COVID-19 India Data

Parsing Detailed COVID-19 Data in Daily Health Bulletins from States in India

Homepage
ibm.biz/covid-data-india

GitHub
https://github.com/IBM/covid19-india-data

IndoML Hackathon
ibm.biz/covid-indoml

Whitepaper
arxiv.org/abs/2110.02311

Contact
mayank.agarwal@ibm.com
tchakra2@ibm.com

IBM Research

Availability of COVID-19 data is crucial for researchers and policy makers to understand the progression of the pandemic and react to it in real time.

Despite pleas from researchers in India for the urgent access to COVID data collected by government agencies, such data is not readily accessible in structured form.

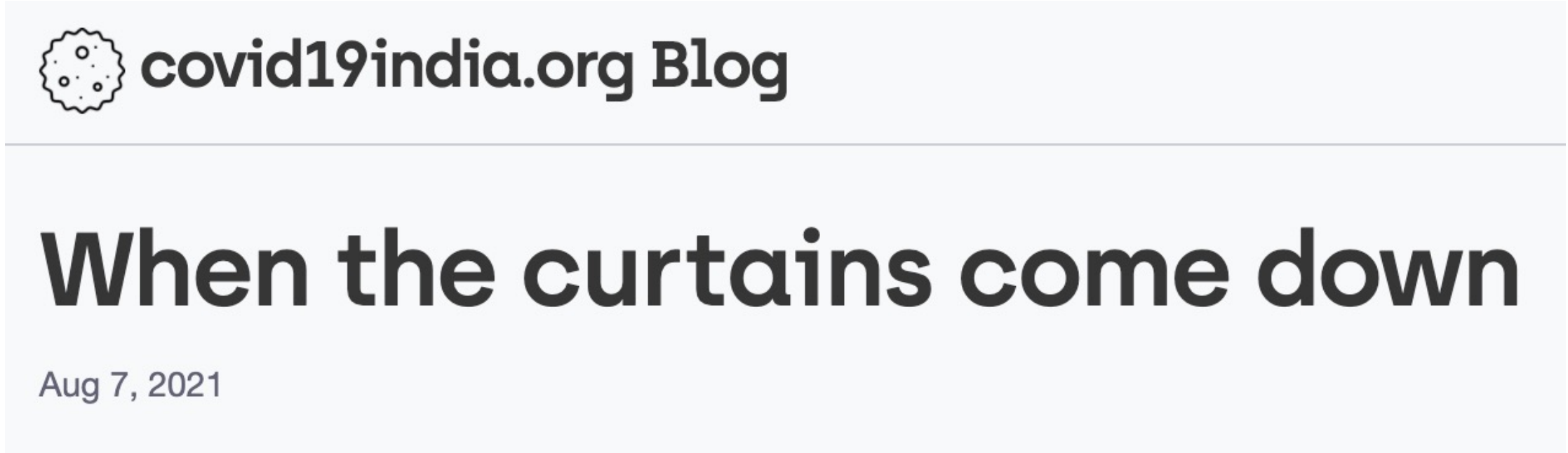
While there are fantastic crowd-sourced efforts underway to curate such data, manual approaches cannot scale to the volume of the data produced over the long term.

SCIENCEINSIDER | ASIA/PACIFIC

‘There are so many hurdles.' Indian scientists plead with government to unlock COVID-19 data

