NIST Internal Report
NIST IR 8479

An Infrastructure for Curating, Querying, and Augmenting Document Data: COVID-19 Case Study

Eswaran Subrahmanian
Guillaume Sousa Amaral
Talapady N. Bhat
Mary C. Brady
Kevin G. Brady
Jacob N. Collard
Sarah Chouder
Philippe J. Dessauw
Alden A. Dima
John T. Elliott
Walid Keyrouz
Benjamin Long
Rachael Sexton
Nicolas J. Lelouche
Ram D. Sriram

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8479

NATIONAL INSTITUTE OF
STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Internal Report
NIST IR 8479
An Infrastructure for Curating, Querying, and Augmenting Document Data: COVID-19 Case Study

Eswaran Subrahmanian
Guillaume Sousa Amaral
Mary C. Brady
Kevin G. Brady
Jacob N. Collard
Sarah Chouder
Philippe J. Dessauw
Alden A. Dima
Walid Keyrouz
Benjamin Long
Ram D. Sriram
Software and Systems Division
Information Technology Laboratory

Talapady N. Bhat
John T. Elliott
Biosystems and Biomaterials Division
Material Measurement Laboratory

Rachael Sexton
Nicolas J. Lelouche
Systems Integration Division
Engineering Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8479

August 2023

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

Publication History
Approved by the NIST Editorial Review Board on 2023-08-04

How to cite this NIST Technical Series Publication:
Eswaran Subrahmanian, Guillaume Sousa Amaral, Talapady N. Bhat, Mary C. Brady, Kevin G. Brady, Jacob N. Collard, Sarah Chouder, Philippe J. Dessauw, Alden A. Dima, John T. Elliott, Walid Keyrouz, Benjamin Long, Rachael Sexton, Nicolas J. Lelouche, Ram D. Sriram (2023) An Infrastructure for Curating, Querying, and Augmenting Document Data: COVID-19 Case Study. (National Institute of Standards and Technology, Gaithersburg, MD), NIST IR 8479. https://doi.org/10.6028/NIST.IR.8479

NIST Author ORCID iDs
Eswaran Subrahmanian: 0000-0002-4639-627X
Guillaume Sousa Amaral: 0000-0001-5241-6508
Talapady N. Bhat: 0000-0002-8396-6622
Kevin G. Brady: 0000-0002-9271-1568
Jacob N. Collard: 0000-0002-4794-8363
Sarah Chouder: 0000-0002-8434-5166
Philippe J. Dessauw: 0000-0002-4565-1232
Alden A. Dima: 0000-0003-0547-3117
John T. Elliott: 0000-0001-5739-7000
Walid Keyrouz: 0000-0003-3807-813X
Benjamin Long: 0000-0002-4340-7674
Rachael Sexton: 0000-0001-5904-2887
Nicolas J. Lelouche: 0000-0002-7031-4887
Ram D. Sriram: 0000-0001-8602-4748

Contact Information
ram.sriram@nist.gov

Abstract

With the advent of the COVID-19 pandemic, there was the hope that data science approaches could help discover means for understanding, mitigating, and treating the disease. This manifested itself in the creation of the COVID-19 Open Research Dataset (CORD-19) which aggregated COVID-19-related scientific literature for use by the data mining community. As a group of interdisciplinary informatics researchers at NIST, we embarked on an effort to use our experience and previously developed systems to explore whether we could enhance the CORD-19 data set and facilitate its use. This effort produced a prototype scientific informatics system that extended data curation, data repository, resource registry, term extraction and indexing systems and resulted in a repackaging of CORD-19 as a Python data package. This paper documents our efforts, provides lessons learned, and proposes a general architecture for these types of systems.

Keywords

CORD-19; COVID-19; curation, infrastructure; informatics; search.

Table of Contents

List of Tables

List of Figures

1. Introduction

Considerable progress has been made in software technology, natural language processing and AI in the last 15 years. Several organizations such as Google and Amazon have incorporated these technologies in their search and query software. Nevertheless, there are still opportunities for improving architectures for managing, curating, and querying data repositories in specific communities of practice. The need for supporting such communities was recognized in 2011 with the inception of the Materials Genome Initiative (MGI) [1]. In response to MGI, the Information Technology Laboratory (ITL) and Material Measurement Laboratory (MML) of the National Institute of Standards and Technology (NIST) collaboratively initiated projects to address user needs for data and document repositories in a scientific community. The COVID-19 pandemic made the need for scientific data informatics infrastructure even more urgent. In this paper, we describe our experiences in addressing research informatics needs with NIST created components. We then propose a generic architecture for curating, querying and augmenting domain specific document repositories.

The pandemic brought a rapid increase in the number of publications addressing COVID-19 and the novel coronavirus (SARS-CoV-2). At the request of the White House Office of Science and Technology Policy [2], the Allen Institute for AI, the Chan Zuckerberg Initiative, Georgetown University's Center for Security and Emerging Technology, Microsoft, and the National Library of Medicine at the National Institutes of Health released the COVID-19 Open Research Dataset (CORD-19) on March 16, 2020. When it was first released, this dataset indexed 29,000 articles that were published in peer-reviewed journals, preprint archives, as well as those available in the public domain. It contained full text for 13,000 of them. It has been growing at a rapid rate, with well over a million articles as of June 2022.

