EXECUTIVE SUMMARY

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 scientific progress vis-à-vis Covid-19 in particular, as well as to biomedical and overall scientific computing methodology in general. The software used to curate **CR2** data has diverse applications 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).

The sudden emergence of Covid-19 as a global crisis has cast a spotlight on computational and technological challenges which, in the absence of a catastrophic pandemic, would rarely rise to public attention. In particular, an effective response to the dangers of SARS-COV-2 requires coordinated policymaking integrating different levels of government as well as diverse modes of scientific inquiry. Genomic, biomolecular, epidemiological, sociodemographic, clinical, and radiological information are all pertinent to Covid-19. In this environment, it is important that the empirical foundations for expert recommendations — which in turn drive public policies of enormous social and economic consequence — be transparently documented and critically examined. The proper synergy between government and science depends on data centralization: given the gaps in our current Covid-19 knowledge, it is appropriate that different jurisdictions craft responses to the pandemic in different ways. There is no central authority with sufficient epistemic force to legitimize homogeneous mandates across the entire country. But it is also important that policy differences reflect alternative interpretations of our scientific knowledge, or the diverse needs of local communities — rather than being a haphazard consequence of governments merely working with divergent, competing, and poorly integrated data.

The current administration, along with numerous corporate and academic entities, has clearly recognized the need for a more centralized paradigm for sharing Covid-19 data. For example, the White House spearheaded a scientific initiative to develop **CORD-19**, an open-access corpus of over 33,000 peer-reviewed publications related to Covid-19, which were transformed into a common machine-readable representation so as to promote text and data mining. Similarly, large institutions such as Google, Johns Hopkins, and Springer Nature have all implemented some form of coronavirus data-sharing platform targeted to both scientists and policymakers. However, these two aspects of the corporate/academic contributions to Covid-19 data sharing have been incomplete, for opposite but complementary reasons. **CORD-19** is highly structured and tightly integrated, but it focuses on text mining and scientific documents, not research data. While it is possible to find data sets about Covid-19 through **CORD-19**, the techniques to do so are cumbersome and non-scalable. Conversely, projects such as the Johns Hopkins coronavirus "dashboard" provide accessible data sets, but these projects are isolated and do not offer the level of structure and integration evinced by **CORD-19**. In short, an optimal Covid-19 research platform would merge the structural text-mining rigor of **CORD-19** with the data-centric focus of several isolated projects sharing Covid-19 data with the general public.

These are the principles which have guided the design of **CR2**. In particular, **CR2** can provide value at different scales of realization. Relatively small data sets serve several scientific and computational purposes: (1) they can provide researchers with a mental picture of how data in different disciplines, projects, and experiments is structured; (2) they can serve as a prototype and testing kernel for technologies implemented to manipulate data in relevant formats and encodings; and (3) they can lay the foundation for data-integration strategies. For example, when designing a representation format and/or implementing code to merge different data formats into a single structure (or meta-structure), it is useful to work with small, representative examples of the data structures involved, so as not to complicate the integration logic with computational details solely oriented to scaling up the data-management logistics. As a result, **CR2** can provide a testbed for implementing data-integration technologies which can scale up as needed. To fulfill this mission, **CR2** can aggregate relatively small data sets which have until now been published on academic and research portals, such as Springer Nature, Dryad, and DataVerse.

At the same time, a more substantial (and not necessarily fully open-access) Covid-19 information space would also be beneficial to the scientific and policymaking community. Ideally, then, **CR2** will be paired with a larger technology sharing a similar implementational strategy but with different accession paradigms, allowing for an open-ended collection of Covid-19 data which users may selectively access, instead of a single package that users may acquire as an integrated resource. The common denominator in both cases is the importance of deploying novel and contemporary data-integration techniques to centralize Covid-19 research as much as possible. Accordingly, this paper will briefly summarize how **CR2** can accelerate Covid-19 data integration on a practical and technological level.

I Methodology for Covid-19 Data Integration

As indicated above, pertinent Covid-19 data derives from multiple scientific disciplines. On a technological level, Covid-19 data is documented via a wide array of file types and data formats. This diversity presents technological challenges: if a Covid-19 information space encompasses files representing 25 different incompatible formats, users need 25 different technologies to fully benefit from this data. In many cases, however, data incompatibilities are merely superficial — an important subset of Covid-19 data, for example, has a common tabular meta-model, even if the data is realized in clashing technologies (spreadsheets, relational databases, comma-separated-value or Numeric Python files, and so forth). One level of data integration can therefore be achieved simply by encoding tabular structure into a common representation: any field in a table can be accessed via a record number and a column name and/or index. In some cases, more rigorous integration is also possible, for example by identifying situations where columns in one table correspond semantically or conceptually to those in a second table. In either case, it is reasonable to assume that a single abstract data format lies behind surface data-expression in patterns such as spreadsheets and **CSV**, so that all files in an archive encoding spreadsheet-like data can be migrated to a common model.

