AMY AFFELT

Location, Location, Location: Where's the COVID-19 Data?

As librarians and information professionals, we tend to meet most statements and proclamations with a healthy sense of skepticism and a question: "Where's the data?" We know not to take information at face value and to make sure that sources and methodologies are sound. However, just as it has with pretty much every other aspect of life, COVID-19 turned information verification and evaluation on its ear.

CONFLICTING DATA

As a May 19 *New York Times* article warned, "Air Travel Surges by 123%! (Beware of Misleading Data Like That)." Typically, looking at a percent change is a reliable guide to the economy. But if an economy is completely shut down, all bets are off—because as it reopens, we are starting from scratch. Similarly, we have never had previous COVID-19 infections, so the percent change comparisons are based on a very small time period (late winter/early spring 2020) and involve numbers regarding something unique. For example, a headline stating a percent change in COVID-19 deaths can be misleading, as some states' current reports include presumed deaths from the past. On June 25, New Jersey reported a 14% rise in deaths, but this was after state officials went back and reviewed death certificate dates from late winter ("New Jersey's Covid-19 Death Toll Rises 14% When Probable Fatalities Added").

In the nascent stages of the pandemic, I had high hopes for Google's Dataset Search. When it was launched in September 2018, Google explained that its repository is based on the metadata received from data publishers: creator, publication date, collection method, and terminology. I decided to give it a test run by doing a search for "coronavirus." The first two results were for generic sites from the CDC and the World Health Organization (WHO), but the subsequent results were competing infection map sites that contained different information, leading to conflicting conclusions. It is unclear how these results are chosen and ranked. For example, another search for "virus peak map" again listed a generic CDC site as the first hit, and *The New York Times'* data map sites as the second and third, but the fourth hit was a weekly national seasonal flu report from the British government.

Clearly, I needed to try regular Google instead in order to find the data that I needed. A search for "percentage of asymptomatic coronavirus" uncovered conflicting results, but they were relevant and easy to parse. Top hits included data from the CDC (35%) and the WHO (80%) and a *Washington Post* article that linked to a study in the *Annals of Internal Medicine*, which cited 16 different sources with data in the 40% range. Although these are all credible sources, the disparate percentages made me pause. The problem of widely varying datapoints being used to answer the same questions is endemic to COVID-19; however, regarding total number of worldwide infections and deaths, most organizations are in agreement, with the WHO reporting around 10 million infections and about 500,000 deaths on June 30, which aligns with Johns Hopkins University's data.

When conflicts arise, it is important to keep in mind that collation tools use data from different sources. For example, when using state data in the U.S., it is important to remember that sometimes deaths are double-counted; for example, when a Florida resident died in Los Angeles, *The New York Times* documented the death as having occurred in California, while the state of Florida counted it as its own. Also, hospitalization data requires a thorough review, as not all states report hospitalizations; reporting is not required by the CDC.

The verbiage used in the metadata is also important. Numbers of "reported cases" are going to be lower than numbers of "actual cases," since many people experience symptoms and do not seek treatment or testing. In June, when a WHO official stated that the spread of the virus by asymptomatic individuals is "very rare," the agency tried to clarify that characterization by saying that the term "asymptomatic" is often used incorrectly and interchangeably with "presymptomatic," which describes patients, who—while they are in stages of infection in which they are not yet experiencing symptoms—will go on to experience symptoms and spread infection. Similarly, officials in some states have documented new cases without making a note of where those patients were being treated.

INCOMPLETE SOURCES

How do we know which sources to use when citing COVID-19 data? For straightforward numbers to answer specific questions, the WHO and the CDC are excellent. For data visualizations surrounding some of the most important concerns (such as outbreaks, risk, and testing), there are several mapping tools that can be used. Covid Act Now provides statistical data on local risk levels by modeling data from *The New York Times* and other sources. It looks at four indicators: decrease in COVID cases, rate of testing, hospital preparedness, and the speed at which contact tracing provides discovery and isolation of new cases. The site offers an open source API, which—along with the fact that it is run by researchers from Stanford and Georgetown universities—lends credibility.

In evaluating Covid Act Now, I was a little concerned that it may be incomplete; for example, on July 6, the state of New Mexico was listed as "at risk," but a search for Taos County, N.M., yielded a result stating, "We don't have enough data to assess COVID risk." Taos County data is not difficult to find; the New Mexico Department of Health updates a dashboard with the information on a daily basis. The Taos County data situation is just one example, but it underscores a best practice: In searching for population- or location-based public health data, go to the most localized, on-the-ground source that you can find (typically, a local health department or city government site).

