Report on Digital Infrastructures Grant: COVID data infrastructure builders

The project team:
Amelia Acker (grant co-I)
Bidisha Chaudhuri
Ryan Ellis (grant co-I)
Megan Finn (grant co-I)
Youngrim Kim
Janaki Srinivasan
Stacey Wedlake

Executive Summary: Learning from COVID

Our project examined the creation and maintenance of COVID data dashboards. We reviewed key dashboard artifacts and interviewed 79 data workers associated with 17 different COVID data projects in the US and India. These projects were run by volunteers, non-profit employees, academics, members of the media, and employees. The core questions that we tackled include: How are COVID data infrastructures created and transformed by builders and maintainers? What kind of support is needed for a diverse community of workers to continue their work?

Our work uncovered the challenges, conflicts, successes, and failures of building public health data infrastructures during a time of political strife and devastating crisis in 2020 and 2021. Our analysis underlines an important lesson: the process of curating data and making infrastructure is always contested and shaped by the varied politics, skills, and expertise of its builders. Specifically, we identify six core lessons:

- *Lesson #1: Long-Term Support and Funding is Vital to Sustainable Data Projects*
- *Lesson #2: Data Projects Can Drive Institutional Change*
- *Lesson #3: COVID Data Projects are Incubators for Skills and Talent*
- *Lesson #4: Data Projects are Social Worlds and Should be Cultivated as Such*
- *Lesson #5: Investing in Data Workers and Management is an Investment in Sustainability*
- *Lesson #6: Data Projects Depend on Institutional Data*

Taken together, the project depicts an ecosystem of data work and projects that is at once hopeful and worrying. One optimistic finding this research underscores is that a number of key pathways that can help support future collaborative open data projects: investing in and supporting workers, cultivating diverse management teams, and focusing on supporting the technical and non-technical work of large data projects are vital to creating sustainable projects. It also identifies key positive spillover effects associated with not just COVID data projects, but likely, other large data projects, including: providing on the job training for the next generation of data scientists and driving real-world change. Yet, there are also notes of caution. Successful projects rely on and require a deep foundation of pre-existing support and infrastructure—enthusiasm, an army of volunteers, and good will can only take a project so far. There are also a legion of barriers to creating sustainable data projects. Additionally, the success of data projects often is tethered to the quality and availability of data produced by state agencies. As COVID data builders discovered, careful parsing, sharp analytic capabilities, and countless hours of hard work can rarely overcome data that is unmade, incorrect, corrupt, or otherwise unavailable. These lessons ultimately suggest a roadmap for how to support future data projects.

Learning from COVID: Supporting and Sustaining Open Data Projects

Overview: Calculated Risk

COVID Dashboards became ubiquitous in the timeframe of our data collection: 2020-2021. Thousands of competing and complementary displays—including official, government-backed, public health dashboards, large volunteer tracking projects, academic efforts, activist interventions, and many others—dotted the landscape. All these systems presented particular views of the pandemic and offered different ways to interpret the unfolding public health crisis. High-profile dashboards, such the Covid Tracking Project in the United States, became near de facto standards, charting the ebb and flow of cases, test positivity rates, and other pandemic data. These popular dashboards were routinely cited by government officials, public health experts, and media outlets; they were authoritative snapshots of the unfolding pandemic read and used by millions. Other dashboards served niche audiences by focusing explicitly on populations or topics that were often overlooked or ignored, such as the spread of COVID among persons who were incarcerated in the USA or the local availability of oxygen tanks in West Bengal, India.

The COVID Data Infrastructure Builders project interrogated the creation and maintenance of COVID data systems. It examined the people, assumptions, and resources that came together to create and maintain these systems. Over two years of work, the project focused on 17 COVID critical data infrastructures in the U.S. and India. Taken together, an analysis of these projects highlights the challenges and difficulties faced by those involved in the provision of these critical data infrastructures—we call these workers "COVID Data Infrastructure Builders." Through detailed interviews with over 79 key builders, the project uncovers the hidden work and resources that made these resources possible. At first blush, COVID dashboards appear to be ad hoc pandemic improvisations: projects built on the fly to serve a pressing public need. But closer examination revealed a more telling picture. COVID dashboards were *always* built on a foundation of preexisting resources. Tools, relationships, and institutions provided the foundation for these novel data projects. Maintaining these projects was at times a herculean task; it involved continually marshaling and organizing resources. Managing these projects required more than simply managing incomplete, error-prone, and ambiguous data: it required managing participants working long-hours under difficult conditions and working with competing publics that were at turns helpful and contentious.

The insights drawn from studying COVID-19 dashboards provide lessons that are relevant for not only public health projects, but apply more generally to large data projects. The challenges that these data projects faced are not unique to public health data. Likewise, the lessons learned can and should help policymakers and funders think strategically about how best to support large data projects and, importantly, the benefits that these projects can deliver.

Methods & Analytic Approach: Disasters, Data, and Dashboards

Disasters are important research sites from which scholars can understand how social worlds and built environments construct the impact and possibilities of something like a virus. To better understand our research questions, the project focused on data projects based in two countries, India and the United States. The unique strength of the project lies in its transnational framing in understanding digital infrastructures in the "global south" and the "global north," with researchers based in both geographies. The US and India, two of the world's largest democracies, also both have robust information technology industries and large open source and open data communities. Both countries had some of the highest recorded COVID case counts and deaths. The extent of havoc wreaked by the pandemic in the two countries resulted in a demand for accurate information about the impact of the virus among their populations. Yet, there was also significant differences in the two countries' histories, administrative practices, work cultures, and the resources available to their governments and residents throughout COVID. Even within each country, the different organizational forms in which these infrastructures are built adds to the range of data projects that emerged in the context of COVID.

The research was divided across three phases: Phase I: Initial Review and Selection; Phase II: Participant Interviews; and Phase III: Analysis and Dissemination.

Phase I: Initial Review and Selection
During this phase, a cross-section of dashboards in both countries were initially archived (capturing basic dashboard data) and reviewed by the project team. The project team cast a wide net, reviewing high-profile projects, municipal dashboards, micro-targeted projects (for example, dashboards associated with individual school districts in the US), and others. A diverse subset was then selected for closer analysis and participant interviews. This sample included both popular and widely-used dashboards, and smaller, less-well known dashboards (See, Appendix: Dashboards at a Glance for a detailed overview of some of the dashboards we looked at). We collected screenshots of the dashboards, documents associated with the different dashboards including blog posts or articles authored by members of the projects, media mentions and podcasts about the projects. Additional primary and secondary data was collected about the dashboards in order to inform Phase II.

Phase II: Participant Interviews
The most important data that we collected were interviews with people who worked on the projects that we described above and were the primary sites of our analysis. During this phase, an interview sample was developed focusing on participants associated with 17 different dashboards. We were opportunistic about getting interviews and sent "cold" emails to many more people than we interviewed. Almost all of the interviews stemmed from cold emails, or from people we interviewed who recommended who we could reach out to next (a snowball sample). In the case of WBCAN and the Karnataka War Room, we reached out to personal contacts initially and then

relied on recommendations from interviewees for who else to contact. In the case of COVID Tracking Project, one of the project members contacted aCOVID Tracking Project project leader over Twitter, and they brought us into the project Slack, and we solicited interviews through the project Slack. We got no responses to many requests for interviews with people on projects that we don't list here or to people on projects that we do list here.

During March 2021- November 2021, we conducted 79 one-to-two-hour interviews with COVID data infrastructure builders across 17 COVID data projects (See Table below). The interviews primarily focused on the work they had done since the pandemic began (in the Appendix: Dashboards at a Glance we describe each of the data projects we investigated; Appendix: Interview Sample and give more details about our interviewees). Our interview sample is somewhat theoretical in that we targeted talking to people on the projects who worked directly with data collection, processing, analysis, visualization, and infrastructures (data workers rather than project leaders). We also talked to people involved in other parts of the projects to fully understand the vast work of dashboards. When projects had people who were paid and unpaid or undergraduate and graduate, we tried to make sure that we talked to people who had different levels of expertise or involvement with the project. For all of the dashboards, we attempted to get to saturation, but we were not always able to do this because people didn't always follow up with us.

COVID data projects (US-based projects in blue; India-based projects in white)	Interviewees
COVID Action Network (CAN)	5
City of Austin, TX	1
CODD-K	3
Consumer Voice	1
CoronaSafe	2
COVID Alliance	1
COVID Tracking Project	23
Emory Health Equity Dashboard	6
Johns Hopkins University's COVID tracker	9
Karnataka State War Room	6

One Point Three Acres	4
Mercy Mission Bangalore	2
The Marshall Project	4
Non-COVID Death Tracker	1
Stop AAPI Hate	1
West Bengal COVID Action Network (WBCAN)	6
Zoho COVID Dashboard	4
TOTAL	79 (25 India & 54 US)

Interviews were generally conducted with one interviewer from our team and one interviewee. In a few cases–with Mercy Mission, CAN, and WBCAN – the interviewees preferred to talk to us in groups of 2-4. We conducted a small number of interviews with people who worked on projects that are not listed above where the interviewee did not consent to be recorded or the interviews were "on background." Typically, the interviews were semi-structured, meaning that we prepared interview guides that we tailored for interviewees depending on their roles in the project and altered during the course of the interview. Interviewees were given the opportunity to use their real name or a pseudonym. All interviews were recorded and transcribed. Interview subjects were asked to review and edit the transcript for accuracy or other reasons. After interviewees edited their transcripts (or chose not to), the interview transcripts were reviewed by the project team and a detailed code book was created. After drafting the code book, each interview was analyzed and coded according to the agreed upon code scheme. Interviewees in the US were given $50 gift cards to Amazon. Some interviewees did not accept the gift cards, which we then used to pay other participants. In India, interviewees received Amazon giftcards of INR 3000.

Phase III Analysis and Dissemination

Immediately after the interviews, we wrote memos about the interview that addressed what we learned from the interview and questions that the interview raised. The compiled memos were about 70,000 words.

The research team listened to interviews that were conducted by other members in order to collaboratively draft a scheme for qualitatively coding the interviews. We piloted the coding scheme, together coding the same interviews and then comparing the work we did. We iterated on the coding scheme and performed the same process of testing the scheme on interviews together two more times. Satisfied that we had a coding scheme that everyone could consistently use, we divided up the interviews among the team and coded them.

Once we coded the interviews, we assigned particular codes to specific team members. We then collectively analyzed the coded quotes for themes within each area. We wrote memos about the themes and organized quotes that illustrated that theme. The themes that we identified in the process of doing this work are the foundation for our findings.

Taken together, by the end of the project we had produced thousands of pages of interview transcripts, 150 pages of research memos, an iteratively created codebook, comprehensively coded and annotated transcripts, and about 15 thematic memos.

Our findings were disseminated across multiple venues. Initial findings were presented at four academic conferences (see "Project Outputs" below). Research results were collected in two research articles and form the basis of a larger in preparation book manuscript.

Key Findings & Lessons Learned

The work of data builders involved a blend of technical and non-technical skills. Creating and maintaining COVID data projects went well beyond the day-to-day work of collecting and curating data. Data builders were organizational entrepreneurs: they marshaled and drew together existing resources to animate their projects. While starting a new project was difficult, sustaining a project was near-impossible. Shifting governmental data practices created frequent uncertainty—stabilizing data practices (and by extension, the larger project) proved difficult. All the projects faced the difficulty of working during a period of emergency and anxiety.

