Proposing a CORD-19 Software Development Kit to Improve Machine Readability

CORD-19 (the "Covid-19 Open Research Dataset") is a new coronavirus data collection which was released in conjunction with a White House initiative to spur Covid-19 research. This initiative is described as a "call to action ... to develop new text and data mining techniques that can help the science community answer high-priority scientific questions related to COVID-19" (see https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/). The White House spearheaded a consortium of industry and academic institutions, led by the Allen Institute for AI Research, who curated a "machine-readable Coronavirus literature collection" which includes article metadata and (in most cases) publication text for over 44,000 coronavirus research papers. This corpus is paired with links to publisher portals (including Springer Nature, Wiley, Elsevier, the American Society for Microbiology, and the New England Journal of Medicine) providing full open access to COVID-19-related literature; these resources collectively constitute CORD-19 (see [2]).

Linguistic Technology Systems (LTS) would like to create a Software Development Kit (SDK) to help scientists utilize CORD-19. This SDK would include new code libraries explicitly implemented for data-management operations specific to CORD-19. The SDK would also include a package of applications, modified to support COVID-19 research, that would collectively create an integrated and self-contained computing environment. These two parts of the SDK — the new code libraries and the application package — are outlined in this paper.

I New Code Libraries within the Proposed SDK

The CORD-19 collection was formulated with the explicit goal of promoting both *text mining* and *data mining* solutions to advance coronavirus research. This means that CORD-19 is intended to be used both as a document archive for text mining and as a repository for finding and obtaining coronavirus data for subsequent research. Because the White House announcement requests institutions to develop additional technologies which would help scientists and jurisdictions to take advantage of CORD-19, CORD-19 was released with the explicit anticipation that industry and academia would augment the underlying data by layering on additional software. Our proposed CORD-19 SDK would do just that: this SDK would be a component providing analytic capabilities which make the raw CORD-19 data more valuable, and a toolkit through which other developers could create new solutions targeting CORD-19.

To accomplish these goals, our proposed SDK would include a collection of new code libraries to aid programmers in the implementation of algorithms to investigate the CORD-19 corpus. These code libraries would enhance the underlying data by providing the following features:

Tools for Correcting Transription Errors Transription errors can cause the machine-readable text archive to misrepresent the structure and content of documents. For instance, there are cases in CORD-19 of scientific notation and terminology being improperly encoded. As a concrete example, "**2'-C-ethynyl**" is encoded in CORD-19 as "**2 0 -C-ethynyl**", which could stymie text searches against the CORD-19 corpus (see [3] for the human-readable publication where this error is observed; the corresponding index in the corpus is 9555f44156bc5f2c6ac191dda2fb651501a7bd7b). To help address these errors, our SDK would augment the CORD-19 corpus by providing alternate machine-readable encodings of the corpus documents in formats such as XML whenever these are available, as a supplement to CORD-19's JSON representation. Because these alternate encodings would be based directly on published manuscripts, they would not be subject to transcription errors. The SDK would then provide tools to cross-reference multiple versions of each document, so as to correct errors in the original JSON encodings.

Tools for Converting Between Data Formats Although the CORD-19 corpus is published as JSON files, many text-mining tools such as those I review in [6] take input in alternative formats, such as XML, BIOC, or JSON trees with different schema than CORD-19. Our proposed SDK would provide libraries to read CORD-19's JSON files and output data in one of these alternative formats, so as to initiate a text mining workflow. The SDK would also include tools for manipulating which *results* from text mining algorithms, which is often represented in formats such as XML and CONLL (Conference on Natural Language Learning).

Tools for Enhanced Annotation The existing CORD-19 corpus does not directly provide a mechanism for asserting annotations related to text mining, such as Named Entity Recognition or formally recognized biomedical concepts. However, the archival schemas supports standoff annotation for intra-document references, so our SDK can provide code for additional standoff annotation categories of the kinds commonly used in biomedical text mining. As a concrete example, the corrected text segment "**2'-C-ethynyl**" mentioned earlier can be annotated as a molecular component.

Tools for Research Data-Mining Although CORD-19 provides a structured representation of a large collection of research *papers*, there is no easy-to-use tool for finding research *data* through CORD-19. However, the CORD-19 resources indirectly include numerous research data sets published alongside the scientific articles. The collection of manuscripts available through the Springer Nature portal, for example, which is linked from CORD-19, included over 30 COVID-19 data sets encompassing a wide 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), CSV (comma-separated values), and tables represented in Microsoft Word and Excel formats. Our SDK would marge these resources into a common representation (such as XML) wherever possible.

