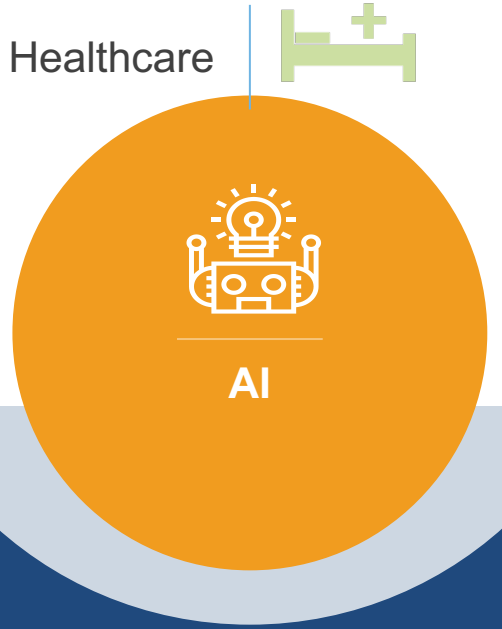
COVID-19 Data Index – Making datasets findable and accessible

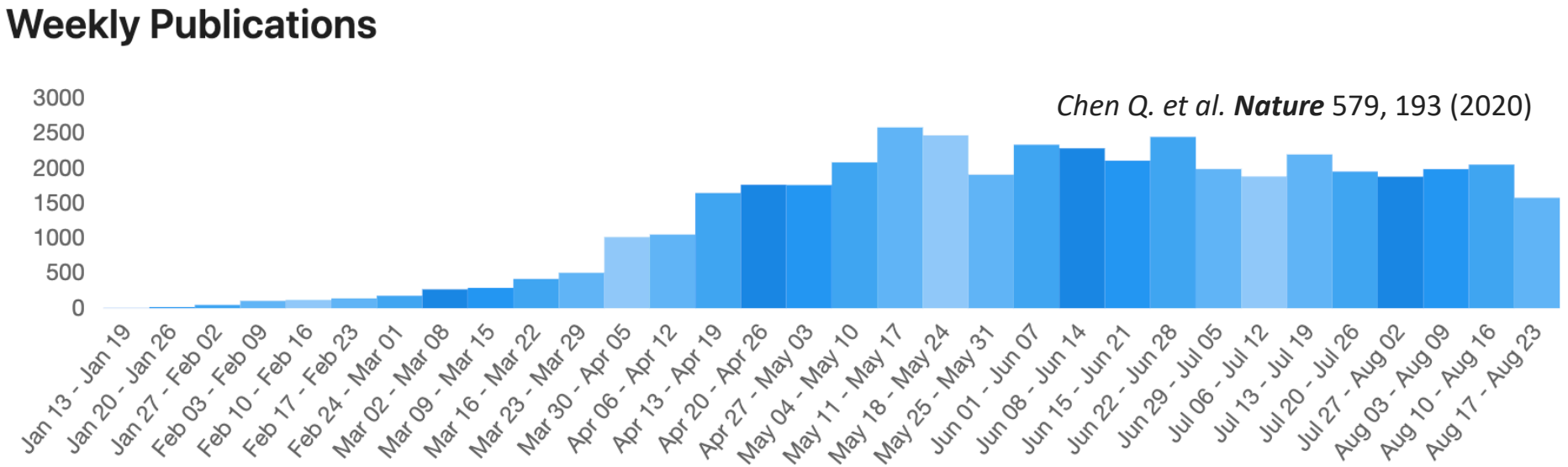


Hua Xu PhD
School of Biomedical Informatics, University of Texas, Houston

August 24th, 2020
2020 KDD Workshop on Applied Data Science for Healthcare

Big Data Generated from COVID-19 Research

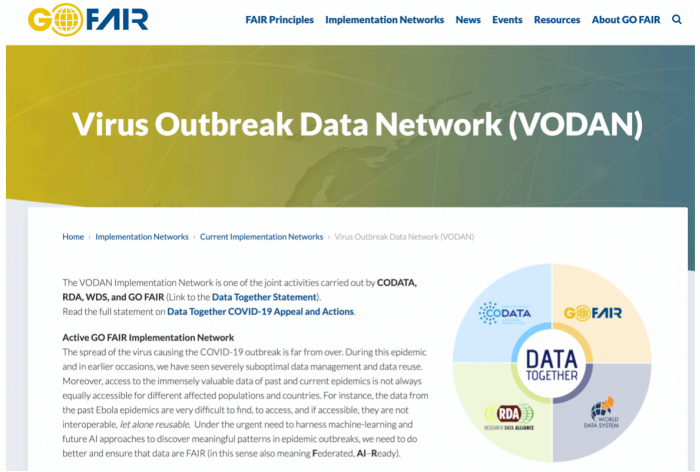
- Large volume of publications: 43,076 (LitCovid, 8/23)

