Developing a Data Mining Repository to Accelerate Covid-19 Research

LTS is founded by Amy Neustein, PhD, Series Editor of **Speech Technology and Text Mining in Medicine and Health Care** *(de Gruyter); Editor of* **Advances in Ubiquitous Computing: Cyber-Physical Systems, Smart Cities, and Ecological Monitoring** *(Elsevier, 2020); and co-author (with Nathaniel Christen) of* **Cross-Disciplinary Data Integration Models for the Emerging Covid-19 Data Ecosystem** *(Elsevier, forthcoming).*

This paper will describe the Cross-Disciplinary Repository for Covid-19 Research (hereafter called **CR2**). **CR2** is a collection of open-access research data sets related to SARS-COV-2 and Covid-19, which will be developed as a supplement to a forthcoming academic volume examining Covid-19 research from the perspective of text and data mining. We believe that **CR2** can accelerate Covid-19 research by (1) pooling a diverse collection of data sets into a single resource which scientists can utilize; (2) serving as the prototype for larger research portals that can aggregate new Covid-19 data that will emerge from hospitals, labs, and academic institutions in the future; and (3) accelerating the implementation of novel data-integration and software-development technologies which can contribute to Covid-19 in particular, as well as biomedical and overall scientific computing methodology in general. The software used to curate **CR2** data has diverse applictions for software and database engineering, and provides solutions to technical problems with broad reach in the private sector. Further documentation of the **CR2** technology and products may be found on the development repository for the aggregation of **CR2** data (Mosaic-DigammaDB/CRCR).