Wrappers for Network Requests Scientific use of CORD-19 will often require communicating with remote servers. For example, genomics information in the COVID-19 data sets is generally provided in the form of accession numbers which are used to query online genomics services. Similarly, text mining algorithms often rely on dedicated servers to perform Natural Language Processing; these services might take requests in BIOC form and respond with CONLL data; or epidemiological studies of COVID-19 may need to access APIs or data sets such as the John Hopkins University "dashboard" (see https://coronavirus.jhu.edu/map.html). To reduce the amount of "biolerplate code" which developers need for these networking requirements, the SDK would provide code libraries based on the QT Networking Module to manage networking requests and responses. Programmers would therefore have a unified framework with which to construct remote queries and route responses, a framework which could be used across disparate scientific disciplines (genomics, NLP, epidemiology, and so forth).

The code libraries just decribed would augment the value of CORD-19 by providing tools out-of-the-box to help scientists (and their codewriters) leverage CORD-19 data. Although we can expect that numerous code libraries will be implemented so that researchers can use CORD-19, a CORD-19 SDK would be beneficial because it would integrate multiple libraries into a single package, designed to be easily interoperable. In particular, these libraries would be implemented in a manner which prioritizes ease of development: the SDK would comprise a *standalone* and *self-contained* development environment with minimal external dependencies. This priority would extend also to software tools that would be bundled together with the new code libraries. These software tools will be discussed next.

II Software Components Within the Proposed SDK

In addition to the code libraries described above, whose purpose would be to manipulate CORD-19 data to prepare for text mining and data mining operations, our proposed SDK would bundle numerous applications used for database storage, data visualization, and scripting. The goal of this application package would be to provide researchers with a self-contained computing platform optimized for science related to COVID-19. The components within this application package would be selected with an emphasis on tools that could be distributed in source-code fashion, and then compiled within the SDK's development framwork with few, if any, external dependencies. In short, the SDK would try to eliminate almost all scenarios where programmers would need to perform a "system install"; for the most part the entire computing platform (including scripting and database capabilities) could be compiled from source "out-of-the-box". The SDK would also modify the included applications to enhance their interoperability and their usefulness for COVID-19 research.

The applications bundled with the SDK would likely include the following:

- XPDF: A PDF viewer for reading full-text articles (augmented with CORD-19 features, such as integration with biomedical ontologies);
- AngelScript: An embeddable scripting engine that could be used for analytic processing of data generated by text and data mining operations on CORD-19 (see [5]);
- WhiteDB: A persistent database engine that supports both relational and NoSQL-style architectures (see [8]);
- IQmol: Molecular Visualization software that can be used to study chemical data presented in formats such as PDB which are employed by some CoVID-19 data sets;
- MeshLab: A general-purpose 3D graphics viewer;
- UDPipe: a C++ library for manipulating CoNLL data;
- LaTeXML: a LaTeX-to-XML converter;
- PositLib: a library for use in high-precision computations based on the "Universal Number" format, which is more accurate than traditional floating-point encoding in some scientific contexts (see [4]).

It is worth noting that a data-mining platform requires *machine-readable* open-access research data, which is a more stringent requirement than simply publishing data alongside which can be understood by domain-specific software. For example, radiological imaging can be a source of CoVID-19 data insofar as patterns of lung scarring, such as "ground-glass opacity", is a leading indicator of the disease. Consequently, diagnostic images of CoVID-19 patients are a relevant kind of content for inclusion in a CoVID-19 data set (see [9] as a case-study). However, diagnostic images are not in themselves "machine readable." When medical imaging is used in a quantitative context (e.g. applying Machine Learning for diagnostic pathology), it is necessary to perform Image Analysis to convert the raw data (viz., in this case, radiological graphics) into quantitative aggregates (for instance by using image segmentation to demarcate geometric boundaries and then defining diagnostically relevant features, such as opacity, as a scalar field over the segments). In short, even after research data is openly published by article authors, it may be necessary to perform additional analysis on the data for it to be a full-fledged component of a machine-readable information space.[1] To deal with this sort of situation, our proposed SDK would include a data-modeling vocabulary that would identify the interrelationships between data representations and define the workflows needed to convert CORD-19-linked research data into machine-readable data sets.