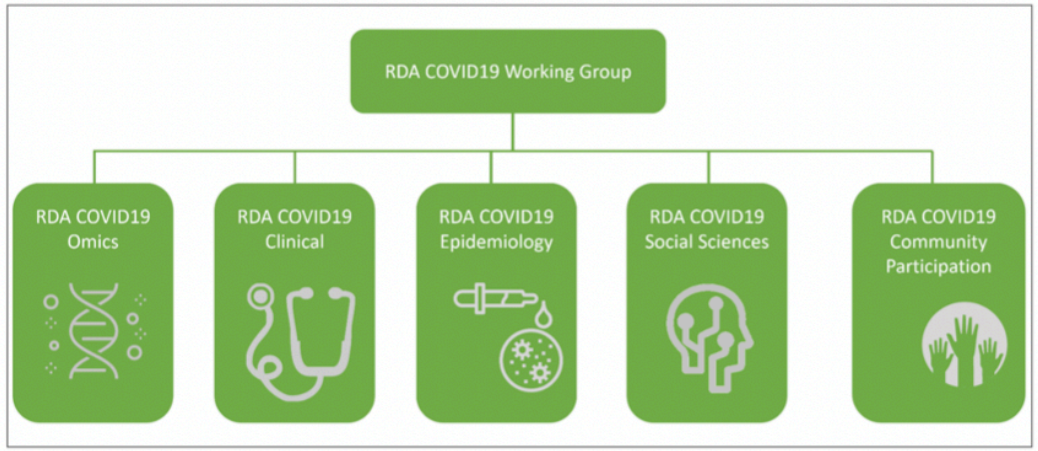


- Heterogenous types of datasets
  - Epidemiological data on testing and case statistics at various locations
  - Omics data from labs
  - Clinical data from surveys, studies (e.g., imaging, assays) or from electronic health records
  - Administrative (e.g., PPE, ventilators, hospitalizations, ICU beds)
  - Socio-demographic, environmental, economic, individual mobility and transportation data.

Effort on Sharing COVID-19 Data

- FAIR Principles (Findable, Accessible, Interoperable, and Reusable)