We observed several informatics challenges with this type of dataset. Most of the scientific information is contained as text in the form of titles, abstracts, sections and subsections. COVID-19 is being addressed by multiple domains such as immunology, public health intervention and policy, and emergency medicine. Each of these domains has its own distinctive concepts and terminology which information systems must accommodate. COVID-19 is not unlike many other domains where problems spill over disciplines requiring additional infrastructure that does not exist in any one of the disciplines. The challenges with COVID-19 are the information explosion and the demand for quick consideration and action. In order to address these demands, infrastructure must be integrated across several disciplines to manage and process the huge amounts of text. The issues of scaling and evaluation under this volume and complexity of information need to be addressed as well.

Information made available by CORD-19 should enable researchers and biomedical practitioners to answer biomedical questions.This constitutes a very small sample of questions:

- What mutations enable amino-acids to affect virus assembly?
- Have any small molecule drugs been specifically shown to affect virus assembly?
- How stable is virus RNA in sewer matter?

From the context of having a large volume of text-based information, the community wishes to ask and answer new questions of relevance. We envisage individuals collaborating to ask and answer questions from multiple perspectives with diverging objectives:

1. Experts who are necessary to tune conceptual structures and relationships.
2. Users who ask/answer questions and who fall into three categories: researchers, administrators, and policy makers.

The remainder of the report discusses the essential themes of problems related to technical data usage, the CORD-19 challenge, and related NIST efforts. It is organized as follows. Section 2 is an overview of CORD-19 and related issues and Section 3 gives related work. In Section 4 states our research goals and in Section 5 details the starting points for our research. Section 6 describes our technical work and Section 7 lists our technical contributions. Section 8 details the impacts that the work presented here has had on our current research efforts. Section 9 proposes a general systems architecture for the types of information systems that we have considered in this report. In Section 10, we draw our conclusions and propose future work.

2. Background

This section provides background information about CORD-19 (Section 2.1) and issues that occur with the management of technical data (Section 2.3), some of which we attempted to address in this work.

2.1. CORD-19

In early 2020, a multi-institutional partnership began releasing a rapidly growing data set of COVID-19-related scientific papers and associated metadata [3]. While CORD-19 was created in response to COVID-19, it also contains earlier scientific papers of interest such as those concerned with past epidemics [4] and some non-scientific COVID-19-related articles. The data set was also introduced with a set of 10 key COVID-19-related questions developed in coordination with the National Academies of Sciences, Engineering, and Medicine's Standing Committee on Emerging Infectious Diseases and 21st Century Health Threats and the World Health Organization with the hope that the AI community would be able to help answer them [2]. These were then made available to the public via the widely-used machine learning community website, Kaggle [5], which we will discuss in Section 2.2.

The goal of the CORD-19 effort has been to bring together the machine learning and biomedical communities in an attempt to accelerate the discovery of COVID-19 etiological, treatment and management strategies. There was also the hope, post-pandemic, that the CORD-19 dataset will provide inspiration for new uses of machine learning in scientific research [6]. CORD-19 has enjoyed wide interest and has played a role in multiple COVID-19 research efforts, resources, and challenges that we describe in the next two sections [3, 4].

2.2. The Kaggle CORD-19 Challenge

The open nature of the Kaggle competition allowed us to observe diverse stakeholders, their goals, and the strategies these stakeholders employed to achieve those goals. We noticed repeating themes that centered on individuals and groups trying to solve problems, sharing resources and knowledge, and helping each other along the way.

2.2.1. Diverse stakeholders, goals, and strategies

The community forum showed participants demonstrating a significant range in technical skill, subject matter expertise, orientation to their given resources, and orientation to the larger problem.

Some sought to share emerging knowledge of tools or resources associated with CORD-19, such as search engines, browsers, and more. Others sought to create or find data sets [7] particular to their questions. Some built one-off tools, like browsers, question-and-answering notebooks, report builders, dashboards, analyses, plots, and more.

The main activities were organized around a competition having several rounds wherein, those participating in the rounds would strive to produce winning submissions for high-value tasks. Some participants who showed up for the CORD-19 challenge appeared to have mixed feelings regarding whether they wanted to compete or simply try to solve particular problems of their own choosing. Folks who jumped into data analysis activities split off into a number of different lines of inquiry. Some spent time on a number of activities such as converting data into word embeddings, using provided or specialized embeddings, filtering, labeling, and searching the dataset, as well as applying various techniques ranging from named-entity recognition, to chemical structural similarity detection, to creating or sharing various resources. Some sought to form or use knowledge-graph based representations of the data or to do topological data analysis. Others sought to identify ways to extract or infer information, such as study designs or statistical data from text.

2.2.2. Duplication of efforts

A number of routine setup and analysis tasks frequently recurred. These included loading the data [8, 9], converting data from JSON to CSV and text formats [10, 11], multiple keyword text searches [10], and using word embeddings [12].

2.2.3. Mechanics

The discussion forum showed a flurry of activity following a number of channels. Many folks sought to plug into the competition platform framework, to use its infrastructure, load the data, install well-known libraries, and start their particular investigations. A number of participants discussed difficulties they experienced when getting started with the dataset [13–22]. Some of these were also discussed as the following kinds of topics: issues with data size being too large [23], corrupted data files [24], duplicate entries [25, 26], confusion about the data [27], parsing errors in the JSON data [28], conversion of JSON data [11], and more. These included questions about the dataset such as how it was originally parsed into JSON [29], how the JSON files were linked to the metadata [30], what tools were used to convert PDF files to text [31], how one might reproduce the entire CORD-19 dataset from scratch [32], how one might enrich it [33], and how one might evaluate it [34].

