Advances in Data Modeling and Text Mining for Covid-19 Research

Amy Neustein
Linguistic Technology Systems, Fort Lee, NJ.
E-mail: amy.neustein@verizon.net
Nathaniel Christen
Linguistic Technology Systems, Fort Lee, NJ.

1. Introduction

The Covid-19 pandemic has inspired an unprecedented convergence of scientific research, driven in part by publishers choosing to allow open public access to many research papers and published data relevant to Covid-19 (the disease) and SARS-CoV-2 (the viral agent). The sheer volume of the data presents both a practical challenge — to help scientists find the information most relevant to them — and a valuable case-study in the challenges and possibilities for curating multi-disiplinary information spaces unified around a common scientific theme and research project (in this case, determining how best to treat Covid-19 and to mitigate the global SARS-CoV-2 pandemic).

In a matter of mere weeks, the scientific and publication community has essentially engineered the origination of an entirely new data ecosystem, unified around the challenges of studying and predicting properties of Covid-19 and SARS-CoV-2 at the molecular, genomic, epidemiological, clinical, and public-health levels. In many respects, this ecosystem has emerged haphazardly, with the rush to publicly share text and data taking priority over rigorous curation and accuracy. In this context, scientists and policymakers may benefit from a volume which provides a critical and analytic overview of the Covid-19 data ecosystem: the different genres of data which are marshalled toward scientific investigation of Covid-19 and SARS-CoV-2; how this data is obtained, consumed, analyzed, and interpreted; how research data supports scientific claims pertaining to Covid-19's biological and epidemiological mechanisms and trajectory; and how these scientific claims should translate into public policy, taking into consideration both the importance of basing government actions on empirical data and the gaps in scientific knowledge which make such data inexact and provisional.

Against this background, our proposed volume will be directed at two distinct audiences. At one level, we intend to examine the operational logistics of Covid-19 data: its structures, protocols, analytic methodology, and empirical significance. That is, we intend to identify the distinct scientific disciplines which each concern one facet of Covid-19 research — molecular biology, genomics, radiology, epidemiology, clinical informatics, and their variatious subfields — and, for each of these disciplines, review their distinct paradigms for data acquisition, analysis, and modeling. The point of these expositions is to bring the reader from conceiving ``data'' as something abstract and amorphous,

Table 1. An example table.

Item	Quantity
Widgets	42
Gadgets	13

to understand data as the building-blocks of scientific research and scientific claims. We want readers to have some awareness, when hearing or assessing claims made by either domain experts or public officials, the provenance and history of the data behind these claims. One way to supply this backstory is to examine Covid-19 data from the viewpoint of software engineering — to demonstrate the methodology for data acquisition and management from the perspective of programmers implementing software which manipulates Covid-19 data. This explication would therefore examine the data structures, file formats, **API** protocols, and other technical details intrinsic to writing code which works with Covid-19 data as a digital artifact. Such an exposition might be of interest to programmers who actually are writing Covid-19-related code, but the primary goal of these discussions will be to help scientists (who may be well-versed in data structures relevant to their specialization but less so vis-'a-vis other disciplines), policymakers, and the general public understand the technical chains which transform Covid-19 information from the realm of laboratories and experiments to the realm of public health and policy.

Apart from that topical focus — the analysis of Covid-19 data as a concrete ecosystem manifest in standardized formats, global identifiers, and other concrete information artifacts — we also intend to write for a more theoretical audience for whom the pandemic is a unique case-study in data curation and integration. From this perspective, the Covid-19 data ecosystem is a concrete example through which theories of data integration can be presented and put to the test.

There are two distinct phenomena which render inter-disciplinary data integration significant for Covid-19.

Although Open-Access research data and text provides useful resources for scientists and government officials confronting the pandemic, this material is often shared and curated in a relatively unstructured and imprecise manner.

2. Some LaTeX Examples

2.1. How to Leave Comments

ere's a comment
the margin!

Comments can be added to the margins of the document using the todo command, as shown in the example on the right. You can also add inline comments:

This is an inline comment.

2.2. How to Include Figures

2.3. How to Make Tables

Use the table and tabular commands for basic tables — see Table 1, for example.

2.4. How to Write Mathematics

Architecting Data Models for Scientific Disciplines Associated with Covid-19

Data structures for molecular biology and virology How genomic data is stored and analyzed in the coronavirus context Structuring of radiographic data in the context of Covid-19 Reviewing epidemiological structures and methodology for SARS-Cov-2 research Modeling clinical data in the Covid-19 patient population

Creating a cross-disciplinary eco f c-19

Chapter: Approaches for merging heterogeneous data sets: Ontologies and Hypergraphs Chapter: Scientific Workflows and Inter-Application Networking: reviewing data pipelines commonly used in Covid-19 research: Chapter: Formal Procedural Models (representing computational procedures applicable to Covid-19) Chapter: Integrating Procedural and Data Models Chapter: Type Theorys for Procedural Data Modeling

Part Three: Software Dev meth for Covid-19

Chapter: Applying Data Mining techniques to Covid-19 Research Corpora Chapter: Text Mining of Covid-19 Publication Archives Chapter: HCI Approaches for Covid-19 software Chapter: Software dev and testing methods for Covid-19 software