Stony Brook University

LOVE DATA WEEK

Stony Brook University | Libraries

Finding Open Data For Your Projects

Ahmad R. Pratama, Ph.D.
Data Literacies Lead

Stony Brook University Libraries

February 10, 2025

Outline

1. What is Open Data & Why It Matters
2. Quick Poll: Who’s in the Room?
3. Key Tools & Resources
4. Tips for Evaluating Data
5. Q&A and Next Steps

What is

Open Data?

Data that can be freely used, re-used, and redistributed by anyone.

Features:

1. Usually accessible online
2. Has an open license
3. May or may not require attribution

Why Use Open Data?

Academic & Research Benefits

- Increases reproducibility
- Speeds up discovery

Professional & Industry Benefits

- Supports data-driven decision-making
- Reduces barriers to entry for new projects

Community & Public Good

- Encourages civic engagement
- Fosters trust and accountability

Reliable Data Sources

Source Credibility
Data from established organizations or institutions (e.g., government agencies, universities).

Data Accuracy
Verified through peer review or by cross-referencing with other sources.

Consistency
Reliable sources update their data regularly and provide historical records.

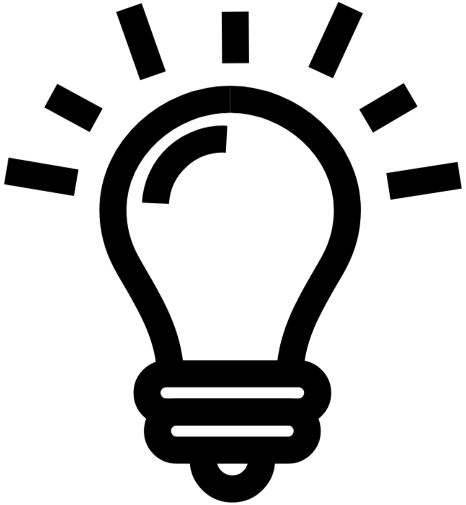
Transparency
Clearly show how the data was collected and what methodology was used.

Quick Poll

Who’s Here?