COVID Data Builders did more than organize data, they organized publics' knowledge of COVID. Data never spoke for itself. It always called for interpretation. As project members worked to collect, sort, and eventually push out into the world reams of public health information, they quickly confronted the realities of dealing with different publics. The data was misread; the data was appropriated and spun into what looked to be deliberate misinformation; the data was confusing; the data was ignored. Collecting and organizing the data was only half the battle: they also had to make sure it made sense.

Data builders confronted the problems of the public in different sorts of ways: they engaged with the public through setting up helpdesks, providing contextual explainers and supplementary materials, drafting F.A.Q.s., and participating in ongoing conversations with members of the public. This work was both personally sustaining—it helped data builders feel their work was important to others and underlined that their projects could really make a difference. But attention also made the COVID Data Builder feel pressure to perform their often uncompensated work with exquisite precision. At the same time the fact that they were doing the work of communicating the impact of the COVID virus dispiriting to individual volunteers: data builders wondered if they should be occupying these roles in the first place. The implication for future data projects is clear: data does not speak for itself; it always requires interpretation. Building in clear engagement strategies to address misinformation, misreading, and user feedback is not simply something additive: it is vital. The pressure to turn bad COVID data into something useful for different publics added to the mental health challenges of the pandemic.

The examination of the lessons that we learned and the builders that we interviewed learned uncovers a number of key lessons and insights for policymakers and funders active within digital infrastructure projects and crisis infrastructures. We distill these below.

Lessons #1: Long-Term Support and Funding is Vital to Sustainable Data Projects

Successful COVID data projects *were always built on existing foundations* of institutional support and resources including different social and financial inequalities. Even seemingly temporary and ad hoc projects sat on a deep foundation: pre-existing technologies, relationships, organizations, and institutions were often central supports for these seemingly improvised projects. At the same time, much of the labor of creating and sustaining COVID dashboards relied on managing relationships. Data workers always were contending with more than just unruly and absent data: they were managing complex social relationships—both intra-project relationships among participants and relationships between the projects and publics that were alternately supportive, contentious, overwhelming, and absent.

These foundations were critical to the success of these projects. COVID data projects were spun up during the first moments of the COVID-19 pandemic. They recognized and responded to a significant gap in public health reporting, cobbling together incomplete data, software tools, and diverse participants to create dashboards that provided the public, policymakers, and public health officials with an evolving snapshot of the unspooling pandemic. The seeming improvised nature of these dashboards and the occasional informal nature of their founding, however, should not obscure a larger lesson for policymakers and funders: pre-existing foundations were vital. The JHU dashboard, to pick but one example, thrived during the early days of the pandemic thanks to the existing resources that were already in place. The Applied Physics Lab—a component of JHU that has been funded for years through federal contracts—had previously created customized software and cultivated expertise in data science, public health, and project management that were vital to the success of the dashboard. Elsewhere, projects such as the Covid Tracking Project (CTP) built their dashboard on a foundation of existing software tools, including free open source projects such as postgres and donated for-profit platforms such as Slack, in addition to accumulated expertise and relationships that allowed the project to thrive. This included free gifts of existing collaboration platforms such as Slack. Projects such as WBCAN relied on a network of friendships and professional relationships to thrive. And groups such as the Karnataka War Room used Rotary Club connections to build their group.

Interestingly, all of the projects that we looked at used Google sheets to collaborate and work on data. Google Sheets were easy for people to learn and very flexible for the builders, accommodating changes in the project as the pandemic data and project personnel changes. Google Sheets allowed for lots of different practices to happen in one place. For example, in the United States, different states may record race data differently; sheets allowed builders to record different kinds of data in one place with notes. However, while Google Sheets was a commonly available tool that had a learning curve that all of the builders we spoke to could learn, it was not without issues. Some builders noted that there was a tradeoff between how easy a technology like Google Sheets was to use (it was easy to bring in volunteers to work with it) and the complexity of the task at hand (Google sheets had limits on the size of the Sheets and issues with integrity compared to databases).

We found that projects typically started or shortly after gathering daily data began to establish protocols for gathering and managing data collection and processing using a standard tool chain. The tool chain consisted of—Slack or WhatsApp for messaging volunteers, GitHub for code versioning and data management, and spreadsheets for gathering and publishing dashboards (usually Airtable, or Google Sheets). COVID Tracking Project, had well-connected leadership who received donations of cloud products or software for free. However, other projects were also circumspect about how much they could really rely on WhatsApp and free software tools.

The larger technical and economic foundations of the projects were not always sustainable: as the pandemic continued, attention and resources became, at times, more difficult to sustain. There were a few different supports that we understood to help COVID data infrastructure builders' work. First, many of the COVID data projects had strong missions that many people on the project discussed and understood. Contributors focused on public impact or communications worked to demonstrate their impact to other people working on the project. This created pressure for people, but also motivated them. Second, some projects paid extra attention to mental health challenges created by COVID and by working on the project. Some projects even had people with social work or psychology backgrounds to provide guidance and support. COVID data builders needed and appreciated these supports. Last, many of the COVID projects drew on existing social networks for volunteers; on other projects, people developed friendships with contributors. These social connections and the project culture supported COVID data builders over periods of time.

Repeatedly, our work uncovers an important lesson: the agile, seemingly ad hoc, projects that leapt to fill an important public need during a moment of crisis sat on pre-existing social, economic, and technical foundations. Supporting these foundations is an investment in supporting the next wave of successful data projects.

Lesson #2: Data Projects Can Drive Institutional Change

COVID data projects drove institutional change: *by pushing for improved collection and distribution of public health data by government entities, some state agencies changed their data practices*. COVID data projects confronted a landscape of fragmented data. It was particularly hard to understand how the pandemic affected marginalized groups of people. Additionally, different jurisdictions followed their own definitions of data units and cadence for releasing data, making aggregation and analysis across regions difficult. COVID data builders worked with the data they had in hand, while agitating directly and indirectly for better data. As some projects got attention, public officials started to work to tailor their data to match the needs of the dashboards. While the dashboards worked continually to manage and accommodate the idiosyncratic data practices of different jurisdictions, this relationship also, in time, worked in the opposite direction. State officials worked to provide the data that the dashboards wanted and needed. COVID data projects became a driver of institutional change—they served as de facto standards bodies, nudging the mishmash of different and incompatible standards closer to something that was coherent. When states started to collect and report vaccine data, they looked to show how dashboards were reporting this data—a clear sign that these private projects could drive serious institutional change.

Along with driving institutional change, the COVID data builders also spoke about the feelings of immense responsibility that they felt. The builders could see that the data was valuable to decision

makers, themselves as citizens, and other people who reached out to them because they worried about COVID. Some dashboards such as JHU, CTP, and The Marshall Project put pressure on government officials in public health offices and in the prison system for data which led to more data transparency, data accessibility, and government accountability. When they felt that the data they produced had implications beyond their projects – whether to people in positions where they had power or to individuals trying to navigate their personal lives and workplaces – the builders spoke extensively about how that drove decisions to be purposeful. In the politicized climates in the US and in India, they felt they had to be doubly careful.

In their own projects, builders emphasized the importance of public communication – being open and clear about what was known and not known and figuring out how to translate abstract statistics to different people. The Zoho team explained that while quantitative data reached people in their head, that qualitative data reached their hearts. For some projects this included doing chats on platforms such as Whatsapp and releasing videos on social media to answer people's questions. Some dashboards, such as 1point3acres, made particular efforts to present data and dashboards in different languages and focused on a wide variety of countries.

The drive to be open about data issues and clear in communication was driven by the impact that builders believed they were having both on policies and on people's lives. Workers on the JHU dashboard, for example, noted that when data were not produced by the public health departments in the position to make data, that you might have to "advocate and rattle some cages." And the builders at the Emory Health Equity projects discussed how flexibility and creativity were required because you sometimes had to work with the data that you have, even when it was not what they were hoping for.

While most of the builders we talked to could point to some ways in which they felt their work led to changes, some of the builders felt disheartened due to the lack of engagement with their particular dashboard and data. And while many of the builders were proud of the work that they had done and how they felt that they helped people, some of the builders acknowledged that the data work had led to many research questions – some of which they could attend to, but many of which they could not.

Lesson #3: COVID Data Projects are Incubators for Skills and Talent

COVID data projects provided on-the-fly training in technical, distributed team projects. COVID data projects, in many instances, relied on volunteer labor to collect, review, and analyze incoming data. Teenagers, college students, and other curious amateurs without any pre-existing data science flocked to COVID data projects during the early days of the pandemic. They received an invaluable education. Working on these projects provided participants with skills in data science and, in many cases, a new passion. Through formal and informal mentoring and on-the-job training, participants learned new skills in data management, software design and use, and project management. The development and acquisition of these new skills is a key—and easy to overlook—success story of the COVID dashboards.

Many of our builders came to work with COVID data with no experience in the domains in which they were working. And many people discovered that working with the datasets that they were using or making took a lot of specific knowledge about the data. We heard from many people that

they had to spend a lot of time understanding their data before they could show it to people; some called the data "nuanced" to describe how it couldn't be interpreted without work. Data doesn't talk or mean anything without the context of the data. For example, builders had to learn about how different regions of the country reported testing results – while some reported the number of positive tests based on the number of specimens tested, others reported the number of people who tested positive. These kinds of vexing problems with counting different units required COVID data builders to come up with a myriad of approaches for learning how to be about to count a total number of positive tests or compare one region or another. They had to burrow into public health websites, learn how testing worked, and experiment with different statistical methods. Tamara explained:

> the importance of data standardization. I think there were a lot of, as we experienced this, there were a lot of, I want to say confusion on say what to report, how to report. What's the definition of a probable case. What's the definition of a confirmed case. None of that was established and standardized such that you can have a full understanding of grasp of what these states were reporting. I think standardization, data standardization is another I say impact of this dashboard (Tamara, JHU).

Many of the builders that we spoke to who focused on data collection and analysis learned about data standards and public health data.

While the skills of learning about data standards were quite specific to the projects that they worked on, builders learned skills that they believed would be useful later such as working with data, google sheets, and making data visualizations – all technology-oriented skills that would be helpful for collaborating on other open data projects. Some projects gave volunteers a direct opportunity to contribute to open-source software projects.

In addition to the data-related or software-related skills that builders obtained, they also learned organizational and management skills for working in FOSS projects. Almost all of the dashboards involved volunteers with a few 100% volunteer-run. Almost everyone that we spoke to who was in a leadership position working with volunteers explained how challenging it was to work with people freely giving away their time. It was hard to ensure that volunteers were motivated enough that they could be counted on. Additionally, volunteers' skills that they came to projects with didn't always match up with what the project needed at the time and the required training (often from other volunteers) on how to be productive members of their teams. These COVID data builders learned about project management, motivating volunteers, task distribution, and matching expertise with projects – all valuable skills beyond COVID data projects.

Lesson #4: Data Projects are Social Worlds and Should be Cultivated as Such

COVID data projects were more than workplaces—they were sources of sociability and community for participants. Countless volunteers poured hours of their lives into making and sustaining COVID dashboards. While initial motivations for joining these projects varied, the importance of community was crucial in keeping them engaged and involved. Over time, participants became friends—a community with their own in-jokes, shorthand, shared triumphs, and communal frustrations. While this development was in many ways organic, it still required tending and support. Managers worked to carve out space where new communities could take root.