2.2.4. Collaboration: Helping and sharing

Collaborations arose during the competition. Some external organizations sought to donate free computational resources to the effort [35–37]. An epidemiologist and other subject matter experts offered to help annotate data [38]. Others sought to mentor and guide individuals by giving examples of particular problem templates.

A number of individuals repeated and shared snippets of code [39], mostly from one notebook to another, while striving to get more quickly to their desired activities: to analyze the data. Some GitHub repositories also linked to these efforts.

2.2.5. Curation, Organization, and Reusable Building

At the end of it, a few notebooks sought to collate community responses to some of the challenge questions. The remaining 1600 notebooks serve as a record of specific, mostly individual investigations [40]. But it was not clear how to reuse, access, or expand on what had been done in a more scalable manner. A number of questions remain for this community effort and for many others like it. What is really necessary in order to organize a community of researchers and resources in a reusable and scalable manner toward solving a complex, large-scale analysis problem?

A much smaller set of individuals on the same forum also sought to ask questions about how one might try to engineer solutions to some of these problems more substantially, posting suggestions such as how to move in this direction by building toolkits or frameworks to help reuse solutions for common problems as well as how to evaluate their quality [41–48]. In the same spirit, others suggested gathering user stories about what problems were really being solved and how to validate them with health professionals [49]. Another asked how one might come to know what findings in the forum would be most useful to COVID-19 researchers. Most of these posts did not receive responses as the community seemed mostly focused on the urgency of responding quickly to the emerging problems. In addition to the contributions that resulted from all of these activities, in this paper, we strive to also entertain a view that can devote some attention to the larger question of how to scale such diverse stakeholders, goals, and strategies.

2.2.6. Impact and future of CORD-19-related activities

When attempting to analyze the outcomes and impacts from the CORD-19 data set and its activities, there are some signs of essential outcomes. However no comprehensive analysis has been performed in the literature regarding the ultimate outcomes. Therefore, without direct access to all the details that gave rise to this data set and these activities, any assessment of the genesis of the project, its goals, and its impacts will have to rely, in some way, on historical and contextual speculation. The data set itself was described and introduced [43, 50] and was also documented on both GitHub [51] and SemanticScholar [52]. Wang and Lo [7] performed a retrospective analysis following the 2020 CORD-19 Kaggle competition [53]. The CORD-19 maintainers also provided commentary as they notified the CORD-19 community of the sunsetting of the CORD-19-related activities [54].

From these resources, one obtains a minimal insight into these activities regarding their target audiences, their primary users during the lifetime of the dataset [43], and what they saw as ultimate outcomes, products, and results of these activities [7]. In all, their self-evaluation was positive and encouraging, noting the degree to which the project facilitated the engagement of biomedical-researcher communities, the AI/informatics research communities, AI/data-mining competition participants, and related tools and projects. They also noted additional advances such as the outcome of being able to include and encapsulate lessons learned from these activities into increased functionality in existing search engines and services (like SemanticScholar) and other tools (such as LitVid). In contrast to the above context, if all of these activities and their outcomes were to be evaluated from the standpoint of an enhanced R&D-community that could scale to handle such problems, to our knowledge, such topics have not yet really been pursued in depth in the primary CORD-19 literature. As a result of the availability of world-wide COVID-19 research literature, it became clear that it was difficult for researchers to keep up with the latest developments [55]. Summarization tools were created to strive to help researchers in summarizing emerging highlights as quickly as possible [56, 57]. However, even with these tools, to our knowledge, no up-to-date analysis has yet been performed that comprehensively unifies what is known about COVID-19 relative to this literature base.

2.3. Gaps in Technical Data Management

CORD-19 is a very large and complex data set and anyone working with it will immediately face data management issues. Data management often refers to planning to meet well-defined intent and maintenance goals for static data sets (as used in reference to organization or research-level data management plan requirements [58, 59]). Yet, practical data management planning is increasingly plagued with curation and integration issues. Organizations following best practices in comprehensive data management still fall prey to inefficient workflows involving manual cleaning, pre-processing, and tool integration. These issues are compounded at the individual research project level, wasting further time and effort. In turn, this has prompted large-scale efforts to promote better planning practices (for instance, the FAIR data principles) under which data discovery and reuse is planned and measured more effectively [60]. CORD-19, being a continuously released cumulative data set, promises to only compound data management issues.

Even with a better set of guiding principles, a lack of comprehensive tooling makes implementing them difficult [61]. What is desired is automation that supports end-to-end data management. Modern data management stakeholders are beginning to adopt operations methodologies from other digital domains (e.g. Development Operations, or “DevOps”) in an attempt to fill this need for automation. Data Operations (DataOps) attempts to treat data as a software dependency, as it often is in the current data-driven climate [62–65]. It describes tools and approaches to data management which seek to enable a holistic approach toward continuous, iterative, and modular data usage [66]. One of the contributions of this work is cv-py (Section 6.3) which strives to provide this automation for CORD-19 end users.

3. Related Work

In the sections below we will look at related scientific informatics efforts to address the COVID-19 pandemic. Section 3.1 focuses on work related to information retrieval, while Section 3.2 looks at other COVID-19-related data curation efforts. Section 3.3 concentrates on term extraction and Section 3.4 will look at scientific discovery techniques and tools.