1. What is your main role?

2. Which type(s) of data are you most interested in?



Key Tools and

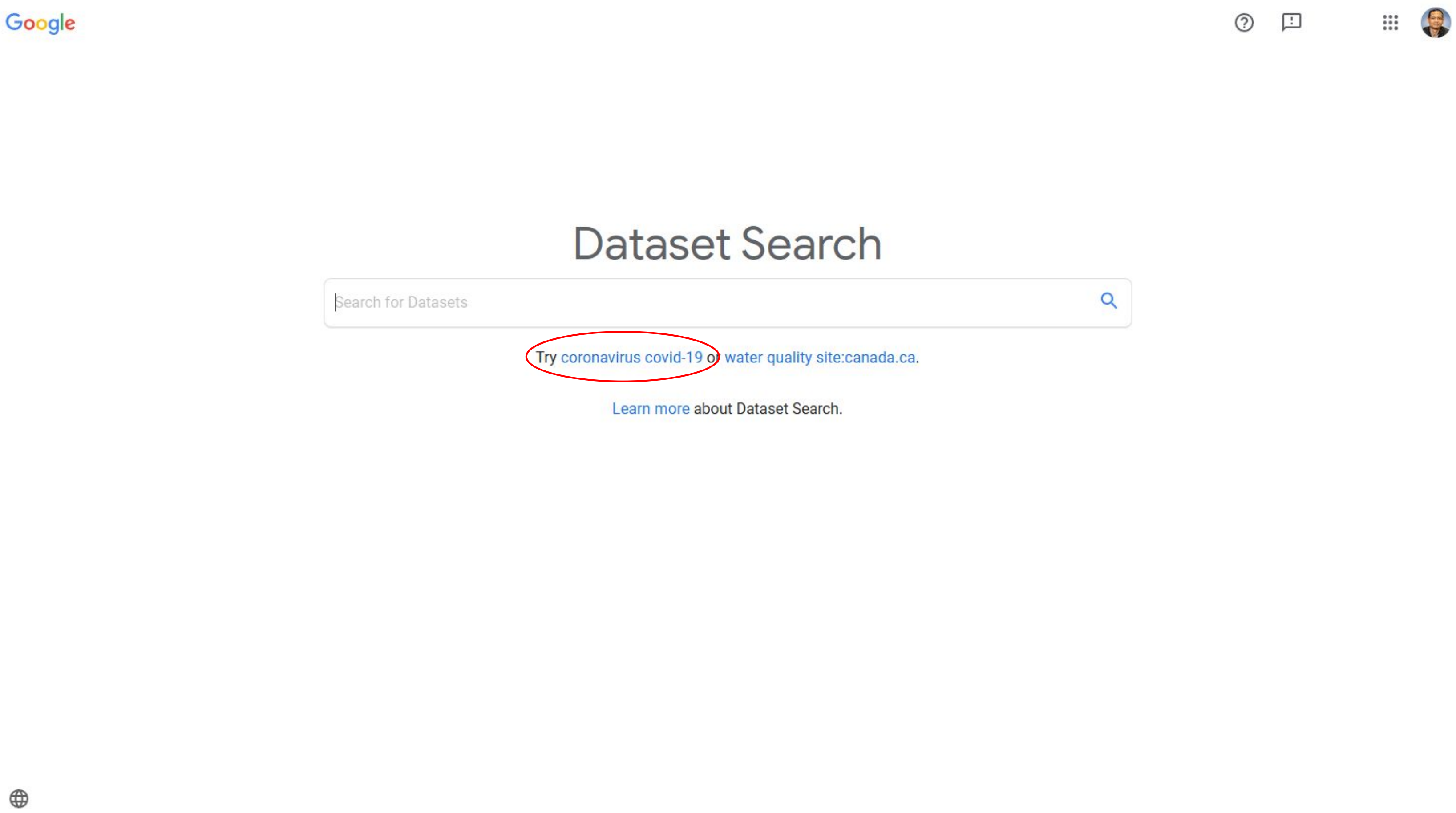
Resources

1. **Google Dataset Search**
   A broad, user-friendly starting point.

2. **Government Repositories**
