???

...

As a supplement to the main book, several data sets will be published (or republished) relevant to subjests covered by chapters in the text, such as signal-processing, bioacoustics, and Natural Language Processing. These datasets' primary purpose is to demonstrate data-management and software-development techniques discussed in the volume, particularly in the third chapter (the data sets are being curated by that chapter's author). These data sets introduce a new protocol for publishing research data, which essentially extends existing initiatives such as Research Objects and **FAIR** (which stands for "Findable, Accessible, Interoperable, Reusable". In principle, these initiatives — which have been pushed by publishers, academic instutitions, and certain government agencies (such as the NIH) — are intended to make it easier for scientists to find, assess, reuse, and re-produce research data. In practice, however — due partly to technological limitations and partly to authors and publishers being slow to adopt the new protocols — the ecosystem of published data sets remains far less rigorously organized than the ecosystem of published scientific/academic books and articles.

The sudden emergence of Covid-19 as a medical and governmental priority presents a unique case-study of the limitations of our existing research-data platforms. Publishers have, to some degree, recognized the extra-ordinary nature of the new coronavirus crisis and taken some commensurate mearures; for example, Springer Nature has committed to releasing as open-access documents a collection of papers potentially helpful for doctors and policymakers responding to the pandemic — some newly published and some dating back several years (the earlier research involving viruses biologically similar to Covid-19). As of mid-March, this portal encompassed 43 articles, of which 15 were accompanied by research data that could be separately downloaded (this number does not include papers that document research findings only indirectly, via tables or graphics printed inline with the text). Collectively these articles referenced over 30 distinct data sets (some papers were linked to multiple data sets), forming a data collection which could be a valuable resource for Covid-19 research — not only through the raw data made available but as a kernel around which new coronavirus data could accumulate. However, there is currently no mechanism to make this overall collection available as a single resource.

This problem demonstrates, among other things, how the Research Object protocol is limited in applying only to *single* articles. There is no commensurate protocol for publishing *groups* of articles which are tied to groups of data sets unified into an integral whole. Open-access Covid-19 papers also reveal limitations of exiting online document portals, particularly with respect to how publications are linked to data sets. There is no clear indication that a given paper is associated with downloaded data; usually readers ascertain this information only by reading or scrolling down to a "supplemental materials" or "data availability" addendum near the end of the article. Moreover, because the Springer Nature portal aggregates papers from multiple sources, there is no consistent pattern for locating data sets; each journal or publisher has their own mechanism for alerting readers to the existence of open-access data and allowing them to download the relevant data sets.

Aside from these user-interaction issues, the collective group of Covid-19 data sets illustrate the limitations of information spaces pieced together from disconnected raw data files with little additional curatation. The files inluded in this group of data sets encompass an array of file types and formats, including **FASTA** (which stands for Fast-All, a genomics format), **SRA** (Sequence Read Archive, for **DNA** sequencing), **PDB** (Protein Data Bank, representing the **3D** geometry of protein molecules), **MAP** (Electron Microscopy Map), **EPS** (Embedded Postscript), and **CSV** (comma-separated values). There are also tables represented in Microsoft Word or Excel formats. Although these various formats are reasonable for storing raw data, not all of them are actually machine-readable; in particular, the **EPS**, Word, and Excel files need manual processing in order to use the information they provide in a computational manner. A properly curated data collection would instead, as much as possible, unify disparate sources into a common machine-readable representation (such as **XML**).