3.1. Information Retrieval

One of the goals for CORD-19 was its use as the basis for a variety of COVID-19 related search engines. Google's COVID-19 Research Explorer [67] uses neural network-based information retrieval to perform focused semantic searches of CORD-19 articles to answer natural language questions. Its focused searches reduce a user's need to sift through non-relevant articles and it provides further assistance by highlighting evidence in the results and by allowing for follow-up questions. Amazon Web Services has also made available a CORD-19 search engine, AWS CORD-19 Search (ACS) [68] that uses natural language queries. It is built from existing AWS technologies including a semantic question and answer system, medical entity recognition and relationship extraction, graph models, and topic modeling. Neural Covidex [69] is a CORD-19 search engine that uses a neural ranking architecture to provide paragraph-level retrieval of relevant articles with salient sentence highlighting. It was rapidly developed in about two weeks and forms the basis for two interrelated online search tools, one is a faceted keyword-based search engine and the other deep learning-based. COVID-SEE [70] uses Neural Covidex and attempts to go beyond article retrieval to support literature discovery from multiple perspectives via domain-knowledge-based visualization techniques. CO-Search [71] performs ranked document retrieval using keyword and transformer-based models. While it does not provide new methods, it is made available as an open-source project with the hope that others will build on it. SPIKE-CORD [72] provides an advanced query system for CORD-19 that allows users to search for specific linguistic patterns and expands on simple Boolean searches with sequential and structural query options.

The importance of information retrieval in combating the pandemic is underscored by the TREC-COVID [73] community evaluation which is creating test collections to evaluate different modes of information retrieval from the CORD-19 data set as it evolves. The goal is to help discover new means of managing rapidly changing information during this and future pandemics. Participants in TREC-COVID have published their lessons learned [74].

CORD-19 does not provide relevancy measures that help users identify which papers are relevant to its associated questions. Heaton and Mitra [75] have investigated re-purposing the TREC-COVID annotations to address this. They trained a BioBERT model to predict article relevance that achieved a Cohen's kappa of 0.443 with the most commonly selected human annotations (majority agreement).

3.2. Data Curation

In the context of this paper, the Configurable Data Curation System (CDCS), which is discussed more fully in other sections, is a primary component of our approach to curation problems facing informatics researchers. It provides an online, federated platform for entering, structuring, validating, finding, sharing, accessing, presenting, linking, and using research data via manual or automated means, using XML and JSON as primary data formats. There are some systems that seem to have similar subsets of functionality such as CORDRA[1] which offers a JSON-based serialization of objects, 4Ceed[2], and T2C2[3]. More broadly speaking, many tools and platforms provide

[1] https://www.cordra.org/

[2] https://4ceed.github.io/

[3] https://t2c2.csl.illinois.edu/

one or more of the above-mentioned general functions to access, transform, or manipulate data in many different contexts in some manner. However, it is much less common to find all this functionality provided together, as open-source, easily tailored and scaled across communities, in the manner provided by CDCS. The ideas in CDCS are fairly simple such as providing basic hosting, searching, and manual/programmatic manipulation of data objects in XML and JSON. Over time, the general functions and capabilities in CDCS and others will likely diffuse across similar projects with the CDCS core functionality being continually evolved and differentiated over time by its unique community of user research-projects. CDCS and other systems are all part of the portfolio of tools that are in use by researchers at NIST. As such, the CDCS only represents a system that is being developed by this team and applied at NIST in an eco-system supporting many other tools with some similarities and differences. While CDCS has a number of applications, such as its use in various laboratory information management system (LIMS) scenarios[4], some other systems, such as 4Ceed, have tailored their functionality toward LIMS applications as well[5]. CDCS strives to stay in close-contact with its research-project base and evolves more in line with researcher applications or needs rather than striving to specialize in any one application area. Thus, while the CDCS shares some similarities with some of these platforms – such as its similarity providing persistent identifiers and linked-data functionality to CORDRA's functionality for digital object identification and handle-related operations – CDCS also continues to grow additional analytical functionality in its core applications (e.g., support for ElasticSearch indexing, custom R&R integration) as well as to expand its integration with scalable middleware in its architectural development.

3.3. Terminology Extraction

Terminology extraction or keyphrase extraction is the process of identifying important phrases from a document or corpus. There are generally two parts to terminology extraction: identifying valid units and identifying which of those units are important. There have been many proposals for handling each of these subtasks, which can often be combined in different ways. For example, Tomokiyo et al [76] compare a few statistical methods for both identifying phrases and ranking them by informativeness. Chuang et al. [77], however, use regular expressions over parts of speech and identify other useful features such as frequency, position, and phrase length. TextRank [78] uses a graph-based algorithm to identify important phrases.

More recent methods combine these two basic steps by borrowing from named-entity recognition, in which phrases are identified directly by a single algorithm. For example, DyGIE++ [79] uses contextualized embeddings to identify the start and end of important phrases.

3.4. Scientific Discovery Techniques and Tools

Many efforts have investigated the use of CORD-19 to accelerate COVID-19-related scientific discovery; we can only describe a few below. Tyagin et al. [80] describe two CORD-19-based automated hypothesis generation systems. These two systems, one based on graph-embedding, the other on transformers are adapted for queries at different scales to address pressing pandemic-related research questions.