This work—cultivating community—was not extraneous to the success or sustainability of the project: it was indispensable. The sense of growing community kept participants engaged and satisfied.

There were sites that cultivated sociality and built online social life at a time of pandemic-related anxiety. The communities that builders made during the time of isolation was important to our builders. Our interviewees spoke about the gratitude they felt for the generosity of their peers and the collaborative work environment. Projects with workers from different generations found older participants feeling optimistic about the dedication of young people and younger people were excited to work side-by-side with experienced older people. Underlying much of these positive feelings was that our participants were enthusiastic about how much people wanted to help; this is what one interviewee called a "culture of willingness" and fearlessness about driving forward.

But all this good will was not without a lot of good faith communicative labor from both volunteers and team members who led projects. Builders in management positions had to deal with political differences on the team. Many people we spoke with stressed that team communication is crucial as well as a "time soak," particularly when it was unclear what is going on with the project – which we found was typical for a pandemic filled with so much uncertainty. And, when there were many workers on projects, and different teams, workers found that communicating across teams was more challenging, but even more important. The builders that we spoke to said that working on the COVID projects taught them the value of organizational transparency. Across many different projects, the builders that we spoke to emphasized the need for high quality documentation to help with team communication.

The projects faced several challenges that created real pressure: burnout was a common theme across dashboards. Maintaining a dashboard required ongoing affective work: this work was, in many ways, just as important as maintaining the data pipeline. Without it, the projects would fracture and sputter. Dealing with data about sickness, death, and discrimination was an emotional burden to the builders. While many of our interviewees could easily speak about the emotional challenges of working with COVID data, fewer spoke of the pandemic itself as an emotional challenge. Yet the people we spoke to got sick, lost loved ones, and were socially isolated and dealing with the challenges of caring for others during a pandemic as well. The COVID data builders told us about feeling burnt out and working a lot on their projects. One project brought in a psychologist to help project volunteers manage their immense anxiety and suffering. Yet builders on the different projects were bound by a shared mission because of the bounded scopes of the projects. Many builders felt the adversity brought them together.

Lesson #5: Investing in Data Workers and Management is an Investment in Sustainability

COVID data projects were, in some instances, intensely hierarchical. Despite low-barriers to entry, the underlying organizational structure of some projects placed important decisions in the hands of a few, select, project managers. At the same time, a large pool of voluntary workers performed important day-to-day tasks and key roles. Supporting a robust management structure that is not tied to initial project founders is likely crucial to the long-term success and ongoing sustainability of projects. Additionally, supporting workers that populate these projects is vital: it can vitiate the cross-pressures of competing obligations and, importantly, the specter of burnout.

COVID data projects mirrored, in many cases, other large collaborative projects: they drew together a variety of people with different backgrounds and expertise in ways that were beneficial while also creating coordination challenges. Some of our interviewees (sometimes in leadership roles themselves) emphasized that project leadership was important–they played a key coordinating role. Many projects thrived because its leadership had personal connections that made particular resources available to the project. But we observed that leaders needed successors for when they got bored, couldn't sustain their efforts, or when the vicissitudes of life simply made it difficult to continue work. Once these individuals decided to move on, the projects ceased (at times over the desires of day-to-day workers). This provides an important reminder: *data projects need to invest in redundancy not just in their technical infrastructure but in their organizational structur*e. Senior management and project leadership are often critical vulnerabilities and single points of failure. An inability to grow the next cohort of managers and leaders can prove fatal to the long-term success of the project.

Investing and supporting the day-to-day workers that populate these projects is, however, just as important. In most of the cases we examined, COVID data projects relied on large pools of unpaid labor. Many COVID projects relied, at least initially, on manual data entry. For some projects, automation would later complement but not replace the necessity of relying on a large number of voluntary workers. Questions of equity and sustainability circled in the background of these projects. Who got paid and who did not was a question for some projects that strongly shaped what could be done. For some builders, working with volunteers was a challenge that presented a growth opportunity. Despite the esprit de corps that we observed among data workers, burnout was often hovering in the background. Long-hours and the difficult and often upsetting nature of the data, made this work difficult for many. At the same time, for voluntary workers, there were other cross-pressures: school, day jobs, and other priorities all bid for time. Supporting the pool of workers that made these projects possible is an investment in long-term sustainability.

Interestingly, one of the most long-running projects was JHU. This project was housed within a stable and supportive institution, with a defined and diverse set of managers, and it was run and staffed by paid employees, not volunteers. The workers faced pressures and difficulties that mirrored other projects–this was, afterall, difficult and draining work. But, the guardrails and support that were afforded to all participants in the project likely account for its endurance.

Finding ways to pay maintainers and key participants can extend the life of data projects. For the most part COVID data projects shutdown before the pandemic concluded. Different COVID data projects sunset for different reasons, but the fragility and down-stream pressures of relying on voluntary workers is one important reason. Overall, supporting diverse management and supporting project staff is an investment in creating sustainable programs.

Lesson #6 Data Projects Depend on Institutional Data

All of the data projects that we examined were an implicit critique of the status quo. In nearly all cases, the data projects were taken to be a corrective: filling a gap that ought not to exist. Projects devoted to social justice most readily reflexive attempts to illuminate inequalities that were overlooked or ignored. In these instances, the builders hoped that the project would allow the public to see aspects of the pandemic that were otherwise shrouded from view. Despite the implicit

critique at the core of all of the COVID data projects, they had to rely on a foundation of data practices that were beyond their control, and often in the hands of the state – the very entity creating the gap that motivated the COVID data projects.

Most, if not all, of the projects worked with data that was generated from other sources—they were layered on top of other data practices. They developed their own practices to work with this data, but contending with the limits of these practices was a central and common ongoing challenge. At the same time, as some projects became large and popular, they confronted a different sort of problem: working as an unofficial standards body. Here, the challenge was not bumping up against the limits of officialdom, it was recognizing and reconciling the power of standing in place for a public governing body. The choices these dashboards made were consequential: this realization was both empowering and worrying.

COVID data projects worked to analyze incomplete and limited data. A key task for these projects was to work with data sets that were, in one way or another, lacking. They developed novel workarounds for some of these failings—and drove better data in some cases. However, in the end, data projects are at the mercy of their sources. COVID data projects highlight the importance of standardized data across domains and better sourcing. This is true not only in the domain of public health but more generally. Without accessible data sets, the work of data projects will always be limited. *Improving data standardization is an investment in the future success of data projects.*

Many of the builders learned how to interrogate data to learn about what policies created it, and how to understand its limits. Hermione explained:

> I don't know if maybe this is obvious, but I think that for someone just pulling up the page and looking at it, this is... I don't know. This is maybe the truth, this is what happened. And this, I would say this is our best attempt to document what happened…. But states changed the way that they reported things over time. I'm sure that things got lost at some... I'm not... Not wanting to like, you question the validity of the data set. We did a lot of work checking and double checking and everything, but just knowing how much policies change over time, I'm sure that at least one thing may not have been captured. So I think just recognizing that we were not the ones producing the data but merely collecting it, and that we're limited by those that produced it (Hermoine, CDT).

There was a lot of unpreparedness in the data infrastructures of the governments that builders needed data from. But the government also needed resources and time for institutions to react to a novel pandemic.

Most builders felt their government was unprepared to adapt infrastructures based on the particularities of emergent health crises. Systems designed for previous pandemics of seasonal viruses did not work for COVID. Most builders needed better public health data because it was crucial to the projects. For example, it was sometimes hard to simply figure out how many people were actively sick at a given time. This was, in part, because parts of the government released data about the number of people who were sick, but different agencies measured the number of people sick with COVID using different techniques.

Lack of data granularity was limiting for the different projects and for the possible impact of their work sharing data with non-expert publics. And the different COVID data projects encountered this lack of granularity in ways that were specific to the work they were doing. For example, one builder found that using data found that distributing food aid to specific ward-based requests not to the whole city was a best practice, but this required specific data. For all of the projects, the fact that people had to make data (by doing things like getting tested, in the case of Stop AAPI Hate, report hate crimes) made the projects hard to quickly scale and made data integrity issues paramount. One technically-skilled builder suggested that data security and persistence were issues and that blockchain might be a viable approach. Other areas of COVID data builder work, such as work on prison data by The Marshall Project, showed that prisons demonstrated asymptomatic spread. But while people in prisons are members of the community, Maurice observed that we have way more data about baseball than prisons.

Builders on many different projects felt that the most important contribution that their data projects made was the data annotations and metadata that allowed people to understand the data; as some of our participants summarized: the numbers are limited without the curated data. Yet, these projects, particularly those with social justice missions, that sought to highlight the plight of people who suffered disproportionately from COVID – for example, the elderly or people who were racial or religious minorities – found that they were completely limited by the state's decisions about whether or not to collect data

Attachments

Appendix: Project Outputs

Project research findings have been shared widely with communities of interest. To date, we have presented results at multiple academic conferences and in scholarly publications (see below). Additionally, we are currently preparing a book manuscript and proposal for potential publication with an academic press.

Research Presentation:

- "Data Care During Crisis: A Comparative analysis of COVID data infrastructure builders in India and the USA," Society for the Social Studies of Science (4S). Online. 2021
- "Unseeing the pandemic: The making of "missing" data and the Indian state" at Digital India and State Making Virtual Workshop,University of Pennsylvania's Center for the Advanced Study of India (CASI), held on 28-29 June 2022.
- "Unpacking the Worlds of COVID-10 Dashboards Through a Sociotechnical Lens," 6th Nordic STS Conference, Oslo, 2023.
- "Epistemologies of Missing Data: COVID Data Builders and the Production and Maintenance of Marginalized COVID Datasets," Association of Internet Researchers, Philadelphia, October 2023.

Research Publications:

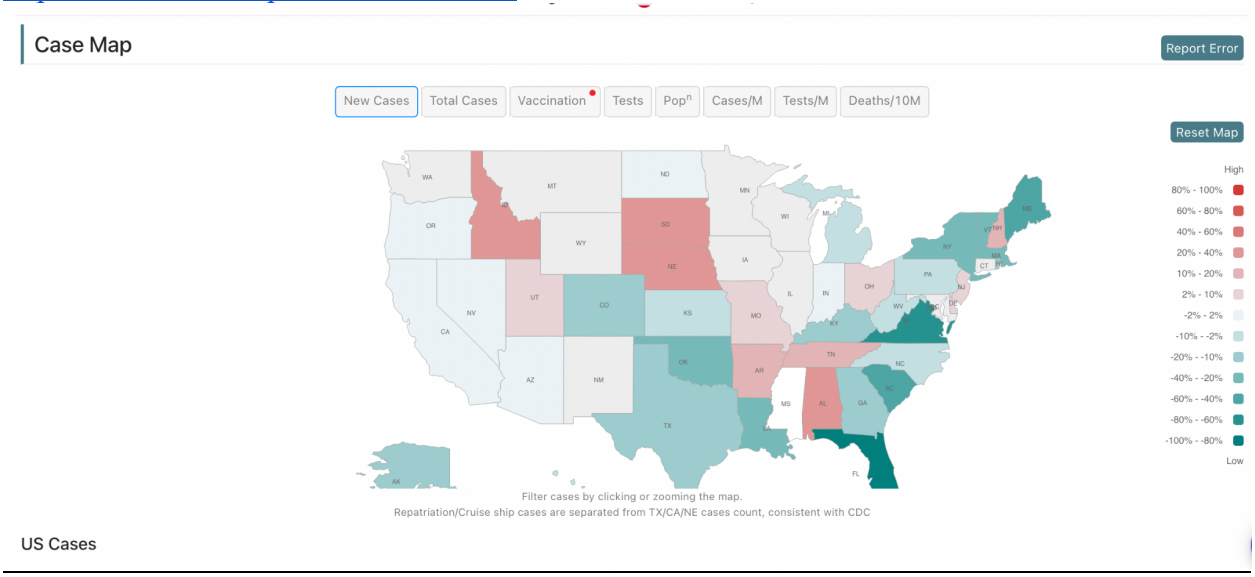
- Youngrim Kim, Megan Finn, Amelia Acker, Bidisha Chaudhuri, Stacey Wedlake, Ryan Ellis, and Janaki Srinivasan, "Epistemologies of Missing Data: COVID Dashboard Builders and the Production and Maintenance of Marginalized COVID Data," *Big Data & Society* (2024) https://journals.sagepub.com/doi/full/10.1177/20539517241259666
- Megan Finn, Youngrim Kim, Amelia Acker, Bidisha Chaudhuri, Stacey Wedlake, Ryan Ellis, and Janaki Srinivasan, "Affective Experiences of Error," *International Journal of Communication* (Special issue on "Sociotechnical Error") (Under Review)
- Megan Finn, Youngrim Kim, Amelia Acker, Bidisha Chaudhuri, Stacey Wedlake, Ryan Ellis, and Janaki Srinivasan, *Untitled Book Manuscript* (In Preparation)

