HUM 346: Digital Humanities
Colin Brown and Melissa Woo

The COVID Tracking Project Datasheet

The COVID Tracking Project was born in early March 2020, when two journalists at *The Atlantic*, Alexis Madrigal and Robinson Meyer, began investigating lagging COVID-19 testing rates. They quickly realized that public data streams were limited by overtaxed infrastructure, data reporting across state lines took vastly different forms from press conferences to online portals, and many public health workers and politicians were hesitant to make their data public to begin with. The patchwork of inconsistent definitions and varying reporting methodologies across county and state lines made tracking real-time trends incredibly difficult, confusing even the CDC. But, having reliable information about testing, hospitalization, patient outcomes, and racial and ethnic demographic information was critical for public health officials and policymakers to make key decisions – and for the public awareness needed to motivate widespread adherence to these public safety guidelines like masking and social distancing.

In response, **The COVID Tracking Project** (owned by *The Atlantic Monthly)* collected and published vast amounts of data on COVID-19 testing and patient outcomes from all 50 states, 5 territories, and the District of Columbia. The collection methodologies and data itself were made available through a data API, used by national and local news organizations across the United States and by research projects and agencies worldwide. The project also launched the COVID Racial Data Tracker on April 15, 2020 via a partnership with the Center for Anti Racist Research to collect, publish, and analyze racial data on the pandemic within the U.S. On September 1, 2020, they launched the Long-Term-Care COVID Tracker to consider data on the pandemic in nursing homes and assisted living facilities in the United States.

During the active data collection period from early March 2020 up until March 7, 2021, the COVID Tracking Project relied on a volunteer corp to analyze and extract data from the official COVID-related websites of all 56 U.S states and territories. Interestingly, unlike most number-based datasets, this project relied almost entirely on manually-entered data. When the dataset was actively being updated, about 15 volunteers on a rotating schedule would work for three hours a day to update new data between 5:30 pm and 7:00 pm Eastern time. During their work sections, volunteers claimed a state to analyze over a shared Slack channel, looked for specific values (“Total antigen tests administered”, or “Total PCR positive results”, for

instance) on the state website, and entered them onto a shared spreadsheet that funneled in their central database. The team incorporated an element of redundancy by having volunteers double-check each others' entries, and then once most of the data was entered, there would be another round of checks for large abnormalities in the data.

The primary difficulty that the volunteers faced during these daily data collection periods were inconsistencies in how different states reported the same or similar data because the federal government never administered guidelines for uniform reporting methods. Often it could be that some states have different qualifications for being a "probable" positive case. In response, some of the volunteers were part of a designated Data Quality team. When an inconsistency like this would arise, the volunteer collecting the data would message the Data Quality Slack channel for the team to deliberate and decide on. Once a decision was reached, the volunteer was alerted on how to update the data. If the abnormality had an outsized effect on created discrepancies between states in the data, the team would append a public note to it describing the inconsistency and some of its effects.

By addressing a glaring need with hard work and an innovative approach, the COVID Tracking Project proved to be extremely valuable in informing the public on the status of the COVID-19 pandemic. The project was cited in more than 1,000 academic papers, including major medical journals like *The New England Journal of Medicine*, *Nature*, and *JAMA.* It was used by two presidential administrations and an array of federal agencies, including the CDC, HHS, and FDA in key public information campaigns and policy decisions that had wide ranging implications for our daily lives and the course of the pandemic. Federal lawmakers used figures in at least 11 letters demanding answers on the pandemic response from government leaders and commercial labs. The data was cited in over 7,700 news stories in publications from *The New York Times*, *The Washington Post*, *CNN*, *Vox*, *ProPublica,* and many more.

The broad impact of the data in the multitude of ways in which it has been used is underscored by specific details of how the data was collected and reported. Every single point of data is the result of human decision-making. Decisions about how to define metrics, what to collect, how to group and publish the data,

and how to label and interpret it are deceptively central in shaping what kinds of questions can be responsibly asked of the dataset.

For example, the team's decision to approach data collection with manual brute force uniquely positioned them to effectively aggregate non-uniform data. As described earlier, this involved visiting a number of various sites and sources to observe all of the initial data, investing significant manual research to define and understand each published statistic, and performing manual data entry to then scrape and aggregate the metrics. This became especially valuable as the demand for nationwide COVID-19 statistics and trends increased, with the project's ability to navigate the inconsistent data definitions and varying reporting methodologies across county and state lines allowing them to provide a comprehensive and nuanced view of the pandemic's impact; the body of prominent citing new sources attests to this.

