Government | University | Industry
RESEARCH ROUNDTABLE

Resilience of the Research Enterprise During the Covid-19 Crisis

Virtual Meeting Series of the Government-University-Industry Research Roundtable

New Models of Open Research, Data, and Collaboration

May 28, 2020

Abstracts from presenters:

Oren Etzioni, CEO of the Allen Institute for AI

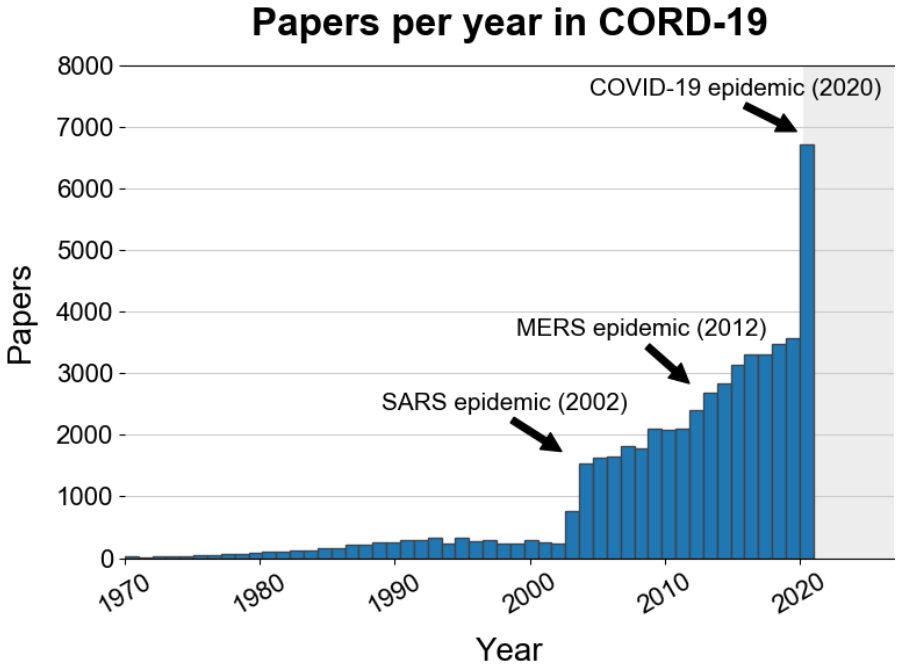
Barend Mons, Leiden University Medical Center

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

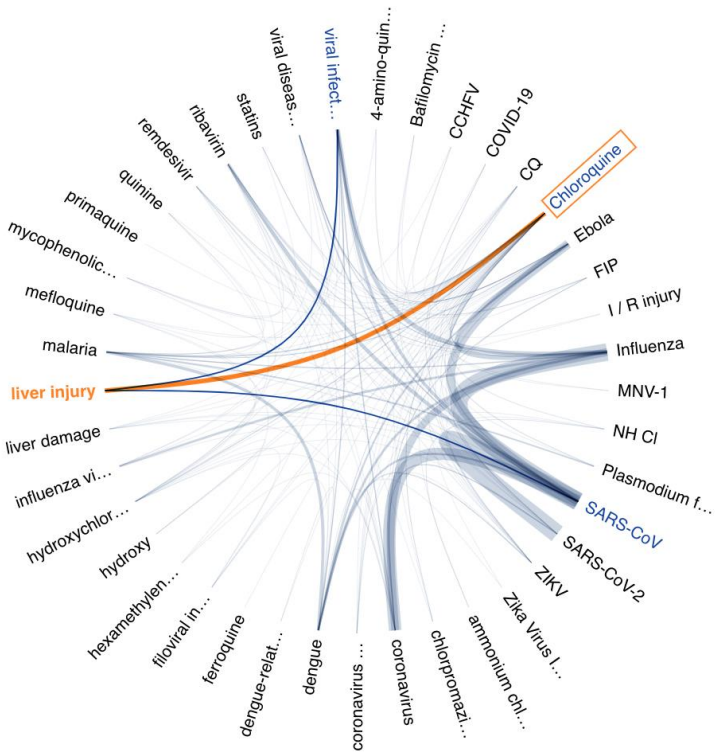
Semantic Scholar and the CORD-19 Dataset
Oren Etzioni
May 2020

The mission of the Semantic Scholar project at the Allen Institute for AI (AI2) is to accelerate scientific breakthroughs by helping scholars locate and understand the right research, make important connections, and overcome information overload. Building on this work to support Covid-19 research is a natural progression. Which is why we jumped into action when the Chief Technology Officer of the United States called in early March with an audacious challenge: put together a machine-readable corpus of all available scientific articles directly relevant to Covid-19 so the AI and IR communities can develop and utilize cutting-edge techniques to obtain insights on the disease, potential treatments, and paths towards a vaccine.