Appendix: Dashboards at a Glance

We interviewed people from a number of different dashboard projects. Below we describe most of the projects that we interviewed people from. We note that our selection of these projects pushed the idea of what could be a "COVID" specific project intentionally to understand the broad ways that people were engaging with data to understand the impact of the pandemic. We have included some images of dashboards, but these are a limited descriptions of the work that these different projects were undertaking or communicating with the publics making use of their work.

1point3acres

https://coronavirus.1point3acres.com/en



1point3acres is the largest Chinese online community in North America that offers a space to exchange information on various topics such as studying abroad, job opportunities, career development, immigration issues, dating, and financial investment. Starting in January 2020, they maintained a real-time, global COVID-19 tracker called CovidNet. The COVID project was founded by Yu (a founding member for the 1point3acres company back in 2009) and Yixin. About 5 paid 1point3acres engineers worked on the project and about 60 volunteers helped as well. In the beginning of the outbreak, data were scattered across numerous sources. They felt the need to integrate the scattered information on a single platform with consistent quality and credibility control. According to one of their published papers, their project was "the only platform providing real-time global case information of more than 4,124 subdivisions from over 27 countries worldwide with multi-language support." In the first two months, they followed media reports and organized them in a spreadsheet. 1point3acres collected data from local health authorities and, when official data was significantly delayed, trusted media sources which were cross-checked with official data afterward. They built an efficient workflow to stream information from media and user generated leads that were fact checked by their data team. Starting in February and March 2020, the public health departments started publishing information that they scraped. The

CovidNet tracker offered full historical case trends around confirmed, deceased, recovered cases. Their interactive visualizations focused on presenting the temporal trends (i.e. epidemic curves -- time series lines and burn down charts) as well as geographical distribution (i.e. epidemic maps, doughnut charts, and tabular views). They updated data real-time through crowd-sourcing and automated data collection which was checked (different from WHO and CDC whose updates were delayed by days). 1p3a didn't use any machine learning techniques such as NLP. Their data was collected via a mix of scripts, API calls, and user submissions. The data team then jumped in and fact checked, cross referencing different sources, de-duping duplicates and ensuring that they presented the most accurate data from credible sources. The original purpose of the project was to create a data website that provides real-time case information in North America to compensate for the delayed official reporting; it became one of the few independent reporting sources of COVID-19 cases across the world. They collaborated with CTP to integrate testing/hospitalization stats and testing locations in the U.S. and their data was used by JHU at times: "Many sites just scrape our data without permission or copy from the public data set published by JHU. JHU world map is widely recognized including the US government but JHU uses our data (US cases). That's why JHU also falls behind us with numbers." CovidNet stopped updating on February 13, 2023, but 1point3acres continues to serve the Chinese diaspora in North America.

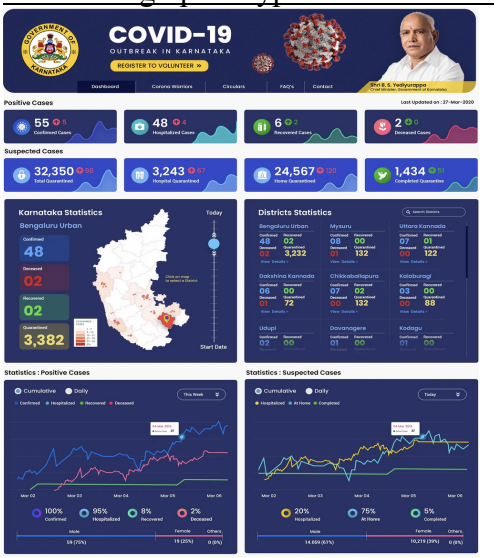
Emory Health-equity dashboard

The Emory dashboard was a nationwide, general population dashboard explicitly focused on equity with county-level data, a contrast to the national dashboards that may have had some equity analysis as a subsection. The dashboard used open-source tools such as Real Simple Maps, Victory, Semantic UI React, and Create React App. We interviewed the PI of the dashboard as well as members of the data team, the communications group and the Georgia-focused dashboard. The team were all faculty or students at Emory University in Atlanta, Georgia. In March 2020 the dashboard was started by the PI, a public health expert on mortality related to cardiovascular disease and diabetes, to understand how COVID-19 was related to social determinants of health. The dashboard went live in June 2020. In the summer of 2020, the team started working on the Georgia focused dashboard with GA Department of Public Health using their data and equity data from the main dashboard – this dashboard went public in early 2021. The project organization was not hierarchical, particularly unusual for its location in an academic setting with students of all ranks paid hourly for their work. But the academic setting also shaped how involved people could be in the project because faculty and students had to manage the demands of their regular jobs and courses. Students drove many of the implementation related decisions and learned "on the job," sometimes discovering and learning technologies for the first time. A PhD student who joined the project about a year into its existence helped to shape the code, documentation, and meeting structure. They had a GitHub site that was internally focused and not public. Communication happened via weekly teamwide meetings and email. Written communication was over email until early 2021 when a PhD student introduced Slack. The PI got initial seed funding from Emory and then substantial funding from the Robert Wood Johnson Foundation for website development. She has gotten additional funding since from Emory and contracts with agencies such as GA Department of Health. In March 2023, Emory posted a note on their website that data was no longer updated due to the "end of the COVID-19 Public Health Emergency and changes to the data pipeline."

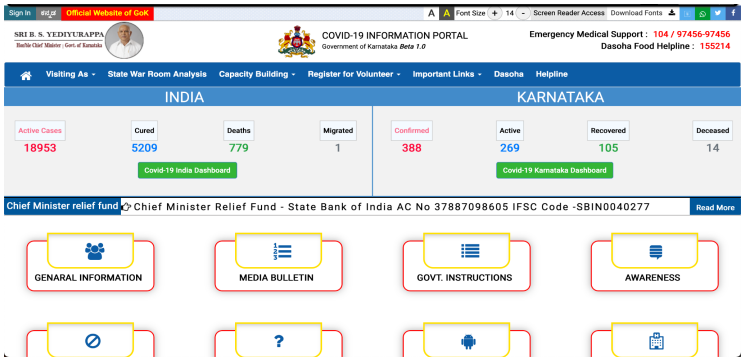
Karnataka State Covid War Room

https://covid19.karnataka.gov.in/English

Initial design/prototype of dashboard:



The dashboard in April 2020:



The Karnataka Covid war room was a prominent state-run dashboard in Bangalore, the "IT capital of India." The War Room, started in March 2020, led many efforts beyond dashboard work, including food relief, migrant transport, quarantining, in addition to building/maintaining a dashboard for internal use and for public consumption. People who worked with the War Room integrated in different functions, focused on both on-the-ground relief work as well as on digital activism. Several departments were involved; the beginning of the War Room and dashboard were led by the Karnataka Dept of Information and Public Relations (DIPR), an agency headed by the popular Principal secretary of DIPR who spearheaded the War Room effort. The War Room worked actively with volunteers outside of the government, involving Rotarians, academic institutes, corporations, prominently, Swiggy, the food delivery platform which provided some software for figuring out how to reach food where needed. Staring in March 2020, DIPR started a Telegram group and a Twitter campaign to answer questions about COVID-19. The physical War Room and later the dashboard came into being in March around the time that the national lockdown was announced on March 24th, 2020. Over time, other departments (Health, Family Welfare, eGovernance, BBMP) and people took over. While some form of the site and the dashboard existed for a while after (at s-https://covid19.karnataka.gov.in/English) – the War Room and the team managing it had changed almost completely by August 2020. We interviewed a variety of War Room workers including volunteers and DIPR employees who were involved in a variety of roles including coders, modelers, designers, infrastructure managers, and food delivery organizers. As a state project, the project had to contend with different state departments as well as different orientations between volunteer groups like the Rotarians and the state employees. Data that was often difficult to obtain so it was gleaned from whatever public office was contactable/responsive in a region. Information was not just disseminated via a dashboard, but also on Whatsapp and Telegram.

WBCAN (West Bengal COVID Aid Network)

West Bengal Covid Aid Network was a volunteer-led initiative that provided access to verified data about Covid-19 care resources available within the state of West Bengal. It emerged from a small WhatsApp group of a few friends who were also doctors/public health professionals and found it difficult to sit idle during the catastrophic second wave of the pandemic in India in April and May 2021. We interviewed 6 volunteers from this project who handled data processing, website development, general project coordination and data verification processed within the project. The project founders said they initiated WBCAN as individuals and healthcare professionals responding to a civic sense of duty to their fellow citizens. Many volunteers engaged with the initiative as a way of coping with their grief of having lost a loved one or having suffered themselves due to resource crisis during the second wave. A seemingly small group of friends recruited a network of volunteers, with the help of social media influencers who amplified their work. The volunteers worked together in Google Sheets and coordinated their work in a number of different WhatsApp groups. Entering data into spreadsheets was difficult for volunteers who don't have access to laptops and only use phones, so WBCAN used Google Forms for data entry. At its peak, the project founders said WBCAN spanned hundreds of people. Social media became the main source of data in absence of comprehensive data published by the state, but because of this unreliable source of information, data verification was intense and tedious involving calling hospitals to verify the number of hospital beds or availability of medicine and oxygen tanks. While it started with medical professionals, as the network rose in scale they also figured the need for different kind of expertise, such as journalists to verify data sources, data scientists to process data, tech developers for better dissemination platforms, legal experts to deal with data fraud and privacy policies. WBCAN attempted to structure documentation and sorting of information into basic categories (hospital beds, oxygen, medicine, food) mainly on Google Sheets. The basic categories of data collection also corresponded to WhatsApp groups. Project members volunteered to take leadership roles in different areas, but there was no formal hierarchical structure outside of the founders who oversaw all the projects. The Google Sheet slowly grew as they covered more districts of West Bengal and created more tabs -- they struggled to improve functionality. Volunteer burnout was a major issue for this project so the volunteers kept recruiting new volunteers to avoid burdening existing volunteers. The entire initiative was conducted without raising or spending any money on part of the volunteers. The data in the worksheets was finally converted into a dashboard that was updated daily, though the project interface was problematic. The people that we interviewed, which included founders and people working on the data infrastructure felt that WBCAN never had enough technical expertise to fully support the project. Since none of them were trained to handle such data projects or crisis situations, they felt that they lost a lot of valuable time just figuring out organizational processes. The project ended at the end of the COVID wave due to volunteer leader burnout, lack of organizational structure, and disagreement among the leadership team about to what to do next.