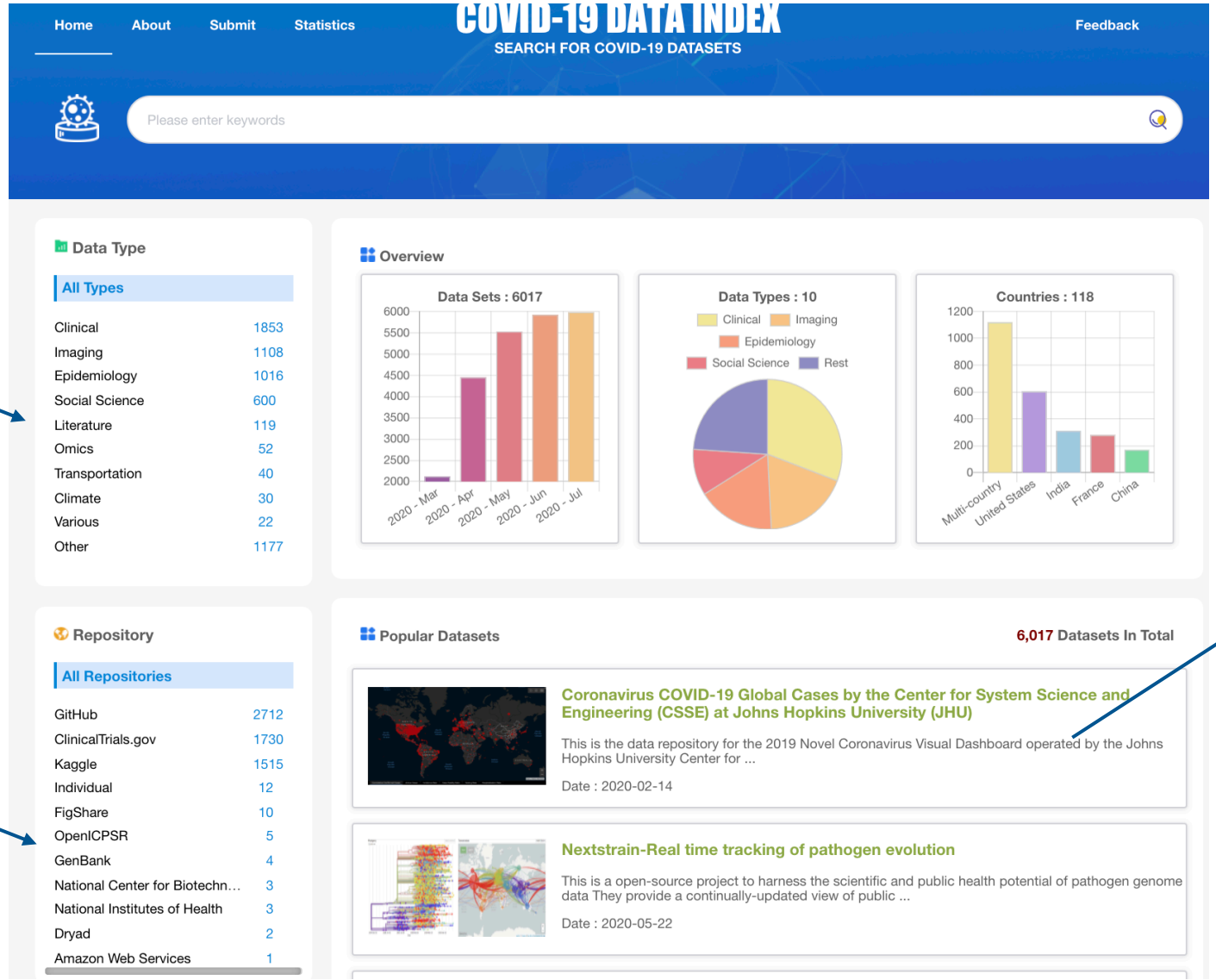


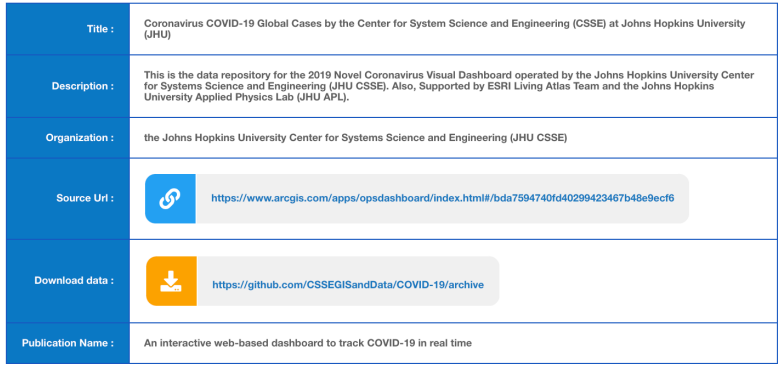
- Other data sharing initiatives

Our Approach – COVID-19 Data Index (www.covid19dataindex.org)