Within a week, in partnership with an impressive roster of partners, we released the Covid-19 Open Research Dataset (CORD-19) resource which has now been viewed close to 2,000,000 times. Starting from 28K papers in the initial release, CORD-19 has grown to contain more than 59K scientific papers about the current Covid-19 pandemic and other coronaviruses, from new publications to historical data stretching back 50 years.



The goal of CORD-19 is to spur the creation of automated systems for interpreting scientific literature and to leverage these systems to improve discovery for biomedical researchers, clinicians, and policymakers. So far, so good — CORD-19 has been downloaded over 75,000 times. The ongoing Kaggle challenge to extract useful findings from CORD-19 has over 550 participating teams, and the TREC shared task for retrieving relevant CORD-19 papers had over 50 group submissions in its first week. Dozens of research groups in both academia and industry have released systems, each showcasing a unique combination of document retrieval, information extraction, and question answering methods. Many of these projects have spurred successful collaborations with biomedical and clinical researchers and other domain experts.



The network of diseases and chemicals associated with Chloroquine, an example of the kinds of insights that can be extracted from CORD-19 — this visualization was produced with the CoViz tool from AI2.

This moment in time will serve as a testament to the potential of machine learning and NLP in science. Recent successes with CORD-19 are a demonstration of how text mining and NLP can be used to advance the pace of scientific discovery. Broad access to scientific literature for automated analysis and discovery could accelerate advancements in all aspects of research, in times beyond the crisis situation we find ourselves in today.

We can also see the limitations of the status quo. The primary distribution format of scientific papers, PDF, is not amenable to text processing. Paper content (text, images, bibliography) and metadata extracted from PDF are imperfect and require significant cleaning before they can be used for analysis. There is also no standard format for representing paper metadata. Existing schemas like the NLM's JATS XML or library science standards like BIBFRAME or Dublin Core have been adopted as representations for paper metadata. However, there is neither an appropriate, well-defined schema for representing paper metadata, nor consensus usage of any particular schema by different publishers and archives. Finally, there is a clear need for more scientific content to be made easily accessible to researchers. Though many publishers have generously made Covid-19 papers available for text mining, there are still bottlenecks to information access. For example, papers describing research in related areas (e.g., on other infectious diseases or relevant biological pathways) are not necessarily open access and therefore not available to the community.

While researchers are churning on Covid-19 research, the Semantic Scholar engine continues to enable advancements in all of science. The rate of scientific publication is increasing every

year, with more than 3 million papers published across 42,500 journals in 2018 alone. This unprecedented flow of information makes staying up-to-date with the scientific literature an increasingly pressing challenge for scholars. Semantic Scholar was created in 2015 in response to this information overload. It is the leading AI-powered search engine for scientific literature which has houses over 180M papers covering all scientific disciplines and over 50M annual users. Semantic Scholar search has helped researchers uncover a range of groundbreaking findings, including gender trends in computer science authorship and supplement-drug Interactions.

A global, equitable and rapid-response data infrastructure for the COVID-19 and future epidemics

Barend Mons
May 2020

The Virus Outbreak Data Network (VODAN) Implementation Network (IN) was conceived to kick-start a 'community of communities' that could design and rapidly build a truly international and interoperable, distributed data network infrastructure that supports evidence-based responses to the viral outbreak. The implementation network has a longer-term goal to reuse the resulting data and service infrastructure, also for future outbreaks. As a GO FAIR IN, VODAN will restrict itself to projects that are directly associated with FAIR data and services relevant to COVID-19. This also means that VODAN will not concern itself with projects purely aimed at studying, for instance, transmission and severity genetics, experimental drug development, vaccine development or actual clinical interventions currently executed to control the epidemic. Only insofar as such activities are in need of access to, and processing of, FAIR data, VODAN partners can add FAIR data and services value.

Much COVID-19-related 'Real World Observation' (RWO) data have political or institutional sensitivities, some have personal components. Sensitive and personal data cannot be 'open' and the sensitive and personal data associated with COVID-19 cannot leave the country, and in most cases, the institution that controls them. Such data can only be accessed partially and in controlled circumstances, and for this to be achieved, the data must be FAIR. We need, therefore, to make a rigorous choice in favor of FAIR (as opposed to always Open) while continuing to emphasize the policy: As open as possible, as closed as necessary.

In such circumstances (and many others) a centralized, data warehousing approach is not fit for purpose, if possible at all. As the data are de facto distributed, rich FAIR metadata is necessary to enable controlled, computational access for analysis or visualization. Traditional article publishing is put to its limits in this crisis and the anachronistic features have become painfully obvious.