CODD-K (Collective for Open Data Distribution-Kerala)

Covid19Kerala.info

Kerala, the state that reported India's first case of COVID-19 on 30 January 2020, was home to a group of scientists, academicians and health professionals who started conceptualizing Covid19Kerala.info in March 2020. The online platform mapped the daily trajectory of COVID-

19 and tracks its progression in different parts of the state. Its makers wanted to overcome the limitations of poorly structured data released by the government and generate open and reusable datasets for analysis and visualization. Over 60 volunteers worked on the bilingual citizen-centric dashboard to provide real-time analyses and updates accessible to non-specialists. Besides the website, they periodically submited the datasets collated from various public sources to Zenodo.org for use by researchers, academics and policymakers. We interviewed three members of this team, including the general project coordinator who was the founding member, the head of the data team and the web developer. Data collection and data verification were laborious processes. CODD-K worked in close collaboration with the state governments as their main data source. They deliberated about what data was missing and what the state dashboard should look like. CODD-K fashioned their dashboard after an open source project in Japan because of the work experience of founding members of the team based in Japan, Korea, US and India who were all Kerala natives. The main motivation was to create documentation for the pandemic and their point of departure was the massive flood that hit the state two years prior to that. This is a long-lasting project starting in March 2020 and continued until August (or even more) in 2021. After the second wave subsided in July 2021, they began updating vaccination data and also into data analytics and training. Their work was published in academic journals. The website was last updated on March 22, 2022.

Non Virus Deaths tracker

https://thejeshgn.com/projects/covid19-india/non-virus-deaths/

Table

You can sort by clicking on the column names and you can filter using the search box.

Show 25 entries Search:

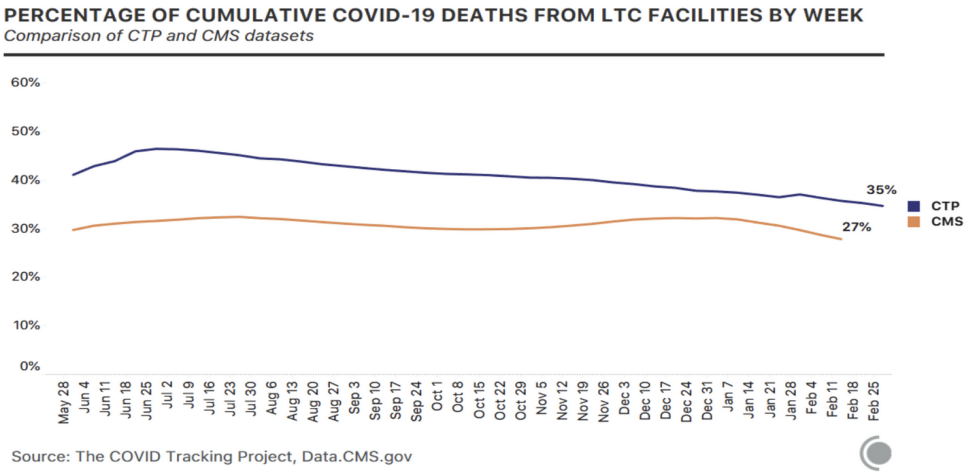
Date	Deaths	Reasons	Location	State	Source	Name	Occupation
2020-07-22	1	Suicide Financial distress Debt	Ahmedabad, Ahmedabad	GJ	ahmedabadmirrordiatimes	Sushi Tibrewal (68)	businessowner
2020-07-22	1	Suicide Financial distress	Khapeyabera village, Giridih	JH	telegraphindia	Binod Roy (18)	business owner
2020-07-18	1	Suicide Financial distress	Kundapura, Udupi	KA	coastaldigest	Nilesh Shetty (31)	business owner
2020-07-16	1	Denial of timely care	Vishwanath Nagenahalli, Vishwanath Nagenahalli	KA	thenewsminute	Shilpa Prasad (28)	sanitation worker
2020-07-16	1	Suicide Financial crisis after being fired during the lockdown	Nalasopara East, Mumbai	MH	mid-day	Ravi Jayaram Shetty (36)	Bar Manager
2020-07-13	4	Suicide and murder Financial stress Loans	Solapur, Solapur	MH	bhaskar	Amol Jagtap (37)Mayuri () Aditya (7) Ayush (4)	Hotel owner
2020-07-11	1	Suicide Financial distress Debt	Berbai village, Tehrauli	UP	amarujala	Premnarayan Raikwar (45)	Migrant labourer
2020-07-11	1	Denial of timely care	Kolkata, Kolkata	WB	newindianexpress	Subhrajit Chattopadhyay	student

Non Virus Deaths Tracker was recording fatalities between March and July 2020 in India that could be traced to the Covid lockdown imposed by the Indian state in late March 2020. The implications of the lockdown were hardest on migrant workers who were stranded without work, earnings, savings or a support system in the province where they worked. With little resources, migrant workers had to return to their home states suddenly in March 2020 after the imposition of the lockdown. Several of the records on the Tracker reflected this. Deaths on the Tracker are categorized as taking place due to starvation and financial distress, exhaustion, accidents during migration, lack or denial of medical care, suicides, police brutality, crimes, and alcohol-

withdrawal. As of July 4, 2020, the tracker recorded 684 deaths. The database was maintained voluntarily by academics based in the USA and India with support from Roadscholarz, a group of freelance scholars and student volunteers interested in action-oriented research, socio-economic rights and related issues. The data on the Non Virus Deaths tracker was compiled manually from newspapers in English and Hindi as well as other languages including Kannada, Marathi, Gujarati, Tamil, Bengali, Odia, Telugu and Malayalam. The small team was alerted to cases by friends and family and on the basis of daily news searches. Volunteers translated the news to enter it into a mastersheet into English/Hindi and checked for duplication of reports/count. The project website said that there may be some duplication despite these checks and also explains the challenges and limitations of relying on news reports alone. Additionally, the categories of death were applied by the people who recorded the deaths based on their interpretation of media reports. The project website had a table that records deaths and displays some details on each death (Incident Location, Date of incident, Number of deaths, Reason of death, Source, Date of publication, Link to publication, Name(s) and age, Occupation Category). A bar graph describes deaths by category (exhaustion, starvation, police brutality. death by walking, deaths in quarantine centres, and uncategorised). This data was made available as CSV, and in open format on Datameet (http://projects.datameet.org/covid19/non-virus-deaths/) using a CC license.

CTP (COVID Tracking Project)

https://covidtracking.com/



The project was started in March 2020 by reporters from the Atlantic, Robinson Meyer and Alexis Madrigal. According to the CTP website and a podcast they created documenting their project, they were working on a spreadsheet and joined forces with Jeff Hammerbacher who was working on a similar spreadsheet. The project team leads and volunteers we spoke to all credited either Erin Kissane or Erin and Alexis as the overall project leaders. The COVID Tracking Project (CTP) included many sub-projects including at least three that reported results for a sustained period during the pandemic. From March 2020 to March 2021, one dashboard reported on data about COVID cases, COVID Testing, COVID hospitalization, and deaths from COVID in the United States. Another dashboard, which started publicly reporting data in June 2020, reported on the impact of COVID in nursing homes and long-term care facilities. The last project reported on the impact of COVID based on race and ethnicity, and also started a few months into the pandemic.

Each of the projects collected data from US states and territories and attempted to standardize the data so it could be compared across the entire country. A bulk of the work for each project was trying to understand how each state reported the specific data of interest, advocate with the states that they open and standardize their data, and to attempt to make data from each state comparable to the data from other states. They took pains to document every single data point and all of the aberrations from their norm. They collected data mostly by hand and used automation tools to check the work of people collecting data by hand. The main rationale for collecting data by hand was to notice the changes in data definitions or reporting irregularities. CTP had over 800 volunteers cycle through their project over the year in which it was operating. Though by number, most of the people working on CTP were volunteers, most of the people who were "team leads" were compensated for their time and some of these team leads worked at CTP as a full-time job. The project listed financial support from the Chan Zuckerberg Initiative, Emerson Collective, The Rockefeller Foundation, Robert Wood Johnson Foundation, Patrick J McGovern Foundation, and Beneficus Foundation. CTP also received software donations from Netlify, Slack, Contentful, Calibre, Algol, Donut, Tableau, Headway, Mapbox, and 1Password. We interviewed 23 CTP participants (8 of whom were listed as "team leads" on the project website) in March, April and May of 2021. One of our project team members (who did not lead interviews) was also part of an advisory board for the CTP archive at UCSF. CTP team communication took place on Slack — a familiar tool for those who had experience in tech and new for many others. They integrated Slack automations with air table to create complex workflows. Slack was also an important place to socialize. COVID data for all of the projects was collected with the interface of a Google Sheet for all projects. However, the main project focus on cases, testing, hospitalization and death had a Postgres database behind it. They used software tools such as Github to make their code available and prioritized accessibility in their data distribution and website, which meant avoiding many of the fancier dashboard interfaces that were not accessible (or easily preservable). Since the focus of the visualizations was accessibility, most of the data was publicized as tables as well as graphs. The dataset was accompanied by extensive notes about irregularities in the data or decisions CTP made about how to best count.