Going further, productive data curation would also aspire to *semantic* integration, unifying disparate sources into a common data model. For example, multiple spreadsheets among the Springer Nature Covid-19 data sets hold sociodemographic and epidemiological information relevant to modeling the spread of the disease. These different sources could certainly be integrated into a canonical social-epidemiology-based representational paradigm which recognizes the disparate data points which might be relevant for tracking the spread of Covid-19 (with the potential to unify data from many countries and jurisdictions). This is not only a matter ofe data *representation* (viz., how data is physically laid out), but also of data types and computer code. According to the Research Object protocol, data sets should include a code base which provides convenient computational access to the published data. In the case of Covid-19, this entails creating a sociodemographic and epidemiological code library optimized for Covid-19 information, which would be the primary access point for researchers seeking to use the data which has been published in conjunction with the 43 manuscripts examined here that were aggregated on Springer Nature, along with any other coronavirus research which comes online. Similar comments apply not only to tabular data represented in spreadsheet or **CSV** form, but to more complex molecular or microscopy data that needs specialized scientific software to be properly visualized.

Considering the overall space of Covid-19 data, it is unavoidable that some files require special applications and cannot be directly merged with the overall collection. For instance, there is no coherent semantics for unifying Protein Data Bank files with social-epidemiology; files of the former type have specific scientific uses and can only be understood by special-purpose software. Nevertheless, a well-curated data-set collection can make use of such special-purpose data as convenient as possible. In the case of Protein Data Bank, a downloadable Covid-19 archive can include source code for **IQMOL**, a molecular-visualization application that supports **PDB** (among other file formats) and has few external dependencies (so it is relatively easy to build from source).

Indeed, a curated Covid-19 archive might include an enhanced version of **IQMOL** prioritizing Covid-19 research, with the goal of integrating biomolecular and social-epidemiological data as much as possible. For example, as Covid-19 potentially mutates in different ways in different geographic areas, it will be important to model the connections between "hard" scientific Covid-19 information and sociodemographics. As the pandemic evolves, genomic and biochemical information may be linked to particular strains of virus, which in turn are linked to sociodemographic profiles: certain strains will be more prevalent in different populations. Consequently, models of Covid-19 variation will need to be formulated and then integrated with both chemical/molecular data and sociodemographic/epidemiological data. Different Covid-19 strains then form a bridge linking these different sorts of information; researchers should be able to pass back and forth from molecular or genomic visualizations of Covid-19 to social-epidemiological charts and tables based on viral strains. Ideally, capabilities for this sort of interdisciplinary data integration would be provide by a Covid-19 archive as enhancements to applications, such as **IQMOL**, that scientists would use to study the published data.

Logistically speaking, not all Covid-19 data is practical to reuse as a downloadable package. This is especially true for genomics; several of the aforementioned 43 coronavirus papers included data published via online data banks capable of hosting data sets that are too large for an ordinary computer. In these situation scientists formulate queries or

analytic scripts that are sent remotely to the online repositories, so that researchers access the actual published data only indirectly. Nevertheless, access to these data sets can still be curated as part of a Covid-19 package; in particular, computer code can be provided which automates the process of networking with remote genomics archives through the accession numbers and file-formats which those archives recognize. A case-study in this type of technology is provided by Verily, the Alphabet division which (according to news reports, such as a New York Times cover-page story on March 15th) is developing an online platform for the public to check symptoms and locate providers prepared to treat Covid-19 cases. One Verily project is PurpleData, esssentially a miniature version of Google's "BigData" platform; unlike its larger archetype, PurpleData is designed to run and host data sets small enough for a single computer, so as to develop and test analytic queries which are eventually posted to remote BigData services. In short, PurpleData is a simulation of BigData designed for prototyping code which will interface with BigData. This prototyping/simulation paradigm is a useful model to apply to other remote-analytics services, such as the National Center for Biotechnology Information (**NCBI**) genomics archives where some of emerging Covid-19 data has been stored. The point is not to recommend a direct generalization of PurpleData — indeed, PurpleData, implemented in python, is not an ideal case-study in prototyping/simulation techniques. One could argue that frameworks based on WhiteDB — an under-appreciated database engine (an **SQL**/**NoSQL** hybrid) currently used for certain CyberPhysical and telemedical systems — would be more definitive prototyping solutions. WhiteDB is a standalone **C**/**C++** engine which can be dropped as source code into any compiled project; it therefore offers fully transparent access to data-serialization code, and would be especially useful for emulating the structural and persistence methodologies implemented at large scale by cloud-analytic services such as BigData or **NCBI** Nucleotide/GenBank.