   e.g., Data.gov (U.S.), EU Data Portal (EU).

3. **Institutional Repositories**
   University/College repositories or library services.

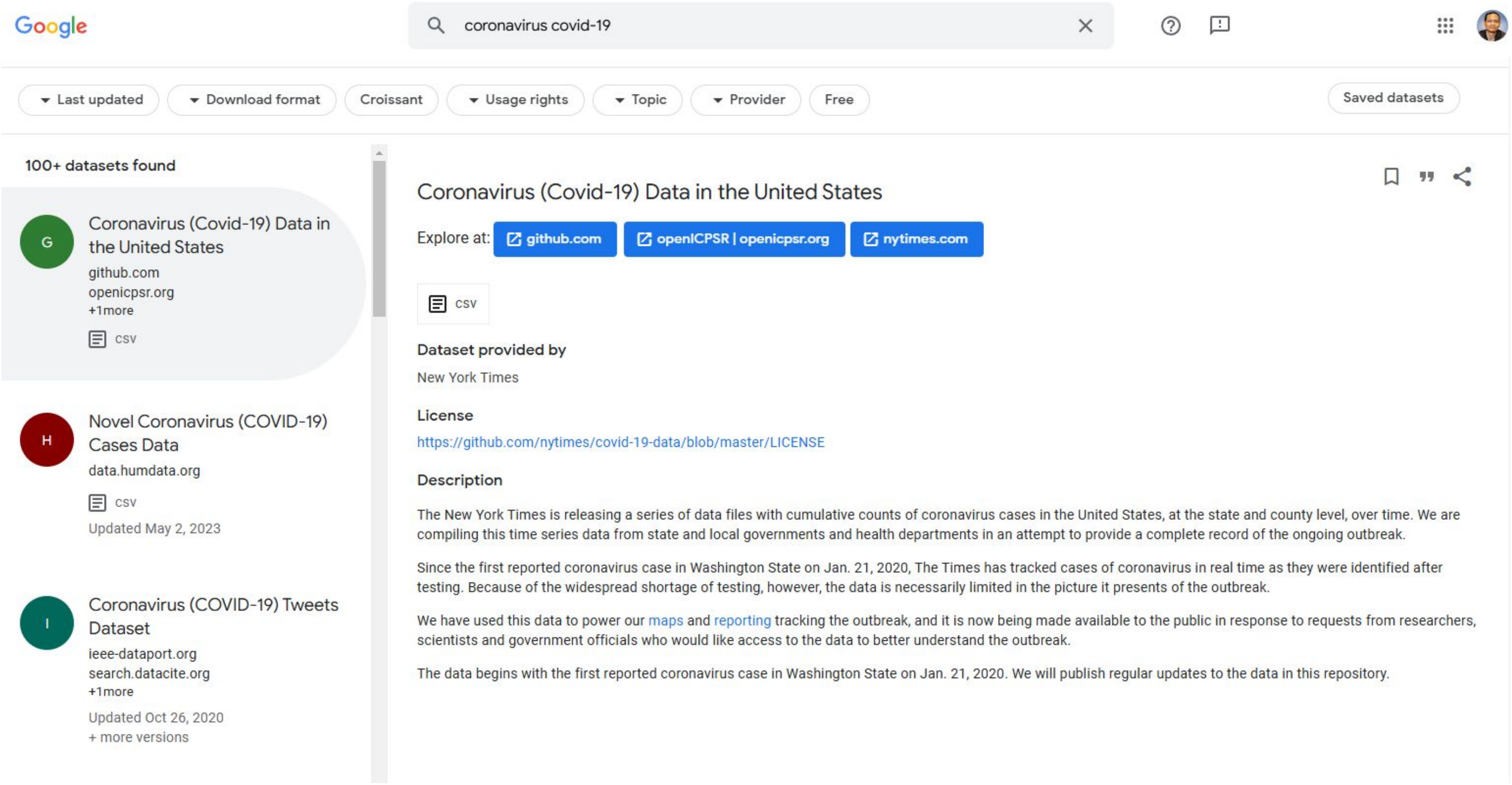
4. **Third-Party Repositories**
   e.g., Kaggle, Figshare, Zenodo.