Another concern in developing an integrated CORD-19 data collection is that of indexing CoVID-19 data for both text mining *and* data mining. In particular, our proposed SDK would introduce a system of *microcitations* that apply to portions of manuscripts *as well as* data sets. In the publishing context, a microcitation is defined as a reference to a partially isolated fragment of a larger document, such as a table or figure illustration, or a sentence or paragraph defining a technical term, or (in mathematics) the statement/proof of a definition, axiom, or theorem. In data publishing, "data citations" are unique references to data sets in their entirety or to smaller parts of data sets. A data microcitation is then a fine-grained reference into a data set: for example, "the precise data records actually used in a study" (as defined by the Federation of Earth Science Information Partners; see [7]), one column in a spreadsheet, or one statistical parameter in a quantitative analysis.

The unique feature we propose for our SDK would be to combine the text-mining and data-mining notions of microcitation into a unified framework. In particular, text-based searches against the CORD-19 corpus would try to find matches in the data sets indexed by our SDK. As a concrete example, a concept such as "expiratory flow" appears in CORD-19 both as a table column in research data and as a medical concept discussed in research papers; a unified microcitation

[1] This does not mean that diagnostic images (or other graphical data) should not be placed in a data set; only that computational reuse of such data will usually involve certain numeric processing, such as image segmentation. Insofar as this subsequent analysis is performed, the resulting data should wherever possible be added to the underlying image data as a supplement to the data set.

framework should therefore map *expiratory flow* as a keyphrase to both textual locations and data set parameters. Similarly, a concept such as *2'-C-ethynyl* (mentioned earlier in the context of transcription errors) should be identified both as a phrase in article texts and as a molecular component present within compounds whose scientific properties are investigated through **CORD-19** research data, so that a search for this concept can trigger both publication and data-set matches.

III Conclusion

The vision of a *standalone* and *self-contained* **COVID-19** data-set collection is consistent with new publishing initiatives such as Research Objects (see [1]) and **FAIR** ("Findable, Accessible, Interoperable, Reusable"; see [10]). Indeed, our **CORD-19 SDK** would function as a macro-scale Research Object, which would be (1) *self-contained* (with few or no external dependencies), (2) *transparent* (meaning that all computing operations should be implemented by source code within the bundle that can be examined as code files and within a debugging session), and (3) interactive (meaning that the bundle does not only include raw data but also software to interactively view and manipulate this data). Research Objects which embrace these priorities will try to provide data visualization, persistence, and analysis through **GUI**, database, and scripting engines that can be embedded as source code in the Research Object itself. Our proposed **SDK** would embrace the same paradigm, only translating it to a larger data space integrating the information contained in dozens of **COVID-19** data sets as well as the corpus of **CORD-19** articles.

References

[1] Khalid Belhajjame, *et. al.*, "Workflow-centric research objects: First class citizens in scholarly discourse". https://pages.semanticscholar.org/coronavirus-research

[2] "COVID-19 Open Research Dataset (CORD-19)". 2020. Version 2020-03-13. Retrieved from https://pages.semanticscholar.org/coronavirus-research. Accessed 2020-03-20. doi:10.5281/zenodo.3715506 https://pages.semanticscholar.org/coronavirus-research

[3] Ludëk Eyer, *et. al.*, "Nucleoside analogs as a rich source of antiviral agents active against arthropod-borne flaviviruses". https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5890575/

[4] John Gustafson, "Beating Floating Point at its Own Game: Posit Arithmetic", http://www.johngustafson.net/pdfs/BeatingFloatingPoint.pdf

[5] Andreas Jönsson, "AngelCode Scripting Library", www.AngelCode.com/AngelScript/

[6] Amy Neustein, *et. al.*, "Application of Text Mining to Biomedical Knowledge Extraction: Analyzing Clinical Narratives and Medical Literature", https://www.researchgate.net/publication/262372604_Application_of_Text_Mining_to_Biomedical_Knowledge_Extraction_Analyzing_Clinical_Narratives_and_Medical_Literature

[7] Mark A. Parsons and Ruth Duerr, "Data Identifiers, Versioning, and Micro-citation", https://www.thelancet.com/action/showPdf?pii=S1473-3099%2820%2930086-4

[8] Enar Reilent, "Whiteboard Architecture for the Multi-agent Sensor Systems", https://www.thelancet.com/action/showPdf?pii=S1473-3099%2820%2930086-4

[9] Heshui Shi, *et. al.*, "Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study". https://www.thelancet.com/action/showPdf?pii=S1473-3099%2820%2930086-4

[10] Alina Trifan and José Luís Oliveira, "FAIRness in Biomedical Data Discovery". https://www.researchgate.net/publication/331775411_FAIRness_in_Biomedical_Data_Discovery