[4] e.g., https://nexuslims.nist.gov/

[5] https://doi.org/10.6028/jres.124.034

Cernile et al. [81] built a network graph representation of CORD-19 article title and abstract data using additional information from medical databases to identify relationships between diseases, medications and procedures. The goal of their study was to demonstrate the rapid development of such a network graph, its usefulness in understanding a large corpus and the hidden relationships that it contains.

Reddy et al. [82] demonstrates an end-to-end question answering system based on information retrieval and natural language understanding. They addressed the lack of annotated COVID-19-related text by using a synthetic example generator as part of a scheme to fine tune an existing neural model to CORD-19.

3.5. Data Operations and Infrastructure

Several tools have already been developed to support data infrastructure around biomedical and scientific text. In our opinion, a very good example of this is scispaCyhttps://allenai.github.io/scispacy/, a package developed by the Allen Institute for AI.

ScispaCy is a Python package that contains pre-trained spaCy models for scientific corpora. SpaCy is a Python-based natural language processing package that makes it easier for Python developers to pre-process and analyze general text. ScispaCy specifically applies the spaCy toolkit to biomedical text by pre-training spaCy NLP models with different annotated corpora. This gives machine-learning researchers an easy way to process other biomedical and clinical data as a significant portion of the pipeline is already pre-built.

These out-of-the-box pre-built pipelines and models are of specific interest to us since we believe that this is an important part of solving issues of transparency and reproducibility that often plague researchers. The common base offered by scispaCy is an excellent baseline for researchers who want to gather insight. The base should be constantly be updated and improved (as is done by the sciSpacy developers) so that researchers and other developers are always working with state-of-the-art data and pipelines.

Other more general data infrastructure tools are Hugging Face transformers. In this context, transformers are machine learning tools (e.g., tokenizers and models) that can come pre-trained using a variety of training sets for researchers to use out-of-the-box and thereby offer a reproduceable framework for analysis.

One issue with these approaches however, is that the annotated corpora that are the foundation of the pipelines and the pre-built models do not evolve. New models are routinely added to scispaCy for example, but the models that already exist are not updated as the community learns.

4. Our Goals

In the context of COVID-19, our aim is to bring together tools and techniques that can be combined into processes that support:

1. Discovery and use of diverse data and document resources,
2. Enhanced search based on terminological structures, both taxonomic, topical, and ontological, implicit in the corpus, but made explicit for search [83], and

3. Visualization of the corpus landscape using taxonomies, semantic schema, topic models, topic trends, and knowledge graphs.

Beyond COVID-19, we wish to facilitate future research efforts that rely on the analysis of text-based data.

4.1. User Questions, Use Cases and Functions

Given the impact of the pandemic, an immediate goal is to leverage CORD-19 to facilitate the discovery of actionable information about COVID-19. One aspect we wish to address is to enable the quick answering of key questions.

We will now examine potential uses for our infrastructure to address COVID-19. We will begin with some questions that might be posed by an administrator, a decision maker and policy maker. These questions were inspired by use cases for NLP and AI tools in clinical information services [84].

Example Researcher Questions (both historical and exploratory)

- How do we search for specific articles based on the concepts of interest?
- How can we identify concepts and their evolution over time?
- How can we understand concepts from different perspectives?

Example Science Administrator Questions

- What has been the trajectory of the field over time?
- What were the events that drove the field?
- What are promising new areas to be explored?

Example Decision Maker and Policy Maker Questions

- Which strands of emerging work are promising?
- What decisions were made and what evidence was used?
- What are potential cross-disciplinary collaborations?

The above questions are representative examples and are not meant to be an exhaustive list. They only indicate the types of problem-solving that we wish to support. Table 1 provides a list of use cases that our prototype infrastructure supports. These cases provide concrete examples of the different potential uses of our infrastructure involving different components, actors, and types of data. The use cases are explained in more detail in Appendix A. Each use case is given an identifier to allow it to be referenced succinctly.

4.2. Data Operations and cv-py

One of the main issues encountered by the data community is the very few holistic approaches that exist to gather, pre-process, and analyze data in a repeatable and transparent manner. The steps are traditionally split up in distinct code packages and datasets, and it is up to the end user to mix and match the correct packages and datasets to get insight. We believe that there is a need for easy,

ID	Summary	Actors
CURATE-1	Curate published data	Publisher, Curator
PUBLISH-1	Register published data	Publisher
SEARCH-1	Search for datasets	Researcher
SEARCH-2	Search for datasets with metadata	Researcher
SEARCH-3	Keyword search	Researcher
SEARCH-4	Systematic literature review	Researcher
SEARCH-5	Find answers to research questions	Researcher
TAXONOMY-1	Explore conceptual taxonomies	Researcher
TEMPORAL-1	Explore usage over time	Researcher
TOPIC-MODELING-1	Explore topic models	Student, Researcher
TEMPORAL-2	Answer meta-scientific questions	Student, Researcher
KG-1	Explore knowledge graphs	Student, Researcher

Table 1. Some example use cases for our prototype infrastructure.

reliable access to resources with low-friction interfaces to text and associated data. Versioning of data products and data pipelines must co-exist in tandem with the models and frameworks needed to process them. Processing should rely on meaningful pre-configured defaults that support the notion of "optional by design" where features do not get in the way if ignored. We address these issues for CORD-19 with our cv-py package (Section 6.3).