However, the manual collection process also introduced the potential for human error, despite the project's efforts to mitigate this through multiple layers of redundancy and a culture of detail-orientation and care. The reliance on volunteers to extract and input data from various state websites meant that there was a risk of incorrectly-inputted numbers making it into the final dataset. This inherent risk shapes how users should regard and analyze the data; there is a need for robust validation and verification processes when using the data for research or policy purposes, with special attention and scrutiny applied to any particular outliers or counterintuitive trends that are observed. This becomes especially important for any research or data collection on smaller, sub-state populations; in these cases, it may be more reliable to draw straight from state databases themselves rather than this project.

The nature of aggregating data across various initial reporting schemes also influences how data from The COVID Tracking Project can be used; the difficulties regarding these inconsistencies must be understood and processed. The lack of uniform reporting guidelines from the federal government meant that there were significant decisions made on a state-by-state or even county-by-county basis with regard to categorizing "probable" positive cases or aggregating antibody tests with PCR tests (which significantly inflated test counts) in order to address discrepancies in the data. The Data Quality team was responsible for deliberating on how to handle discrepancies and advising volunteers on updating the data. Public notes were

appended to the dataset to discuss any significant inconsistencies and their potential effects. This means that in order for users to fully understand the limitations of the dataset, they must consider these notes to draw accurate conclusions and make informed decisions based on the dataset. Purely quantitative analysis therefore is not sufficient; it must be combined with qualitative awareness of these data inconsistencies and their impacts.

Looking forward, the methodology employed by the COVID Tracking Project prompts important reflections on the nature of data collection and its implications for dataset utility. The reliance on volunteers and diverse data sources highlights the necessity of robust validation processes to ensure data accuracy and reliability. Moreover, the project's handling of discrepancies underscores the importance of transparency and documentation in informing data interpretation and decision-making. The COVID Tracking Project is a testament to the power of collaborative efforts in generating actionable insights from data, and its methodology sparks important conversations about the intersection of data collection, transparency, and effective decision-making in addressing the challenges of our time.

References

"Data API." The COVID Tracking Project, covidtracking.com/data/api

"Data Summary." The COVID Tracking Project, covidtracking.com/about-data/data-summary

Gilmour, Jonathan. "Analysis & Updates | 20,000 Hours of Data Entry: Why We Didn't Automate Our Data Collection." The COVID Tracking Project, 28 May 2021, covidtracking.com/analysis-updates/why-we-didnt-automate-our-data-collection

Kodysh, Julia , and Jonathan Gilmour. "Analysis & Updates | How and Why the COVID Tracking Project Built a Screenshot System." The COVID Tracking Project, 4 May 2021, covidtracking.com/analysis-updates/how-why-covid-tracking-project-built-screenshot-system

The COVID Tracking Project. "The COVID Racial Data Tracker." The COVID Tracking Project, 7 Mar. 2021, covidtracking.com/race

"The COVID Tracking Project." The COVID Tracking Project, covidtracking.com/

"The COVID Tracking Project." Reveal, 15 Apr. 2023, revealnews.org/article/covid-tracking-project/

"The COVID Tracking Project." GitHub, github.com/COVID19Tracking

"The Long-Term Care COVID Tracker." The COVID Tracking Project, covidtracking.com/nursing-homes-long-term-care-facilities.

Dataset Name	Link to data source	Link to storage source	Short Description	Who collected the data	Who owns the data	How was the data collected	Sample Size	Who was included/excluded from sample	When was the data collected	When was the data last updated	Why was the data collected	Notes on data quality	Notes on data usage conditions
The COVID Tracking Project	https://covidtracking.com/	https://covidtracking.com/about-data/data-summary	- Cumulative daily totals of national level metrics for cases, tests, hospitalizations, and outcomes. - State-level metrics for cases, tests, hospitalizations, and outcomes. - Metropolitan level case and death data broken down by race and ethnicity (where available) for 65 cities and counties from May 29 to October 21 (note: not all locations were tracked for this entire time series). - CSV files of every long-term-care facility we collect data for, and every state's total cumulative and outbreak numbers.	The COVID Tracking Ground, a volunteer organization launched from The Atlantic	The Atlantic Mor	About 15 volunte	56 US states and territories, 65 cities and counties	Definitions for each metric collected are stated and aggregated across various data sources	Early March 2020 - March 7, 2021 (daily, weekly)	March 7, 2021	In March 2020, necessary COVID tracking data could be found only as a patchwork across cities, counties, and states. Some local governments launched COVID-19 data dashboards; others reported data in press conferences—or not at all. Volunteers began combing through all of these sources via daily, close contact with the data, which was necessary to understand what states were reporting and aggregate information over time.	Standardized race and ethnicity categorizations based on the US Office of Management and Budget's guidelines (used by the Census Bureau)	Freely available through data API on website