Dataset Search

Try coronavirus covid-19 or water quality site:canada.ca.

Learn more about Dataset Search.



coronavirus covid-19

100+ datasets found

Coronavirus (Covid-19) Data in the United States
github.com
openicpsr.org
+1more
csv

Novel Coronavirus (COVID-19) Cases Data
data.humdata.org
csv
Updated May 2, 2023

Coronavirus (COVID-19) Tweets Dataset
ieee-dataport.org
search.datacite.org
+1more
Updated Oct 26, 2020
+ more versions

Coronavirus (Covid-19) Data in the United States

csv

Dataset provided by

New York Times

License

https://github.com/nytimes/covid-19-data/blob/master/LICENSE

Description

The New York Times is releasing a series of data files with cumulative counts of coronavirus cases in the United States, at the state and county level, over time. We are compiling this time series data from state and local governments and health departments in an attempt to provide a complete record of the ongoing outbreak.

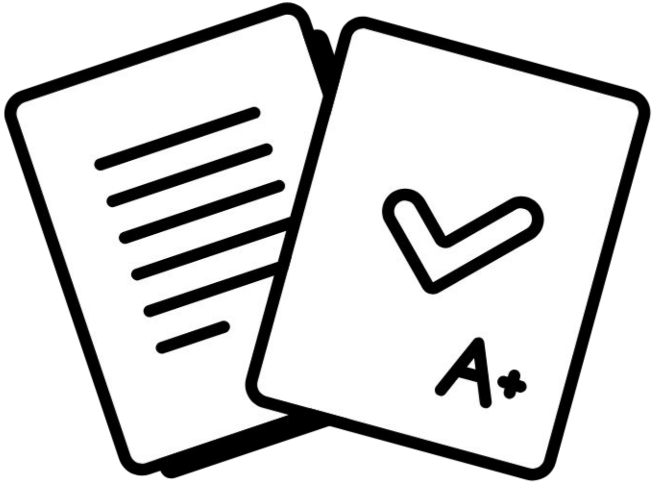
Since the first reported coronavirus case in Washington State on Jan. 21, 2020, The Times has tracked cases of coronavirus in real time as they were identified after testing. Because of the widespread shortage of testing, however, the data is necessarily limited in the picture it presents of the outbreak.

We have used this data to power our maps and reporting tracking the outbreak, and it is now being made available to the public in response to requests from researchers, scientists and government officials who would like access to the data to better understand the outbreak.

