provide a reference dataset for evaluating research
- As a shortcut alternative to creating a large dataset with APIs (e.g. The Echo Nest's)
- To help new researchers get started in the MIR field

The core of the dataset is the feature analysis and metadata for one million songs. The dataset does not include any audio, only the derived features. The sample audio can be fetched from services like 7digital, using code provided by Columbia University.

Size: 280 GB

Number of Records: PS – its a million songs!

5. Free Music Archive (FMA) (https://github.com/mdeff/fma)

FMA is a dataset for music analysis. The dataset consists of full-length and HQ audio, pre-computed features, and track and user-level metadata. It an an open dataset created for evaluating several tasks in MIR. Below is the list of csv files the dataset has along with what they include:

- tracks.csv: per track metadata such as ID, title, artist, genres, tags and play counts, for all 106,574 tracks.
- genres.csv: all 163 genre IDs with their name and parent (used to infer the genre hierarchy and top-level genres).
- features.csv: common features extracted with librosa.
- echonest.csv: audio features provided by Echonest (now Spotify) for a subset of 13,129 tracks.

Size: ~1000 GB

Number of Records: ~100,000 tracks

6. Free Spoken Digit Dataset (https://github.com/Jakobovski/free-spoken-digit-dataset)

Free Spoken Digit Dataset was created to solve the task of identifying spoken digits in audio samples. It's an open dataset so the hope is that it will keep growing as people keep contributing more samples. Currently, it contains the below characteristics:

- 3 speakers
- 1,500 recordings (50 of each digit per speaker)
- English pronunciations

Size: 10 MB

Number of Records: 1,500 audio samples.