Johns Hopkins University's Coronavirus Research Center

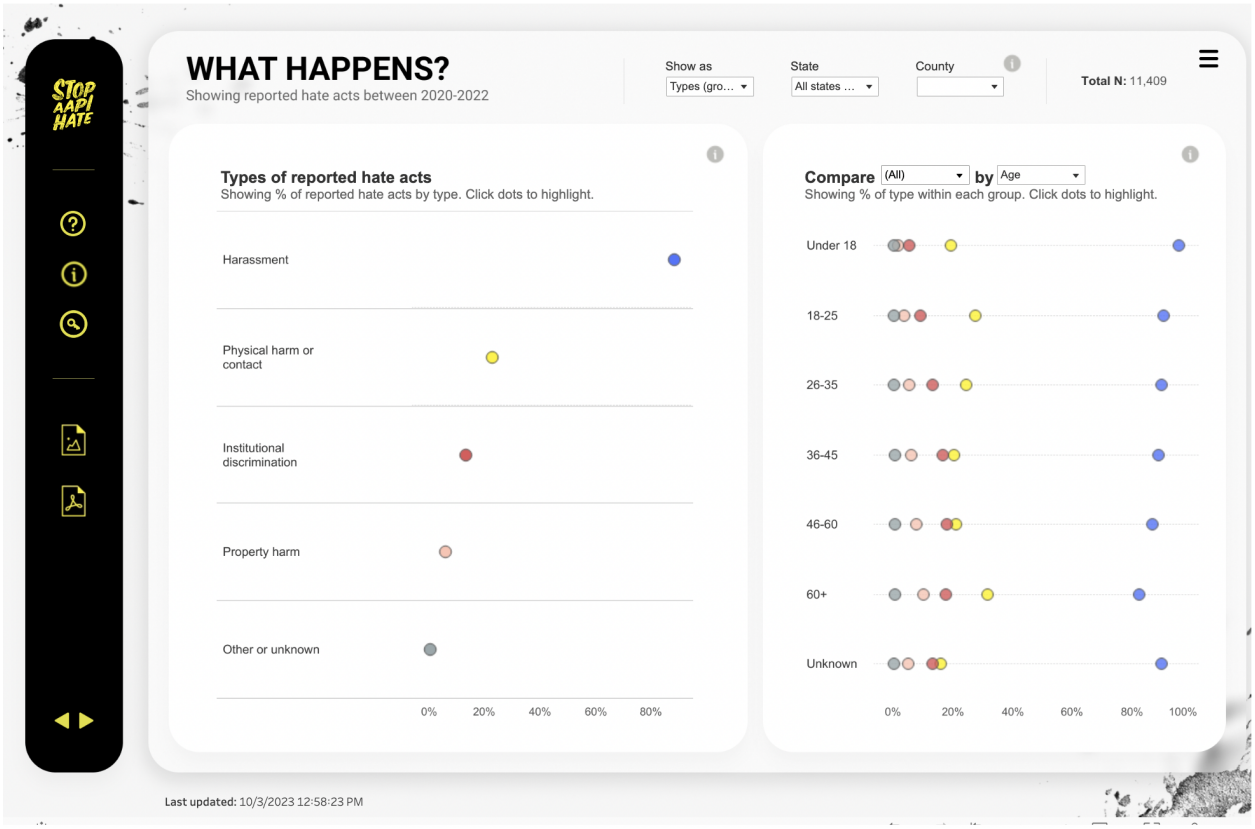
https://coronavirus.jhu.edu/

Johns Hopkins COVID-19 dashboard launched during the early days of the pandemic in January, 2020. Initially, the project began as the work of Lauren Gardner (then an Associate Professor of Civil and Systems Engineering) and Ensheng Dong, one of her PhD students at JHU's Center for Systems Science and Engineering. In media reports, Dong said he was driven to create the dashboard in part due to an early and keen personal interest in the unfolding pandemic—his family living just outside of Beijing was providing firsthand accounts of the spiraling catastrophe while many in the US still only had a hazy understanding of what was happening. The initial map used ArcGIS (Esri would support JHU's work) and data manually culled from international sources. Eventually, the map morphed into a larger operation, the JHU Coronavirus Resource Center (CRC), that included a US-focused map, contextual and interpretive articles and commentary, a helpdesk, and other resources. The larger set of offerings relied on the work of the interdisciplinary JHU community, including the Applied Physics Laboratory (APL), a frequent government contractor focused on providing expert analytics for federal sponsors, and the JHU Center for Civic Impact. APL provided key resources, including novel data scraping

tools that had been developed in-house before the pandemic, and staff. APL managed the data collection and data pipeline for the project, blending automated collection with ongoing manual review of data collected from over 250 different sources. At the beginning of the pandemic in the USA, JHU's COVID project became one of the largest and most popular COVID data resource in the world. It was routinely spotlighted in media outlets across the globe as an authoritative source. At its peak, it received over 18 billion views on March 20, 2020. JHU's work was supported by the University and outside funders, such as Bloomberg Philanthropies. Unlike many other dashboards, JHU's work was largely supported by paid full-time staff members. Users also contributed comments, data, and edits supplied via GitHub. JHU stopped collecting new data on March 10, 2023, effectively shutting down the project. We interviewed 9 participants on the JHU dashboard including project leadership from the CRC. Our JHU colleagues shared their interview transcripts with university legal teams before they decided to give us permission to use them. One interviewee dropped out of the study after the interview.

Stop AAPI Hate

https://stopaapihate.org/

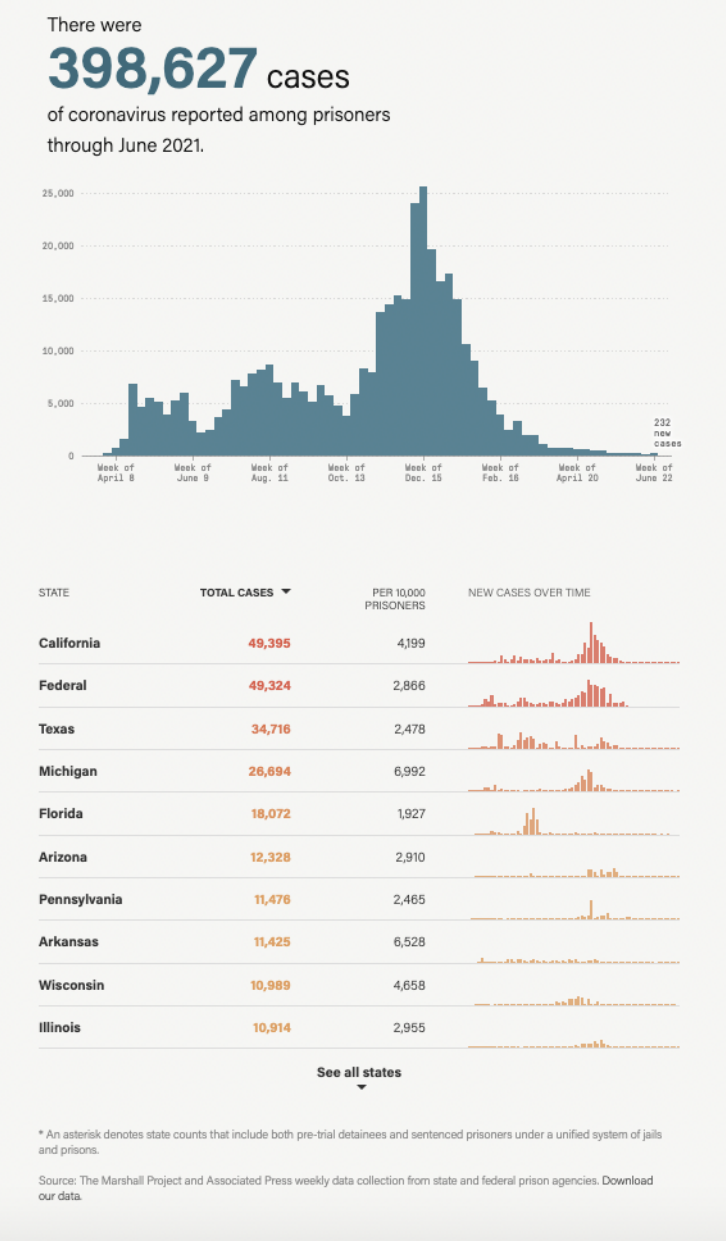


Hate and discrimination against AAPI communities intensified during the COVID-19 pandemic, fueled by President Trump's assertion that COVID was a "China virus" and some US citizens' animosity towards all Asian Americans. In response to the alarming escalation in xenophobia and bigotry a coalition of three AAPI related organizations, Asian Pacific Planning and Policy Council (A3PCON), Chinese for Affirmative Action (CAA), and the Asian American Studies department of San Francisco State University launched the Stop AAPI Hate website on March 19, 2020. This

website aimed to track, raise awareness, and respond to incidents of hate, violence, harassment, discrimination, shunning, and child bullying against AAPI in the U.S. The site asked those who experienced or witnessed acts of hate towards AAPI communities report incidents in a Typeform that was available in 11 languages (English, Mandarin, Cantonese, Korean, Japanese, Vietnamese, Tagalog, Thai, Khmer, Hmong, Punjabi, and Hindi). The form did not collect data about perpetrators because they are focused on centering AAPI experiences and did not share data with law enforcement without consent. They had five overlapping goals: (1) Serve as the leading aggregator of anti-Asian and anti-Pacific Islander hate incidents; (2) Offer multilingual resources for impacted community members; (3) Provide technical assistance from rapid response to preventative measures; (4) Support community-based safety measures and restorative justice efforts; and (5) Advocate for local, state, and national policies that reinforce human rights and civil rights protections. The three project leaders were awarded Time 100 Most Influential People 2021, and were instrumental in winning legislative victories in California addressing Anti-Asian hate. The project recruited 20-30 student volunteers (unpaid undergraduate and graduate students in SFSU, UC Davis, UC Berkeley, and other universities) and hired one paid research assistant. We interviewed one of the project leaders. The project provided a brief summary of descriptive stats/trends on the reports tab as well as longer reports. They published press releases and statements detailing specific actions that should be taken, practical solutions and policy recommendations based on analysis of the self-report data, but did not distribute the data that they collected without analysis.

TMP (The Marshall Project)

https://www.themarshallproject.org/2020/05/01/a-state-by-state-look-at-coronavirus-in-prisons



The Marshall Project, a journalism non-profit who reports on the US criminal justice system, focused on COVID in carceral facilities. The project began in March 2020, after a couple of weeks of pilot testing. The project concluded in June 2021. A TMP senior editor and data journalist were the main leaders of the TMP data project. They managed the whole project and communicated with the journalists who were enlisted to do the data entry work and communicate with prison PIOs. Around 25-30 paid journalists in total were involved in this project. Journalists came in and out of the project based on their working schedule. At any point, there were approximately 18-20 journalists working on the project. Among them, approx. 6 of them were journalists from AP,

which joined the project in April 2020 while the rest of the journalists were from TMP. Based on their experience of writing investigative reports on the criminal justice system, incarceration, and prisons, which particularly lack transparency, they knew from early on that there won't be a good authoritative source of COVID information within prison systems. So they started this data project expecting that it would become a really important source of information that shows how vulnerable populations who have difficulty maintaining social distance are affected by the pandemic. Initially, they recruited 15-16 journalists within TMP to work on data entry while the project leaders did the analysis and presentation. Later, to make this project more sustainable, they partnered with Associated Press, where the project leaders had some colleagues with whom he worked in the past. Roughly, around 20 journalists between 2 newsrooms were collecting data every week – the journalists were familiar with each other, their shared professional norms, were paid for their work, and were very organized. Their shared infrastructure, called Klaxon, an existing infrastructure. TMP worked hard at standardizing data that were released in different formats and with different data definitions. Many times, they had to do the math themselves to make numbers comparable between states. Not only was it challenging to harmonize data from different states, TMP encountered dubious data practices by the state and federal bureau of prison that made it impossible to know the accurate cumulative number of cases who were ever infected in prisons. The project closed in June 2021 knowing that there are similar dashboard projects working on collecting this data (UCLA Behind the Bars, COVID Prison Project) so that TMP could focus on investigative reporting. We interviewed 4 participants from TMP including the project's senior editor and three staff writers.