Other forms of clinical and epidemiological inputs are more amenable to graph-like representations. For instance, trajectories of viral transmission through person-to-person contact is obviously an instance of social network analysis. Similarly, models of clinical treatments and outcomes can take graph-like form insofar as there are causal or institutional relations between discrete medical events: a certain clinical observation *causes* a care team to request a laboratory analysis, which *yields* results that *factor* into the team's decision to *administer* some treatment (say, a drug *from* some provider *with* some chemical structure), which observationally *results* in the patient improving and eventually *being* discharged. In short, patient-care information often takes the form — at least conceptually — of a network comprised of different "events", each event involving some observation, action, intervention, or decision made by care providers, and where the important data lies in how the events are interconnected: both their logical relationships (e.g., cause/effect) and their temporal dynamics (how long before a drug leads to a patient's improvement; how much time elapses between admission to a hospital and discharge). These graph-like representations are a natural formalization of "patient-centered" data models.

A higher level of data integration can then be achieved by merging tabular and graph-like models into a single *hypergraph* format. A significant subset of Covid-19 data (or, more generally, any clinical/biomedical information) conforms to either tabular or graph structures, and so it is feasible to unify all of this information into a common framework to the degree that one works with a meta-model which incorporates both record-sets and graph structures (node-sets and edge-sets) in its representational arsenal. A graph-plus-table architecture is generally considered some form of Hypergraph model, and indeed **CR2** uses a hypergraph paradigm to merge many different sorts of information into a common structure. In particular, **CR2** introduces a new "Hypergraph Exchange Format" (**HGXF**) which can provide a text encoding of many files that, when originally published, embodied a diverse array of file-types requiring a corresponding array of different technologies. **CR2** will include computer code to read **HGXF** files and use them to create hypergraph-database instances. In short, **CR2** will promote Covid-19 data integration by translating a wide range of files into a common **HGXF** format.

Not every format relevant to Covid-19 can be realistically translated to **HGXF**. In particular, scienctific fields requiring substantial quantitative analysis — e.g., biomechanics or genomics — express data via encodings optimized for relevant mathematical operations. In this scenario, **CR2** will not attempt to migrate *all* of a data file to **HGXF**. However, even for these files **CR2** will generally provide a supplemental **HGXF** encoding supplying data *about* the original file, with information about the file type, preferred software components for viewing/manipulating its data, and so forth. In this manner the contents of non-**HGXF** files can be indirectly included into the **CR2** hypergraph-based ecosystem.

1.1 Hypergraph Data Models and Multi-Application Networks

As has been outlined thus far, vast quantities of Covid-19 data can be wholly or partially integrated into a single hypergraph framework, which accordingly simplifies the process of designing software applications and algorithms to analyze and manipulate this data. Specifically, software components can employ a single code library to obtain, read, consume, and store data, rather than needing to re-implement this logic for a large number of different file formats and/or database models.

Quality software — especially in the clinical and biomedical context — demands a careful balance between applications which are either too narrow or too broad in scope. On the one hand, doctors often complain that homogeneous Electronic Health Record systems (where every digital record or observation is managed by a single all-encompassing application) are unwieldy and hard to work with. This is understandable, because the clinical tasks of health workers with different specializations can be very different. Conversely, doctors also complain about software and information systems which are so balkanized that they must repeatedly switch between different, non-interoperable applications. In short, clinical, diagnostic, and research software should be neither too homogeneous nor too isolated; finding the proper balance between these extremes is, one can say, a major challenge to the usability of electronic health systems going forward.

Against this background **CR2** demonstrates novel solutions to this problem: it focuses on the dimensions of data acquisition and management that are specific to individual scientific or medical specializations, while also identifying requirements that are consistent across domains. Scientific software generally needs to hone in on the data visualization and analytic requirements of particular disciplines; biochemists use different programs than astrophysicists. However, much of the code underlying scientific applications has nothing to do with these high-level models or theories, but is simply a fulfillment of basic data-management functionality — data storage, accession,

provenance, searching, user validation, and so forth. In effect, the computational requirements of scientific and biomedical software can be partitioned into two classes: one the one hand, domain-specific logic which reflects the quantitative or theoretical models of narrow scientific fields; on the other hand, data-management logistics which can be realized within a central access hub, rather than being re-implemented by each application in isolation.