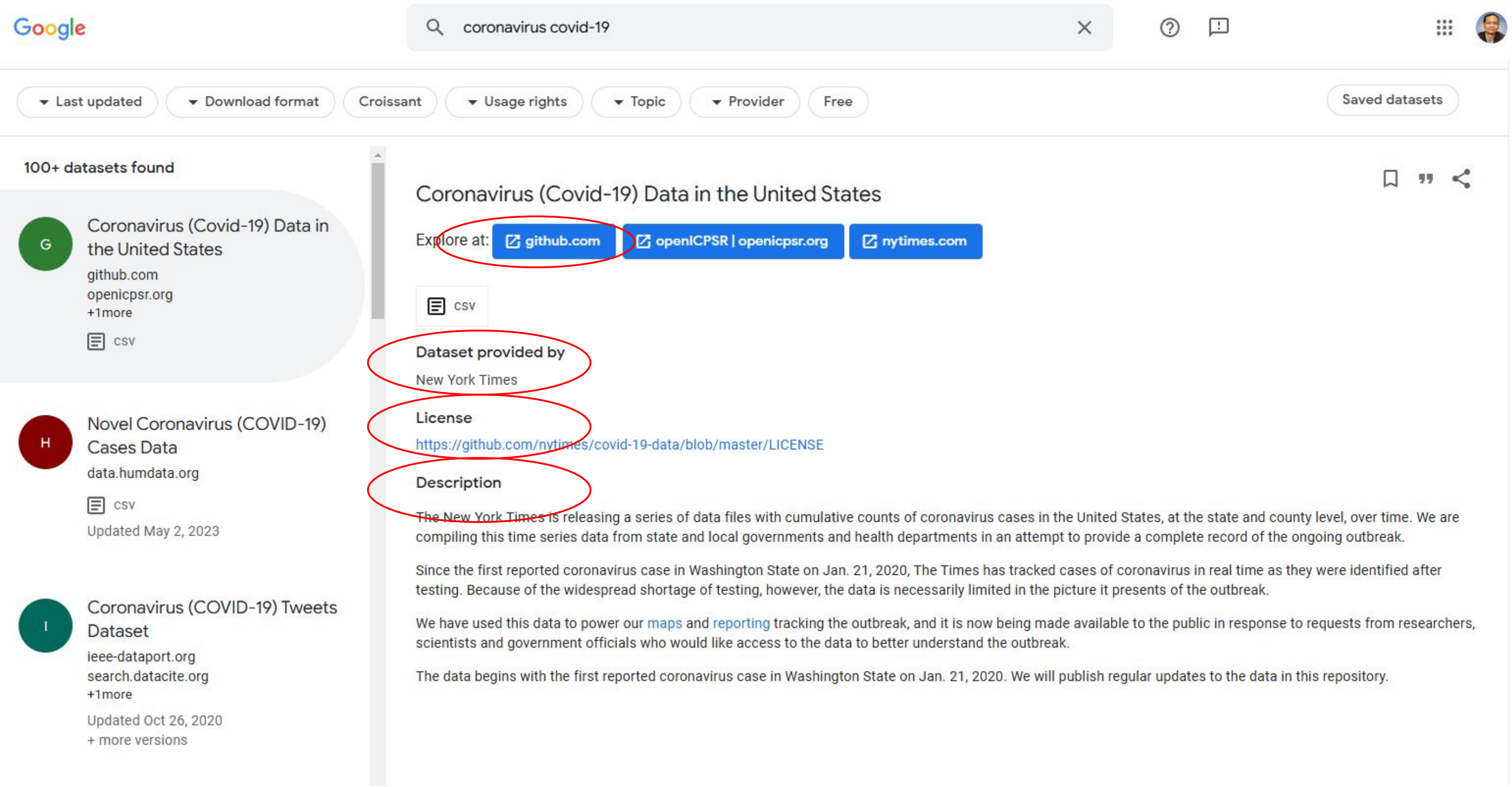
The data begins with the first reported coronavirus case in Washington State on Jan. 21, 2020. We will publish regular updates to the data in this repository.

Open Data: Tips for

Evaluation



1. **Check Source Reliability**
   - Source credibility
   - Data accuracy
   - Consistency (update frequency and recency)
   - Transparency
2. **Metadata & Documentation**
   - Clear column headers, definitions, measurement units
3. **Licensing & Permissions**
   - Creative Commons or Open Data Commons licenses
   - Always note if attribution is required
4. **Format & Ease of Use**
   - CSV, JSON, Excel, etc.
   - Potential for easy manipulation in your workflow



100+ datasets found

Coronavirus (Covid-19) Data in the United States
github.com
openicpsr.org
+1more
csv

Novel Coronavirus (COVID-19) Cases Data
data.humdata.org
csv
Updated May 2, 2023

Coronavirus (COVID-19) Tweets Dataset
ieee-dataport.org
search.datacite.org
+1more
Updated Oct 26, 2020
+ more versions