This discussion has highligted limitations of data sets published in conjunction with coronavirus articles made available as open-access recources on SpringerNature. The point here is not to criticize the work of individual authors, but rather to argue for a distinct data-curation stage in the publication process, with data curators playing a role distinct from that of both authors and editors. Moreover, the discussion has hopefully highlighted problems with current data-sharing paradigms, even those such as the Research Object and **FAIR** initiatives which are explicitly devoted to improving how open-access data sets are published. The Covid-19 case study documents several lacunae in the Research Object protocol, for example, which point to the need for a more detailed extension of this protocol. In particular, an enhanced protocol should encompass:

1. A canonical framework for archiving collections of data sets, not only single data sets (and not only groups of data sets published with a single research paper). For example, all data sets published alongside the 43 Springer Nature articles could be unified into a single collection.
2. A code base accompanying data-set collections designed to help research unify the infrmation provided. Curating the overall collection would involve pooling disparate data into common representation, and implementing computer code which deserializes and processes the unified data accordingly. For instance, **CSV**, **EPS**, and Microsoft Word/Excel tables could be migrated to **XML**, **JSON**, or a more complex common format (Chapter 3 of *Advances in Ubiquitous Computing* presents the theoretical case for a "software-centric" representational format based on hypergraphs). Customized computer code could then be implemented specifically to parse and merge the information present in single data sets within the overall collection. This implementation would reciprocate the Research Object goal of unifying code and data, but again would operate at the level of an aggregate of research projects rather than a single Research Object.
3. A unified data-set collection should provide prototyping and remote-access tools to interface with web-based information spaces that host data sets too large to be individually downloaded. Ideally, these would include simulations of remote services analogous to PurpleData vis-à-vis BigData, which would help scientists understand the design of the remote archives and how to interface with them.
4. A unified research portal should also influence the design of the web portals where associated texts are published. It should be easy for readers to identify which articles have supplemental data files and to download those files if desired. Moreover, textual links should be established between publication content and data sets — for instance, a plot or diagram illustrating statistical or equational distributions should link to the portion of the data set from which that quantitative data is derived.

The above discussion has considered the Springer Nature articles as a case-study, but analogous comments would apply to other Covid-19 related resources. For example, John Hopkins University has created and deployed a Covid-19 "dashboard" tracking the spread of the virus; new data from which the web dashboard is generated is published via a **GIT** archive roughly once daily. If and when the reported Verily portal comes online, hopefully machine-readable access to that public data will be provided either via an anlogous updated archive or via an **API**. Ideally, these disparate Covid-19 projects will be interoperable: any code published in relation to the Springer Nature coronavirus collection, for example, could include components implemented to access the John Hopkins and (anticipated) Verily data sets as well as all the data brought in via the Springer Nature articles. Insofar as conscious effort is made to integrate all publicly accessible Covid-19 data via an overarching toolkit, it will be easier to continually accumulate new data sources as these come online.

This discussion has also used the Covid-19 crisis as a lense through which to examine data-publishing limitations in general. These problems are not specific to coronavirus, but the almost unprecedented urgency of this epidemic exposes how science and the publishing industry are still struggling to develop technologies and practices which keep pace with the intersecting needs of systematic research and public policy. An optimistic projection is that the crisis will spur momentum toward a more sophisticated data-sharing paradigm — perhaps a generalization of the Research Object protocol toward data-set collections, with features as outlined above. We hope to contribute to the emergence of such a protocol, so as to operationalize some of the ideas laid out in *Advances in Ubiquitous Computing*. It would be especially rewarding if an integrated data-set collection devoted to Covid-19 would serve as a first example and a test-bed for this new paradigm, given the potential public benefit of unifying disparate Covid-19 data as effectively as possible, where this technology can then be generalized to other medical priorities and other academic disciplines overall.