5. Starting Points

Three prior efforts at NIST that inform our vision for exploration of scientific/technical literature are:

1. Configurable Data Curation System (CDCS)
2. Root and Rule Based System (R&R)
3. Model Based Enterprise for Manufacturing and Maintenance Data

All three systems are further discussed below.

5.1. Configurable Data Curation System (CDCS)

CDCS [85] is a configurable system for the creation of searchable repositories and registries containing resources curated using user-specified templates.

CDCS was developed in response to the Materials Genome (MGI) and serves as a framework for reusable syntactic components that alleviate the need for developers to create custom data formats for technical data along with their associated parsers and transformation tools [86]. It provides a means for capturing, and sharing that is amenable to transformation to other formats. The data are organized using user-created templates encoded in XML Schema that are also used to create data entry forms whose data are saved in a non-relational document-oriented database. CDCS is designed for a wide variety of users and allows for both manual and automated access to its functionality. It is currently being used by a number of research efforts.

CDCS provides the facilities for composing community data-networks where data can be integrated with custom applications and workflows via unified interfaces for accessing, organizing, sharing, and searching. When configured as a registry, CDCS allows users to find curated domain-specific resources. The information describing these resources is maintained by individuals, projects, and organizations in a community. The registry serves to aggregate and organize a community's resources.

5.2. Root and Rule Based System (R&R)

The R&R terminology generation system [83] is a framework for constructing human- and machine-readable representations of natural language text. The R&R system constructs normalized representations for each phrase in the input text, and indexes the resulting representations to relevant sentences, sections in the original documents.

The representations provided by R&R are constructed according to a linguistically-motivated algorithm to provide useful representations for precise search and discovery. These representations serve to both *normalize* and *disambiguate* natural language, and to make explicit the implicit semantic relationships that hold between words and phrases.

In order to produce these representations, R&R uses a combination of statistical and symbolic methods. These are arranged into a customizable pipeline, allowing different components to be replaced or modified in order to make the system more suitable for new data or domains. Most of this pipeline, other than parsing can be performed without the need for annotated training sets (and many pre-trained parsers are available). Re-training the parser can be a time-consuming and expensive process, since syntactically annotated datasets must be constructed. However, all of the other components of the system provide customization options that allow for the *ad hoc* modification of the results, and the default configuration is often sufficient. This allows for changes to be made according to observation and reasoning, making the changes predictable.

5.3. Model Based Enterprise for Manufacturing and Maintenance Data

Data infrastructure and the issues that it attempts to solve are applicable to more than just the COVID-19 pandemic response. These problems and patterns have been noticed in other areas, and at NIST's Engineering Laboratory, an area where we have encountered these previously is for manufacturing data using the Smart Manufacturing System (SMS) Test Bed.

The SMS Test Bed research project provided valuable experience to our projects regarding needs associated with collecting, sharing, using and re-using data to gather insight. In the context of CORD-19, it was inevitable that some of these same issues would show up, namely:

- Having data sources that evolve, but these changes are not tracked. This means that there can be confusion regarding which version of CORD-19 is in use for a specific pipeline.
- Pipelines themselves can be unclear and disjointed. Pipeline reproducibility (i.e., re-generating insights from data) in general is an area where there is significant room for improvement.
- Potential solutions to these issues often encounter institutional resistance both from developers, since workflow changes are hard to enforce, and system administrators, since benefits

are not necessarily readily apparent from an IT infrastructure point of view.

These are all issues that the Information Testing and Modeling group at the Engineering Laboratory at NIST has experienced first-hand. The issue of transparency and reproducibility is especially important in the context of the COVID-19 pandemic, since policy decisions that are influenced by unreliable insights can have undesirable consequences.

A general conceptual solution to some issues is described in the FAIR data principles outlined by Wilkinson et al [60]. They argue that data should follow four core principles: Findability, Accessibility, Interoperability, and Reuse. These principles, broadly regarded as being a core part of DataOps, will guide any potential solution to these issues.

The DataOps concepts and philosophy more generally offer themselves a potential workable solution to issues encountered here. In fact, at NIST, DataOps has become more prominent in recent years, with the SMS Test Bed being an excellent example [87]. The SMS Test Bed works with the NIST Fabrication Technology Office to mimic the configuration of a contract-manufacturing shop with several fabrication machine tools and inspection equipment.

One of the goals of the SMS Test Bed was to solve issues that come up regularly, which include different data formats and data and communications protocols, obsolete operating systems, large data volumes over a large range of temporal scales, and demanding limitations of physical environments [88].

The data generated by the Test Bed were thus collected, aggregated, and published via three different web services: (1) a volatile data stream, (2) a query-able data repository using NIST CDCS, and (3) pre-compiled data packages.

What is interesting about the pre-compiled data packages is that the pre-processing done before publication does some of the work traditionally done by end users by associating different types of data. This takes out one step in the end-user workflow, and aims for a more transparent and reproducible pipeline to gather insight by lifting the baseline from which end-users start.