Coronavirus (Covid-19) Data in the United States

Explore at: github.com openICPSR | openicpsr.org nytimes.com

csv

Dataset provided by

New York Times

License

https://github.com/nytimes/covid-19-data/blob/master/LICENSE

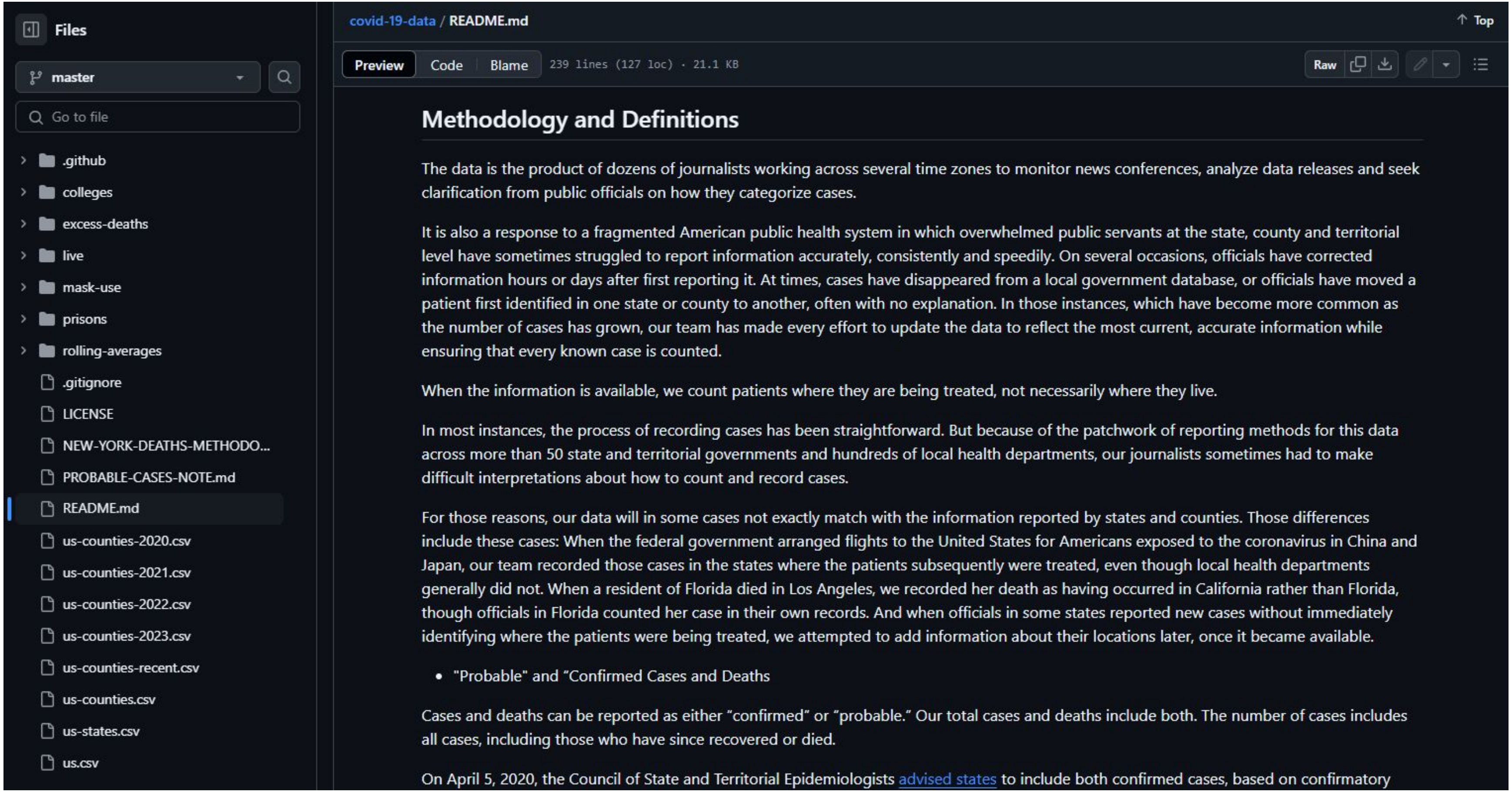
Description

The New York Times is releasing a series of data files with cumulative counts of coronavirus cases in the United States, at the state and county level, over time. We are compiling this time series data from state and local governments and health departments in an attempt to provide a complete record of the ongoing outbreak.