In short, this architecture conceives of a central hub which is responsible for storing data and for serving as a common access point — providing the gateway where authorized users can gain access to heterogeneous information spaces utilized by an array of domain-specific software applications. These peer applications would not be directly responsible for data persistence or user identity management, so they can focus on their specific data analysis and visualization capabilities. The central hub, serving multiple peer applications, is then a heterogeneous data space managing multiple applications' information while also tracking information about the applications themselves: helping users to identify and launch the software which is most directly relevant to their clinical or research needs at the moment. Meanwhile, because peer applications are jointly connected to a central hub, it is possible to implement scientific workflows where one application may send and receive data from its peers, allowing applications to complement each others' capabilities.

The multi-application networking architecture just outlined has precedents in current database and engineering technologies. Many hospitals and medical institutions employ some version of a "Data Lake", pooling disparate data sources into a heterogeneous aggregate which is then accessed by multiple client applications. Similarly, machine-learning and Artificial Intelligence often adopts "software agents", or analytic modules in contexts such as Online Analytic Processing, which again represent semi-autonomous software components sharing an originary data hub. Web applications, too, often act as domain-specific subsidiaries deferring operational requirements, such as user authentication or transaction processing, to a central web service. The limitation of multi-application networks in these existing contexts are that the software agents involved are generally "lightweight", with relatively primitive user-interface design.

By contrast, the technology introduced with **CR2** will develop multi-application networking in the context of more substantial, desktop-style scientific applications. In short, this technology offers a hybrid of the development methodologies employed for desktop scientific software and those applicable to multi-agent heterogeneous data stores, like a Semantic Data Lake. In particular, **CR2** will utilize a new hypergraph database engine, coded in the **C++** programming language, which has a unique focus on supporting native **GUI** applications from the ground up, including persisting application state and storing application documentation within the database itself.

II Augmenting Covid-19 Data with Patient Narratives

In addition to aggregating published data sets, **CR2** may be used as a repository for collecting new Covid-19 information. With that in mind, we are prioritizing the design of a standard for storing and accessing natural-language text representing patients' subjective descriptions, which may be useful both for diagnostic assessments and for gleaning information about Covid's psychological and neurological effects.

Here again **CORD-19** offers a useful case study. The **CORD-19** corpus has archived thousands of research papers in a common format suitable for text mining. This textual data was then made openly available to researchers and programmers in the hopes that developers would implement new text-mining algorithms to mine data or detect patterns in the overall corpus. Just as **CR2** envisions an analogous curation of research data sets for data mining, we can also see the benefit of a text-mining repository of patient narratives submitted by citizens who have been diagnosed with, or believe they may have, Covid-19. **CR2** will not specify how these narratives should be collected, but it will implement a common representational format so that patient narratives can be pooled, analogous to the merging of research manuscripts within **CORD-19**.

In modeling patient narratives, this technology will be oriented toward the scientific-computing ecosystem outlined in the previous section. In particular, we assume that **GUI**-based desktop applications will be the primary instruments for data collection and analysis; this means that patient narrative representations may be paired with **GUI** or multi-media content. For example, the software with which patients submit information could allow them to pair text-form narratives with graphics indicating the source of pain or discomfort; or narratives may be accompanied by audio files where patients cough or speak into a microphone. A patient-narrative representation therefore must be flexible enough to include these possible forms of multi-media content.

As described earlier, an information space adapted for multiple peer applications should encompass capabilities for saving application state, which includes features for modeling instances of **GUI** classes. This technology provides the necessary infrastructure for managing patient narratives; for example, consider a multi-media intake form where patients may describe symptoms by placing icons (representing pain or discomfort) against anatomic silhouettes (back, front, head, and so forth). As patients use such a form, **GUI** state corresponds to the subjective symptomology which should be incorporated into patient narratives and medical records. This is an example of how application-persistence logic can be marshaled to the related project of curating patient narratives.

Once aggregated into research corpora, patient narratives are available to be analyzed by different language-processing and text-mining methodologies. Our strategies for encoding patient texts are based on Sequence Package Analysis, a novel discourse-representation technique developed by Amy Neustein (see Using Sequence Package Analysis to Improve Natural Language Understanding). The patient-narrative models are also designed to facilitate annotation according to techniques described in Dr. Neustein's Application of Text Mining to Biomedical Knowledge Extraction: Analyzing Clinical Narratives and Medical Literature.

Further documentation of text-encoding methodology applicable both to patient narratives and to manuscripts associated with **CR2** research data is available on the **CR2** web site, such as here (this is a downloadable **PDF** link; visit the repository to see the larger archive structure).