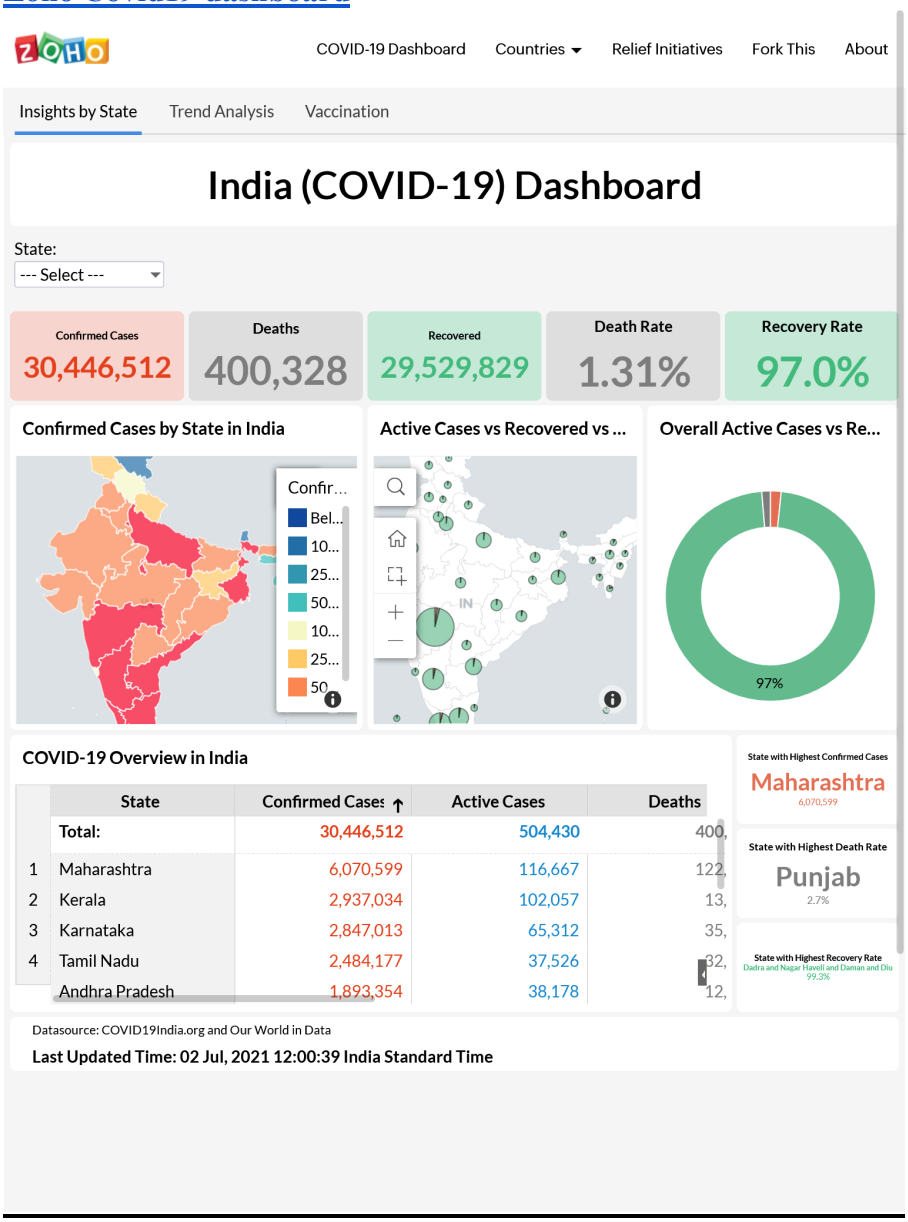
CAN (Covid Act Now)

https://covidactnow.org/?s=28511730

Covid Act Now "CAN" was an early visualization dashboard that competed with, though had less media coverage, than Covid Tracking Project in the USA. CAN published the first version of a model of COVID spread prediction on March 20, 2020 in consultation with university affiliates from Georgetown University Center for Global Health Science and Security, Stanford University Clinical Excellence Research Center, and Harvard Global Health Institute. Using open source tools, CAN published a "risk" map of disease and response, it also allows users to access data from both the website and an API. We found that CAN's iFrame embed feature was being used in other local news dashboards and visualization resources, such as the Navajo Times. CAN has a county level iFrame "sharing" feature for all of its trends, curves, and plots, so users could easily share, curate, or build a data pipeline. Their website reports they have served over "15m website users" and have "200k subscribers" on COVID news. They also said they worked with "more than 100 federal, state, and county officials as well as numerous multinational corporations and NGOs to develop data-driven COVID responses." CAN described their team as having a variety of relevant expertise. There was a CEO and a COO. There was a team of about 30, many of whom have volunteered at CTP and are taking time off of college; others work for large technology companies. We interviewed four people involved with CAN, one with expertise in science communication and the others paid operations, strategy, and partnerships team members. CAN was being run like a start up by engineers that ship "data products." They seemed to be targeting future US/non-profit partnerships in the tech and governance space. No engineers would agree to an interview, but there appeared to be a lot of emphasis on the reuse and access of the data. The Act Now Coalition is a 501(c)(3) which went on to focus on Rewiring America, a green energy initiative.

Zoho

[Zoho Covid19 dashboard]

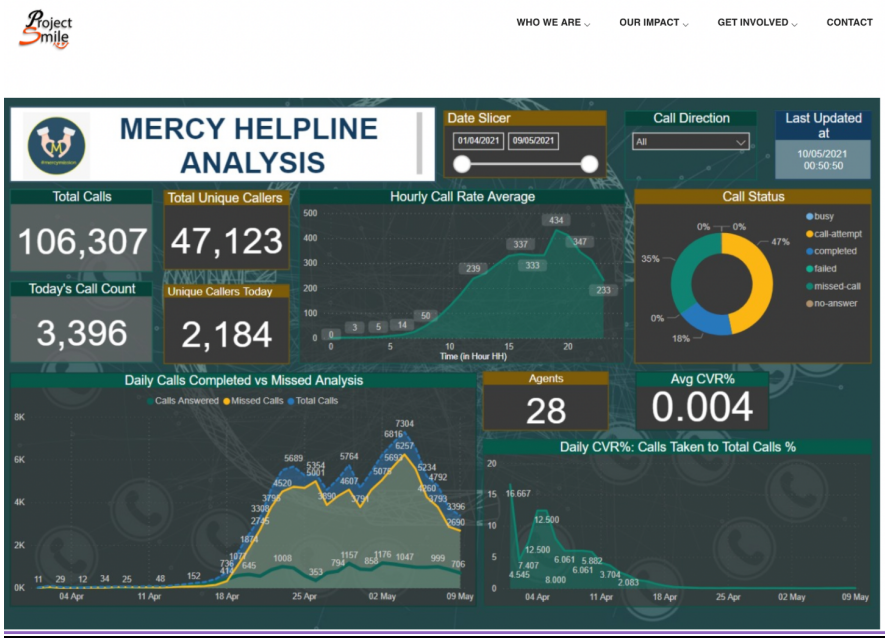


The Zoho COVID-19 dashboard reported on data about the COVID pandemic from around the world starting in March 2020. The project was built and managed by a corporation, Zoho, a software company that was started in India. The dashboard was a side project that was built on Zoho's existing proprietary analytics product. It was built mostly by people from the Zoho Analytics group (Zoho Analytics was a Business Intelligence product), especially its product management and marketing teams. It was common for Zoho's Business Intelligence analytics product to publish a featured analysis of a global event. Zoho employees noticed many people searching for COVID data – this motivated the team to build dashboards, their expertise, for COVID data. They wanted to make an easily accessible portal using Zoho's products where users could gather comprehensive insights about COVID globally. We interviewed four people from Zoho including people from marketing, product management, data analytics, and business

intelligence. While the user interface that Zoho used was fairly stable, finding data sources was a constant challenge and they had to drop sources that were not being updated regularly. Once they identified data sources, they found a lot of heterogeneity and realized that they needed to focus on trends rather than on real time numbers alone. They eventually used Worldometer for real time data and ECDC (European Centre for Disease Prevention and Control) for trend analysis. Over time, they also dropped some data sources when they found the data sources weren't getting updated regularly. Zoho learned a lot about at what level of granularity data was reported and how regularly countries updated data; these lessons shaped the decisions about what to publish, what countries to highlight etc.The dashboard is still active but the underlying source data, WorldoMeter Coronavirus Tracker stopped data updates on April 13, 2024 "due to the infeasibility of providing statistically valid global totals, as the majority of countries have now stopped reporting.

Mercy Mission

[Mercy Mission's Covid Relief work]



Mercy Mission was an umbrella of 25 Non-Governmental Organizations (NGOs) in Karnataka, including Mercy Ambulance, Mercy Helpline, Mercy Oxygen, and Project Smile that worked with marginalized communities. These organizations decided to combine forces during COVID to provide relief efforts. Each initiative had a volunteer team and they coordinated across these different teams. Individuals from some of these non-profits had started discussing this idea of bringing together an emergency taskforce and creating a network in Dec 2019 as they started hearing about the possible spread of COVID. Their idea was to bring together volunteers from different walks of life together: Bangalore coming together to deal with this emergency. Eventually, several non-profits in Bangalore (including Project Smile, whose volunteers we interviewed) with different strengths and who catered to different regions/demographics came together under the banner of Mercy Mission. Of the NGOs that participated, some were faith-based and others were not. Each of these organizations had a history of working with communities in specific regions of the city. They knew these populations well

and were trusted by them in return. Because of this, Mercy Mission was able to reach populations and understand needs that were inaccessible for other groups (such as state-sponsored ones). During the Lockdown, Mercy Mission worked on awareness, helping with last rites, food provision, ambulance and hospitals. We interviewed the director of one of the non-profits that was part of the coalition of NGOs working for COVID-19 relief in Bangalore and one of the volunteers. The initiative was running multiple operations including oxygen centers, ambulances, coordinating last rites, and taking calls for food provisions or from people with symptoms. Data was critical and was centralized. But the last mile operations were handled by the non-profit volunteers that best knew those regions. The activities undertaken by Mercy Mission evolved over time, as the needs dictated by the pandemic did (lockdown, migrants traveling, quarantining of travelers…). The two waves of COVID also raised very different demands. During the Lockdown, Mercy Mimssion worked on awareness, helping with last rites, food provision, ambulance and hospitals. We interviewed the director of one of the non-profits that was part of the coalition of NGOs working for COVID-19 relief in Bangalore and one of the volunteers. The project got donations through Corporate Social Responsibility arms of IT companies and individual donations. The dashboards were last updated on May 10, 2021, but Mercy Mission's work as a non-profit continues.

Appendix: Interview Sample

During March 2021- November 2021, we conducted 79 one-to-two-hour interviews with COVID data infrastructure builders across 17 COVID data projects. The interviews were coded and analyzed. We also inferred a number of different characteristics from the COVID data builders to describe the pool of people that we spoke to in more detail:

- Were the COVID builders we spoke to employed? We asked our interviewees whether they were being paid to work on the COVID data projects
    - 36 people were paid to work on their COVID data project (either as part of their jobs or as employment)
    - 42 people were volunteering their time to work on the COVID data project. Of the volunteers,
        - 25 had other jobs
        - 11 were students
        - 6 didn't have other jobs
- What industry were the COVID data builders part of? During the interviews, we identified the primary industry that out interviewees came from:
    - 26 people were from academia
        - 15 were part of academia in permanent staff, research, or faculty roles
        - 11 were undergraduate or graduate students
    - 4 people worked in healthcare (broadly)
    - 20 people were in high tech
    - 7 people were in communications or journalism
    - 4 people were in not-for-profit work
    - 8 people had jobs in other parts of the private sector
- What did the people we interview do on the COVID data project? Though many people we spoke to often had many different roles, we identified our interviewees as primarily having the following roles:
    - 2 were primarily involved with providing healthcare knowledge
    - 4 interviewees primarily engaged with activism and advocacy
    - 4 were primarily involved with science communication
    - 8 primarily did data analysis
    - 20 primarily did data collection/input/organizing
    - 13 primarily did infrastructure/engineering work
    - 24 were primarily doing some kind of project/program management

Appendix: Coding guide

We developed the following codebook to facilitate analysis of our interview transcripts. We include it below to give a sense of our approach to analysis and because it might be useful to other projects.