LINKS TO THE SOURCE

"Air Travel Surges by 123%!
(Beware of Misleading Data Like That)"
nytimes.com/2020/05/19/upshot/virus-economic-data-upended.html

"New Jersey's Covid-19 Death Toll Rises 14%
When Probable Fatalities Added"
wsj.com/articles/new-jerseys-covid-19-death-toll-rises-14-when-probable-fatalities-added-11593109913
(subscription required)

Google's Dataset Search
datasetsearch.research.google.com

***The New York Times'* death discrepancies**
tinyurl.com/ybjc9cqn

Covid Act Now
covidactnow.org

New Mexico Department of Health's COVID-19 database
cvprovider.nmhealth.org/public-dashboard.html

Pandemics Explained
globalepidemics.org

Business Insider's risk assessment
businessinsider.com/riskiest-to-least-risky-activities-during-coronavirus-pandemic-ranked-2020-5

Scott Gottlieb's tweet about COVID-19 risk
twitter.com/ScottGottliebMD/status/1279108199891767296

CNET's risk assessment
cnet.com/health/what-activities-are-safe-to-do-during-coronavirus-pandemic

NPR's risk assessment
npr.org/sections/health-shots/2020/05/23/861325631/from-camping-to-dining-out-heres-how-experts-rate-the-risks-of-14-summer-activit

Pew Research Center study of COVID-19 research
pewresearch.org/fact-tank/2020/07/08/most-americans-say-despite-ongoing-research-ways-to-limit-spread-of-covid-19-are-well-understood

RISK ASSESSMENT

Harvard Global Health Institute's Pandemics Explained site features an interactive map depicting risk assessment and testing targets, which illustrates states' abilities to test widely enough to mitigate or suppress the virus. Information such as the number of people tested per day per 100,000 people is contrasted with the number of tests per day needed for suppression and mitigation. The numbers are alarming. For example, on July 6, the U.S. was conducting 170 tests per day per 100,000 people, with 355 needed for mitigation and 1,304 needed for suppression. This tool goes a long way toward helping us understand where we need to be and what we need to do to get there, both locally and globally.

Some of the most common questions and concerns regarding the pandemic involve the riskiness of resuming activities after stay-at-home orders are lifted and businesses are reopened. I have seen many charts that rank common activities on a scale of 1 to 10; most are in alignment. For example, data from Business Insider, former U.S. Food and Drug Administration commissioner Scott Gottlieb, the CDC via CNET, and a panel of experts assembled by NPR agree that while opening your mail is very safe, getting a haircut involves medium risk, and going to a bar is a five-alarm fire.

THE FUTURE OF DATA

Ultimately, while risk levels of activities have a general consensus, most other COVID-19 datapoints (such as how the virus is transmitted or how long it remains active on surfaces and face masks) is up for some debate. As librarians and information professionals, we are skilled at choosing credible sources, but how should we advise our constituents when the data from these sources are in conflict? The best practice is to send all available data, explaining any obvious limitations or sources of discrepancy, and let the constituents make the judgment call regarding which source to use. Ideally, they would use a range of sources.

There are few truisms regarding COVID-19, but one is that it is safe to say that the more information that is available, the better. A Pew Research Center survey conducted from late April to early May found that 78% of Americans believe "it makes sense that studies may have conflicting advice because research is constantly improving." Finally, some good news!

AMY AFFELT is director of database research worldwide at Compass Lexecon, a global economic consultancy, where she finds, analyzes, and transforms information and data into knowledge deliverables for Ph.D. economists who testify as experts in litigation. She is a frequent writer and conference speaker on Big Data, the Internet of Things, adding value to information, evaluating information integrity and quality, and marketing information services. The author of two books, *The Accidental Data Scientist: Big Data Applications and Opportunities for Librarians and Information Professionals* (Information Today, 2015) and *All That's Not Fit to Print: Fake News and the Call to Action for Librarians and Information Professionals* (Emerald, 2019), Affelt was the Big Data columnist for *EContent* magazine. She is also an SLA fellow. Affelt has a B.A. in history, Phi Beta Kappa, from the University of Illinois–Chicago, and an M.L.I.S. from Dominican University. Send your comments about this article to ecletters@infotoday.com or tweet us (@ITINewsBreaks).