Open letter asks for better access to genomic and epidemiological databases

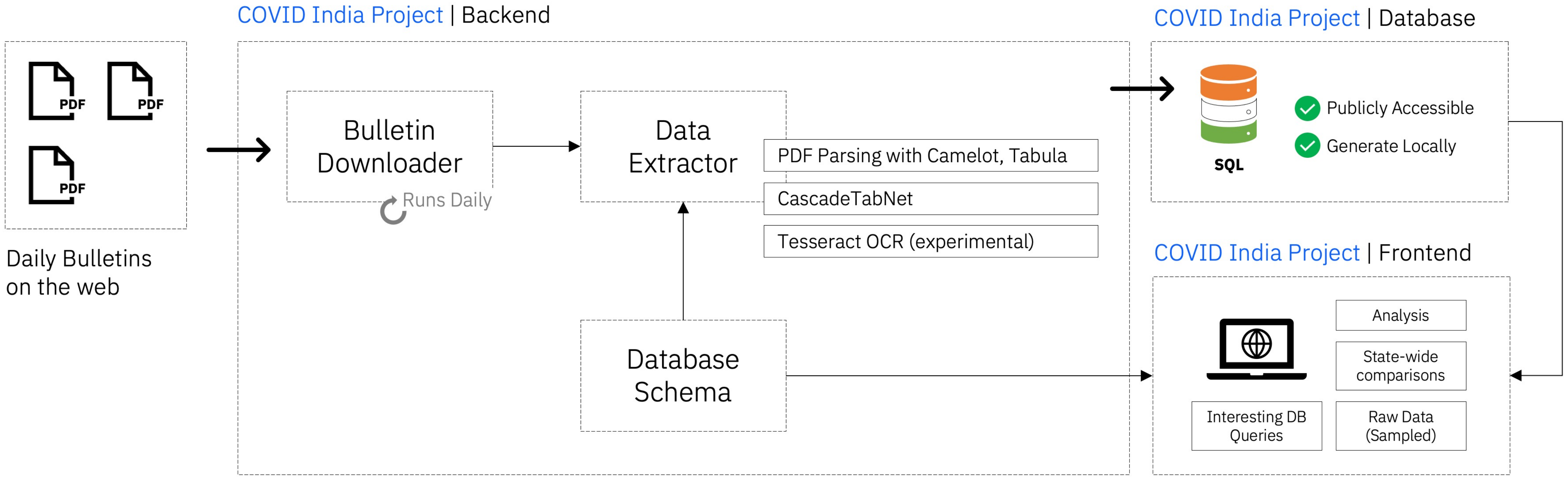
4 MAY 2021 · BY PRIYANKA PULLA



covid19india.org Blog

When the curtains come down

Aug 7, 2021



Individual state governments in India publish daily health bulletins containing the state of the pandemic in that area.

Our system runs automatically every day to fetch newly released bulletins (in the form of PDFs), parse them, and push the new data to a publicly accessible SQL database.

The system is fully automated and thus does not require manual upkeep.

This also means that we can extract and curate data at much larger detail and volume than manual volunteer driven efforts.

Classic PDF Parsing We use Tabula and Camelot Python libraries to extract a substantial amount of the data.

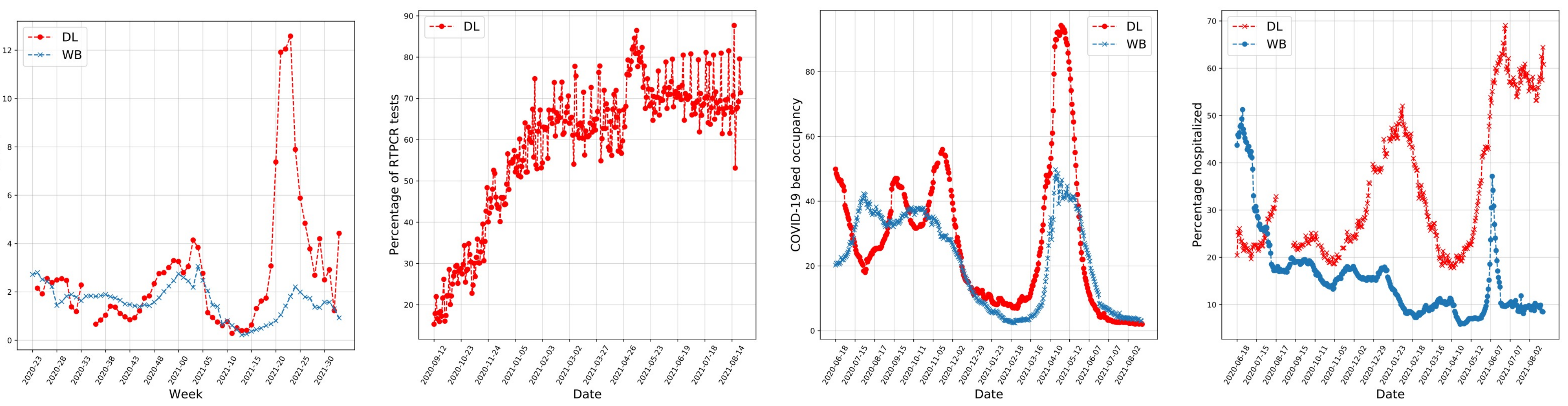
Deep-Learning based PDF Parsing When the above fails, we use CascadeTabNet, a state-of-the-art CNN that identifies table regions and structure, to increase parsing accuracy.

OCR Not all information is embedded as data tables in PDFs. We are experimenting with open-sourced OCR techniques to extract information from images inside PDF bulletins or published on social media.

Automated document extraction allows us to access detailed district level case and vaccination data, vaccination efforts, age-wise and gender-wise distributions, and for certain states, details of individual cases.

More details in the whitepaper.

Dataset (→) **Category** (↓)	covid19india.org	Ours										
		DL	GA	HR	KA	KL	MH	PB	TN	TG	UK	WB
Case information	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Testing	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Vaccination	✓	✓	-	✓	✓	-	-	✓	-	-	✓	✓
Hospitalization	-	✓	✓	✓	-	✓	-	✓	✓	-	-	✓
Individual fatalities	-	-		-	✓	✓	-	-	✓	-	-	-
Age/Gender distribution	-	-		-	-	✓	-	-	✓	✓	-	-
Mental health counselling	-	-		-	-	✓	-	-	-	-	-	✓



(a) Weekly CFR (b) RTPCR tests (DL) (c) Bed occupancy (d) Hospitalization %-age

Call to Action

Currently, we have indexed 12 major Indian states, covering 653 million people, approximately 47% of the population, of one of the largest (as well as worst hit by the pandemic) countries.

The data is publicly accessible and completely free to use. We invite you to contribute to the open-sourced data extraction pipeline or use the data as is in your own research, modeling, and analysis.