Procedural notes:

General rule: Being over-inclusive is better than missing something out!

Segments to Code: Code the entire question and the entire answer (not paragraphs or sentences, unless as noted before). Add all the codes that fit. Include interjections by the interviewer (ex. "Yeah," "Awesome," and other affirmation comments) in the same block until the interviewer poses the next question (including clarifying questions).

WOW and Follow-Up: These codes—and these codes only—should be at the sentence or paragraph level. These are the only exceptions to the note above re Segments to Code.

N.B. The quotes below are often only *hypothetical illustrations.*

Code Name (alphabetical order)	**Description & Examples**
Affective experience	Discussions around burnout, depression, feelings about the epidemic, feelings about the project. This includes conversations about how the interviewee felt or the emotions of others (including coworkers felt). This includes discussions of the work of managing others' affective experience (e.g. emotional labor, affective labor, relational labor). Any reflections around both positive and negative emotions that interviewees or other volunteers felt go in here. (This is an actor category; sometimes tricky if interviews are emotional/crying but it doesn't show up in the transcript/words.) EXAMPLES: 1. "It was depressing" 2. "I felt overwhelmed." 3. "We had a lot of burnout during the first few months." 4. "People were really working to their limit—stress was high." 5. "It was rewarding" 6. "I was impressed by what the project put into place" 7. "There was lots of energy in the room. Everyone was so eager and fearless to get things done."

Audience/Users/Impact	Interviewees describe the project's potential/real impact and the audience they speak to. This includes discussions of the work it takes to create products of all kinds for external users/audience, including reports. This also includes the interviewees' imagination of how the audience will or is using the dashboard. Also include the project's actual interaction with the audience. (e.g. managing help desk; responding to github tickets). This does not just have to be about connecting with users over data (e.g. could be talking about delivering meals or writing blog posts). This does **not** include the project's personal impact on the interviewee. EXAMPLES: 1. "Governors were calling us for updates!" 2. "My mom used the dashboard." 3. "On Github, we could see thousands of people pulling our data." / "People found errors in the dashboard and reported tickets on Github. I had to resolve those." 4. "People complained—they always do—about how we displayed the data. What can you do?" 5. "We were thinking about how local reporters would use our data in their stories." 6. "The states were using our dashboard as a scorecard"
Background, Expertise, & Networks	The interviewees discuss particular expertise or existing skills. Include discussions of their work as paid employees during/prior to COVID project. Also include social "installed base" -- individual/project's relation to external institutions (tech/media companies, universities), social networks, civil society, how people draw on their existing social networks. Any discussions around an individual's background should go here. Also includes the larger background of the organization (ex. "We were funded by the Atlantic"). EXAMPLES: 1. "And also I've been part of the industrial research for television and all that, media organizations and I am media and content and communication. I've been working as a consultant as well. For like two, three years back, actually, I left the mainstream media and I started my own venture." 2. "I got involved in this project through a capstone course in my university."

	3. "I've been working at DataCenter for more than 20 years. DataCenter is a big organization. I think we're somewhere in the neighborhood of 8,000 people within the lab. We're fairly large." 4. "You should talk to Sharon -- she is really a big part of this."
Challenges/Problems	It captures moments of frustration or difficulty ("It was hard"). These are the moments where interviewees talk about the barriers, hardships, or challenges they had to overcome (or note). This is an **actor** category, only code it when the interviewee identifies something as a challenge (both to the individual and to the project). EXAMPLES: 1. "The data was provided in a nice image, but we could not harvest it. We eventually figured it out, but it was tough." 2. "The data was so messy. We had to come up with a way to make it make sense." 3. "The long hours were a grind. It impacted our ability to get the work done." 4. "Coordination was tricky. How to get all people on the same page during the pandemic was tough."
Data	"Data" is the General/Parent Code. The code is split into the following four sub-codes: Data Collection; Data Validation; Data Dissemination; Missing Data. **Do not use the parent code.** **Sub Codes** **Data Collection**: Discussions of data sources ("Where we got it"); and the process of collecting data ("How we got it"). This often describes government sources, but may describe other dashboards, press conferences, newspaper articles, or any other sources for data. Discussions can be positive or negative (i.e. examples of trying to collect data but it not being there). EXAMPLES 1. "New Jersey was always a problem…they did not list probable cases" 2. "We scraped data twice daily"

	3. "My job was to watch the daily press conference and get the updated figures." **Data Processing**: Discussion relating to how data was processed inside the dashboard/project organization. This includes activities like cleaning, validating, standardizing, and storing. EXAMPLES 1. "We double checked the numbers before pushing them out." 2. "We used automated tools to flag anomalies." 3. "One challenge was reformatting the data into a usable format." **Data Dissemination**: How the data was made available or shared with the public. This includes visualizations, the output of the dashboards, posting of data files on twitter, sharing data via Github, creating explainers and other activities to contextualize the data, writing blogposts and reports. This is about what the dashboard/project is disseminating (NOT what the gov't is disseminating). Often double coded with audience/impact. EXAMPLES 1. "Github data posted at midnight every night." 2. "I created a YouTube video to ensure that it was understood." 3. "My job was to make sure that the notes were clear on the visualizations." **Missing Data**: Discussions of data that was not available or that they could not find. This includes discussions of when certain sources did not collect figures on certain cases. EXAMPLES 1. "Virginia never had numbers for likely cases." 2. "Correctional facilities always undercounted…their numbers were not credible" 3. "We just did not have the data."
External Factors	Codes capture events, other organizations particularly the government/regulation/administrative state, or strategies that occurred externally but impacted the organization's workflow. EXAMPLES: 1. "The states were constantly changing their data standards. There was no consistency among them"

	2. "The government was disseminating data in pdf format so we had to scrape them manually" 3. "NHS is a good example and the UK is an excellent example. So we would look at that and like, 'Okay, oh my god. If only our governments could do something like this, it could have been pretty easy for us in terms of effort.'"
FOLLOW UP Tag	Do much more research on the referenced object, thing, resource that is mentioned by the interview. Useful for organizations, technical concepts, political issues that may not be well understood by the coder. Even if the individual coder has some knowledge, if the coder thinks the group as a general would have difficulty understanding it, code it as FOLLOW UP. This is coded at the sentence level or less.
Inequality	Captures discussions about inequalities in how the virus impacted different groups (not just in terms of illness but also hate crimes or shutdowns), different types of access to data, etc, inequality in the virus' impact on different countries. Includes data inequality, language barriers in reading the dashboards, and other accessibility issues. Actor/interviewee describes when they understood themselves as confronting inequality or when the project/data reacts to, addresses, or produces inequality. This is not just an actor category, but also an analyst category. EXAMPLES: 1. "Then the lockdown started and a lot of senior citizens who could not afford medicine." 2. "We needed to show that bad things happen with communicable diseases in nursing home settings." 3. "What do we mean by disparities? How do we illustrate that?" 4. "We needed the information to see how vulnerable people, incarcerated populations, who are incapable of social distancing are affected by COVID." 5. "There was a heightened sense of anti-Asian sentiment around those months. My friends and I were very concerned" 6. "We made our dashboard available in 10 different languages so that ethnic communities could receive information"

Lessons Learned	What the project or the individual learned from this experience (Big picture takeaways and individual lessons). Captures the interviewees' reflections about "take-aways," "best practices." Includes positive and negative lessons. This is an actor category, NOT an analyst category EXAMPLE: 1. "One key insight…we need better standards." 2. "I guess one thing I walk away with is pride in our volunteers." 3. "Now, I want to focus on public health in my next role."
Money	Discussions of project funding, costs, salaries and financial flows generally. Include any conversation around $$ as it pertains to the project management. ● How to code "not getting paid": volunteers as a whole not getting paid is coded under "project org," not "money." However, if individual interviewee comments about frustrations over not getting paid, code it as "money" EXAMPLE: 1. "We started to get donations from our audience." 2. "The company 1p3a was behind this dashboard project so they managed the server costs." 3. "We paid volunteers $12/hour." 4. "Airtable gave us a steep discount"
Motivation	The interviewee describes a reason, value, incentive or experience for creating or joining and staying with the dashboard as a volunteer, leader, employee ("**Why I do it").** This also includes motivations of the larger project (**"Why we do it"**). EXAMPLES: 1. "It was important. I had to do something." 2. "We tried to fill the gap—the CDC was doing a terrible job." 3. "The goal of the project was to provide people with information they could use for their daily decisions" 4. "We wanted to inform policymakers and others."

Project	**<u>Do not use the parent code.</u>** **Subcodes:** **Project - Org Structure:** The roles and responsibilities of different people in the organization; training; discussion of volunteer work; all discussions of hierarchy or lack of hierarchy/leadership and lack of leadership. It's okay to double code org structure and culture when confused. EXAMPLES: 1. "The work was paid for everyone who made big contributions." 2. "Joanne came on and started to do a lot of management of the volunteers." 3. "The hardest part of the project was managing the volunteers" 4. "We would always have two people validate everything manually before pushing it out." **Project - Culture:** includes internal project communication, the culture of the project (e.g. emojis, transparency, vibe/atmosphere, general protocol/rule of thumb, shared responsibility, shared affects etc). It's okay to double code org structure and culture when confused EXAMPLES: 1. "We shared a sense that this was important." 2. "I got on Slack to chat with people and see what they were up to" 3. "We used emojis to signal that we had completed a task" 4. "Each morning we started with a Zoom call." **Project - Evolution**: includes project origin stories, project endings, and discussions of "middle" phases. EXAMPLES: 1. "Early on, the project wasn't centralized." 2. "The second generation of the website came out in March" 3. "We haven't worked on it since that data came out" **<u>Project - "What I did":</u>** The interviewee describes their day-to-day work; how many hours they worked per day; work routines; managing schedules between full-time work and volunteering work. This is strictly *what I did*, not what we did.

	(ex. When management people talk about what the project was doing, code it "project org") EXAMPLES 1. "I was in charge of collecting data" 2. "I worked about 70-80 hours a week on this project…" 3. "I created the training questionnaires for volunteers" 4. "I was the go-to person whenever volunteers were facing challenges."
Software Tools	Includes automated tasks, bots, and other technical solutions to data capture/project organization - many of the interviewees discuss a move to automation, when tasks could be deployed by scripted bots or routines. Discussions of visualization software, google sheets, databases. Also discussions of social media, uses of Slack and other communications tools. EXAMPLES: 1. "We used Slack for internal communication" 2. "Initially, we used Google sheets to collect data. But later we shifted to AirTable." 3. "We published our data on Github" 4. "State prisons purchased Microsoft BI and began to create really elaborate visualizations that became hard to understand." 5. "We didn't consider building an automated scraper because data was released in different formats."
Virus Moment	Aspects of the virus events. Includes the temporality of the virus (like when the virus was bad in NY in March/April 2020 or in India in spring of 2021). EXAMPLES: 1. "Vaccines started to roll out since November 2020 so we began collecting this data" 2. "At these initial months, there was a lot of uncertainty about the virus and scientists/doctors were also trying understand what's going on"
WOW Quote	Everyone needs to see this quote, it explains things. Especially code mentions of other dashboards (ex. "We referenced JHU dashboard a lot"; "CTP dashboard was kind of the standard for other dashboards"; "I really appreciated the Stop AAPI Hate project") This is coded at the sentence level!