Diverse Data Types

Different Repositories

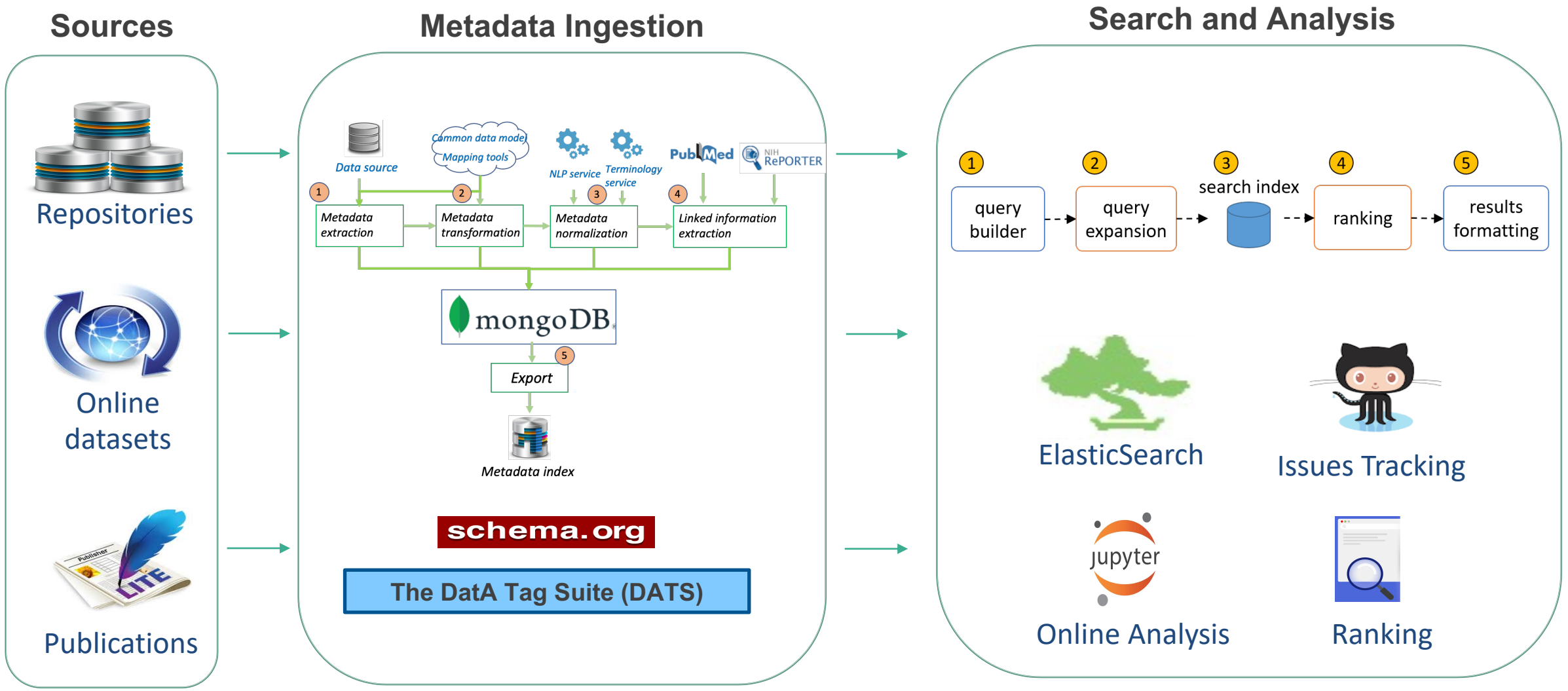




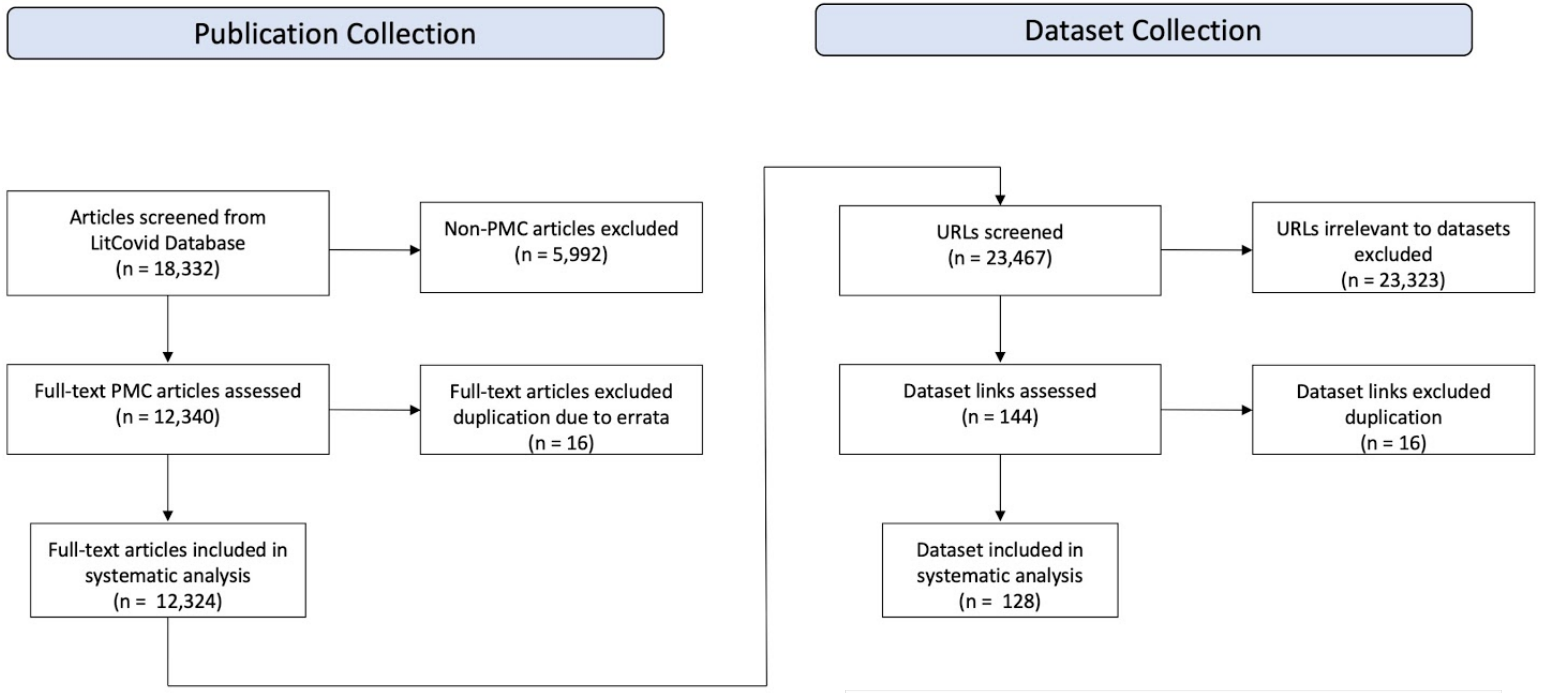
Links to detailed information and direct downloading

Ohno-Machado and Xu. ***Nature*** 584, 192 (2020)

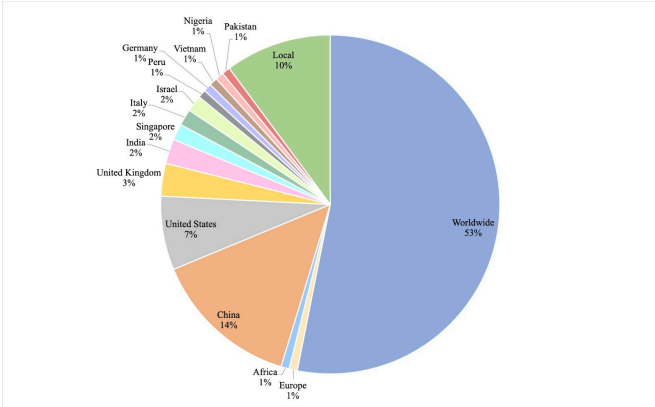
COVID-19 Data Index Workflow



A review of COVID-19 datasets in PubMed Central articles



- Research questions
  - Accessibility
  - Content
  - Citation



Key Findings:

- Only 2% of PMC articles provided Data Availability Statement.
- 28.5% of PMC articles mentioned at least one data link
- 84.4% datasets are immediately downloadable
- Epidemiological datasets accounts for 53.9%
- GitHub is the most popular data repository

Most Cited COVID-19 Datasets in PMC

Dataset	Overall Citations	URL Citations	Article Citations
John Hopkins University Dashboard[2]	454	416	275
Real-time estimation of the novel coronavirus incubation time[3]	239	0	239
Worldometers[34]	231	231	0
Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2)[37]	189	0	189
Estimates of the severity of coronavirus disease 2019: a model-based analysis[38]	132	0	132
Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts[5]	104	0	104
Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus[39]	102	1	102
Early dynamics of transmission and control of COVID-19: a mathematical modelling study[3]	97	0	97
The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak[6]	90	0	90
CDC[35]	87	87	0

Summary

- Data sharing is critical for reproducible research including COVID-19
- FAIR Principles are important for sharing digital assets; but significant efforts are needed to develop FAIR-compliant datasets
- Citation analysis could be the incentive for data sharing; but more work is need for formally citing datasets and measuring its impact

Thank you!

Questions?

hua.xu@uth.tmc.edu