Since the first reported coronavirus case in Washington State on Jan. 21, 2020, The Times has tracked cases of coronavirus in real time as they were identified after testing. Because of the widespread shortage of testing, however, the data is necessarily limited in the picture it presents of the outbreak.

We have used this data to power our maps and reporting tracking the outbreak, and it is now being made available to the public in response to requests from researchers, scientists and government officials who would like access to the data to better understand the outbreak.

The data begins with the first reported coronavirus case in Washington State on Jan. 21, 2020. We will publish regular updates to the data in this repository.



Methodology and Definitions

The data is the product of dozens of journalists working across several time zones to monitor news conferences, analyze data releases and seek clarification from public officials on how they categorize cases.

It is also a response to a fragmented American public health system in which overwhelmed public servants at the state, county and territorial level have sometimes struggled to report information accurately, consistently and speedily. On several occasions, officials have corrected information hours or days after first reporting it. At times, cases have disappeared from a local government database, or officials have moved a patient first identified in one state or county to another, often with no explanation. In those instances, which have become more common as the number of cases has grown, our team has made every effort to update the data to reflect the most current, accurate information while ensuring that every known case is counted.

When the information is available, we count patients where they are being treated, not necessarily where they live.

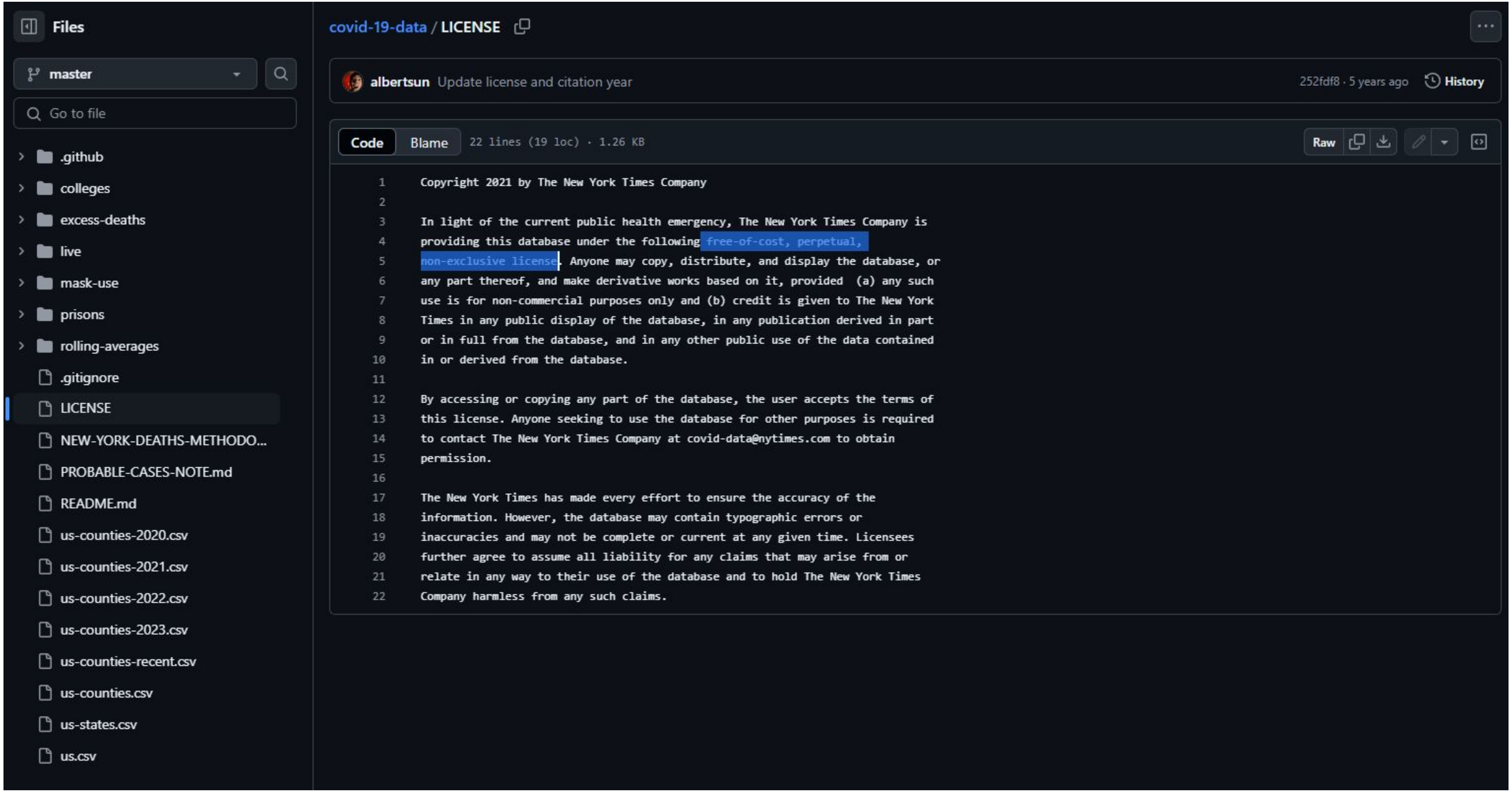
In most instances, the process of recording cases has been straightforward. But because of the patchwork of reporting methods for this data across more than 50 state and territorial governments and hundreds of local health departments, our journalists sometimes had to make difficult interpretations about how to count and record cases.