However, the SMS Test Bed encountered significant issues with the way it published data, since its data were published directly on GitHub. Due to constraints at the time, this means that as data were generated, they were published raw to the Test Bed repository. To this day, there are thousands of files stored this way in the GitHub Test Bed repository. To be clear, this was a significant step in the right direction since data is versioned and released regularly, but using git and GitHub for data sets may not be the best solution for two categories of reasons:

1. From a technical stand-point, git and GitHub are not made to handle heavy data files. GitHub repositories are supposed to stay below 1GB in total size, and GitHub support may ask for corrective action to the repository if the repository size grows beyond 5GB [89]. Additionally, individual files cannot be larger than 100MB. One way to get around this is to use Git-LFS, but file size is still capped at 5GB using the most expensive subscription service offered on GitHub [90].
2. From a researcher stand-point, git loses utility for researchers through this kind of use. The point of using git and a git workflow is to make use of the different tools offered. Diffs offer limited use on JSON and CSV files, and many features that are key to good data usage are

not included in typical git hosting services (e.g., pipelines and metrics) [91, 92].

These issues point to the difficulty of finding a general solution to the issues outlined. This is compounded by the current data management tools, that make comprehensive implementation of the FAIR principles difficult [61].

The experience gained from efforts such as the SMS Test Bed has thus helped inform us of the need for improved data management.

6. Technical Work

In order to accomplish our goals, we need to overcome some hurdles which we will describe in the sections below.

6.1. Data Curation

Data curation is a pillar of our CORD-19 infrastructure effort. While our CDCS has proven capable at data curation, we needed to enhance it to handle CORD-19's scale and text format in addition to utilizing and augmenting the dataset. We describe these enhancements in the following sections.

6.1.1. CDCS Enhancements

At the time of this writing, CORD-19 contains over a million articles. In order to process and provide the CORD-19 dataset as a service, the CDCS system was enhanced to scale in several ways. New modules were introduced to perform tasks such as cleaning or normalizing syntax errors found in canonical CORD-19 records and augmenting that data with keywords, full-text, and links to articles, licenses, authors, and institutions. The dataset was made available for use both online (via the CDCS system) as well as offline through downloads of the updated dataset versions. The R&R process and technology were integrated using the Parmenides search-based application, the introduction of augmented search and indexing capabilities, and the development of streamlined data loading components. The COVID-19 system performance was also evaluated via measurement and profiling frameworks.

6.1.2. Text Reconstruction

CORD-19 was initially provided on Semantic Scholar and Kaggle as five primary files – a metadata file and four files containing article full-text which was updated weekly. The article text was provided as JSON documents in which fragments of text of varying sizes needed to be properly reassembled to be useful for analysis. An early version of CORD-19 contained 29,000 articles distributed among 1.7 million text fragments, many of which were small with less than 20 characters (e.g., "30 countries [1]"), while a few were large with some exceeding 50,000 characters. Multiple languages were present: including English, Italian, French, and what appear to be Unicode escape sequences for Chinese. Though those who analyze data are accustomed to spending the majority of their time preprocessing the data, we saw evidence on Kaggle that those working with CORD-19 eliminated the many articles that they couldn't decipher. We realized that simply converting the CORD-19 JSON documents into XML during the curation into the CDCS would not be helpful. We opted instead to reconstruct the contiguous text for each article.

6.1.3. Normalizing and Augmenting the Data

A necessary and crucial step in preparing data for analysis and interpretation in applications is ensuring that it is appropriately formatted and that it conforms to a well-formed data structure of some type.

In general, CDCS users will often ensure valid data by encoding their data into XML documents that conform to a valid XML Schema. Thus, a CDCS-CORD19 schema was created for initial offerings of CORD-19 and was continually updated until the format reached stability.

After the schema was developed, CORD-19-specific quality control and error correction processes were established to minimize syntactic noise. This involved removal or correction of malformed records, supporting Unicode for textual data, augmenting provided data records with URLs to licenses and article sources, augmenting the base data with interlinked records for publications, authors, and institutions, as well as integrating cleaned text and extracted keywords.

The result is a set of validated XML records that contain augmented information which is uploaded to the CDCS platform for community use.

6.2. Representing the Data

Given both the size and breadth of CORD-19, information retrieval tools for the data need to be able to handle a variety of query types for different use cases. CORD-19 includes data not only from medicine, but also social sciences, technology, and even arts and humanities [93]. While some users may want very broad searches, others may need more control over the context in which their search terms appear.

We chose to implement two different types of query: basic web portal queries and rich structural queries. For web queries, we wanted to provide a simple, minimal interface that provides a basic keyword search. For structural queries, we wanted to allow advanced queries that make reference to linguistic structure in the data.

These two query types require different data representations. The basic keyword search depends primarily on text, while the structural search requires annotations such as part-of-speech tags, syntactic dependencies, and word and sentence boundaries. Some additional data may benefit both searches, such as author names and publication dates.

Though the simple queries require a simpler data representation, it is by no means trivial to construct. The representation needs to be able to handle alternate formulations of text, including differences in inflection (e.g., plural forms or verb agreement) and word order. This information is available in the structural representation, but a performance hit comes from searching over this more complex data.

As a result, we produce two different data representations: first, we produce a rich, structural analysis using the R&R framework. This produces a database describing linguistic features of CORD-19. Then, we construct from this a simplified representation with basic normalization, but no further linguistic annotations.

Our first representation is ideal for researchers with complex needs. It provides context and reusability, making it possible to build new applications off of the initial representation. We have also developed tools for customizing the data representation, since we recognize that most linguistic data representations are not atheoretical and may not meet all needs.