For those reasons, our data will in some cases not exactly match with the information reported by states and counties. Those differences include these cases: When the federal government arranged flights to the United States for Americans exposed to the coronavirus in China and Japan, our team recorded those cases in the states where the patients subsequently were treated, even though local health departments generally did not. When a resident of Florida died in Los Angeles, we recorded her death as having occurred in California rather than Florida, though officials in Florida counted her case in their own records. And when officials in some states reported new cases without immediately identifying where the patients were being treated, we attempted to add information about their locations later, once it became available.

- "Probable" and “Confirmed Cases and Deaths

Cases and deaths can be reported as either “confirmed” or “probable.” Our total cases and deaths include both. The number of cases includes all cases, including those who have since recovered or died.

On April 5, 2020, the Council of State and Territorial Epidemiologists advised states to include both confirmed cases, based on confirmatory



albertsun Update license and citation year

252fdf8 · 5 years ago

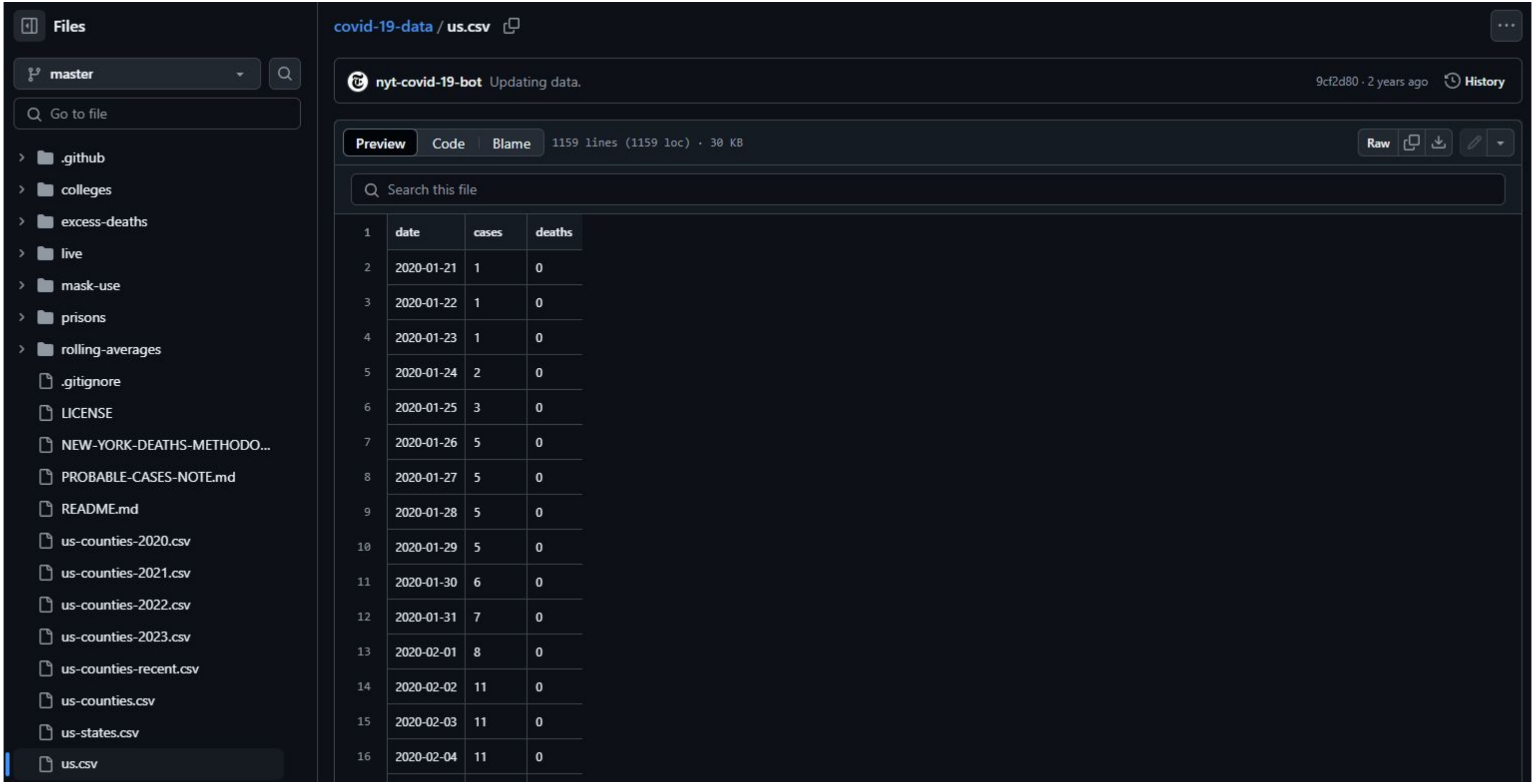
22 lines (19 loc) · 1.26 KB

Copyright 2021 by The New York Times Company

In light of the current public health emergency, The New York Times Company is providing this database under the following free-of-cost, perpetual, non-exclusive license. Anyone may copy, distribute, and display the database, or any part thereof, and make derivative works based on it, provided (a) any such use is for non-commercial purposes only and (b) credit is given to The New York Times in any public display of the database, in any publication derived in part or in full from the database, and in any other public use of the data contained in or derived from the database.

By accessing or copying any part of the database, the user accepts the terms of this license. Anyone seeking to use the database for other purposes is required to contact The New York Times Company at covid-data@nytimes.com to obtain permission.

The New York Times has made every effort to ensure the accuracy of the information. However, the database may contain typographic errors or inaccuracies and may not be complete or current at any given time. Licensees further agree to assume all liability for any claims that may arise from or relate in any way to their use of the database and to hold The New York Times Company harmless from any such claims.



date	cases	deaths
2020-01-21	1	0
2020-01-22	1	0
2020-01-23	1	0
2020-01-24	2	0
2020-01-25	3	0
2020-01-26	5	0
2020-01-27	5	0
2020-01-28	5	0
2020-01-29	5	0
2020-01-30	6	0
2020-01-31	7	0
2020-02-01	8	0
2020-02-02	11	0
2020-02-03	11	0
2020-02-04	11	0

Some Examples

Reliable Data Sources

- **Government datasets**
  e.g., U.S. Census Bureau, EU Data Portal.
- **Established organizations**
  e.g., UN, WHO, OECD, World Bank.
- **Peer-reviewed academic research**
  e.g., Dryad, Zenodo, Figshare
- **Trusted Online Platforms**
  e.g., Our World in Data, Statista

Even More Data Sources…

kaggle

Pew Research Center

DATA.GOV

ICPSR

The General Social Survey
GSS

NOAA

NASA

GitHub
Open Data

ITU

ADB
ASIAN DEVELOPMENT BANK

BANQUE AFRICAINE DE DEVELOPPEMENT
AFRICAN DEVELOPMENT BANK

AFRICAN DEVELOPMENT FUND
FONDS AFRICAIN DE DEVELOPPEMENT

IDB
Inter-American Development Bank

Key Takeaways

1. **Start Broad**
   Use Google Dataset Search or large repositories.
2. **Evaluate Carefully**
   Check source credibility, licensing, and format.
3. **Consider Domain-Specific Repositories**
   For more targeted or discipline-specific data needs.

Stony Brook University
Libraries

Data Literacies

https://guides.library.stonybrook.edu/data-literacy

Thank You!

Ahmad R. Pratama, Ph.D.
ahmad.pratama@stonybrook.edu
Data Literacies Lead

Stony Brook University Libraries
February 10, 2025

@StonyBrook
UniversityLibraries

sbulibraries

stonybrooklibrary