The second representation provides a fast and simple interface for basic search queries. This supports fewer query types than the structural representation, but is much faster and requires no technical background to use. Furthermore, this representation can still provide interesting research analysis of data, such as the evolution of terms over time in the data.

6.3. Data Operations and cv-py

The issue of treating data as code, as a dependency on which the rest of the analysis and code-base relies, lies at the heart of the data management problem.

The goal of the cv-py package was to find a way to merge all the necessary steps to gather insight from data in a single Python package. This was done by implementing DataOps concepts, most importantly including FAIR principles [60], in service of the COVID-19 pandemic response. In doing so, cv-py has scouted out issues with DataOps at both a technical and institutional level at NIST.

One of the main milestones was to provide a data pipeline, including all necessary code and library dependencies to do a base-level NLP analysis of the CORD-19 corpus, the analysis itself and presentable results to the user immediately out-of-the-box.

During this effort, we nonetheless experienced several obstacles:

- The Engineering Lab at NIST has little experience with DataOps, and issues regarding approval for use of different data infrastructure tools have come up. NIST IT administrators are keen on helping researchers accomplish their goals, but there can be a disconnect between the benefits the machine learning researchers see in adopting these tools, and what system administrators can interpret as being the next flavor-of-the-month DevOps gadget.
- The research teams themselves were not used to DataOps workflows, and having to change development habits can be frustrating. This is one of the primary issues that can occur: a lack of adoption by researchers can happen if the shift is too abrupt and the benefits are not obvious.
- Publishing data in a standardized way is not a trivial problem. As discussed previously, the NIST Smart Manufacturing Test Bed (SMS) project's solution was to share thousands of files on a GitHub repository. Our preliminary solution for the CORD-19 use case was to bundle the data as a Python package for analysis and use by other tools, but a general solution is not in the scope of this project.
- cv-py is currently dependent on other packages and third-party code.

An important part of the technical work was the start of the integration of DVC (Data Version Control) [65] to store and version data sets. This tool has begun to help alleviate the issues outlined in 5.3: it allows researchers to version their data exactly like git does with code and text. This helps track data set versions as they are modified and enhanced by researchers, which allows for easier re-use.

The open-source nature of the project allows for incremental improvement of any of parts of the package, so that it may grow with the biomedical community. A short-to-medium term goal would be to grow all three parts of the project infrastructure (data shareability and pre-processing, analysis, result) as a single growing shareable package through community-based code improvement.

7. Contributions

As a result of this effort, we applied our skills in natural language processing, text analysis, and information retrieval to prototype an infrastructure that provides CORD-19-based resources to the COVID-19 R&D community. The essential elements of this infrastructure, presented loosely in the order of the data flow, include:

- **CORD-19 Normalization, Augmentation, and Integration Process** – the collaborative workflow instantiated with our partners to integrate the CORD-19 resource into the CDCS community data-network
- **CDCS COVID-19 Repository** – CDCS data node for hosting each CDCS CORD-19 release
- **CDCS COVID-19 Registry** – CDCS registry for registering and finding important COVID-19 research resources in the COVID-19 community data-network
- **Parmenides** – R&R indexing service integrated into the CDCS data nodes for search by conceptual relationships
- **R&R Term Extraction and Knowledge Indexing Service** – identifies and indexes significant terms and relationships from text
- **R&R Search Engine Service** – for performing searches based on these knowledge indices
- **R&R Knowledge Graph Modeling Service** – for representing knowledge in conceptual-graphs based on extracted knowledge-structures
- **R&R Trend Analysis Service** – for performing trend analyses based on these knowledge structures
- **cv-py** – pipeline for CDCS CORD-19 data release and data science tool integration

The following sections further describe some of these elements.

7.1. COVID-19 CDCS Repository and Registry

With the onset of the COVID-19 pandemic, by leveraging the CDCS, we were able to quickly apply CDCS to the COVID-19 research domain where the situation appeared very similar to that of materials science: very heterogeneous communities and resources that lacked integration and which are rapidly developing and emerging, resulting in a significant fragmentation of data, resources, and computation. The team began to apply the CDCS platform and methodology to this problem by integrating with our collaborators' specialized resources and services for knowledge extraction, modeling, and analytical computation.

With the announcement of the availability of the CORD-19 dataset, the COVID-19 CDCS Repository was created and customized to support the integration of the CDCS CORD-19 research literature into the CDCS data network.

The COVID-19 CDCS Repository is synchronized with the CORD-19 dataset releases through a periodic update process. This process begins with the publication of COVID-19-related research papers via multiple venues. The CORD-19 curators harvest the articles and associated metadata to augment the data set. This data is released in a JSON format. We download this dataset from the CORD-19 site and then process it. Our efforts include reconstructing the article text, extracting keywords using one of the Allen AI language models, normalizing and cleaning the metadata, as well as augmenting it with links to the original papers, their licenses, and institutions. We merge this data and metadata and convert it into XML using a schema that we created. During this process, each paper, author, and institution is assigned a persistent identifier (PID). These PIDs are used to generate inter-record links between authors, authors and papers, and papers and institutions. We also create term/phrase indexes using the Parmenides CDCS implementation of the Root and Rule (R&R) service. These data, metadata, and indexes are loaded into the COVID-19 CDCS repository to create our overall cdcs-cord19-dataset.

Once it is available from the CDCS, this augmented CORD-19 data can be exported in alternative formats such as CSV, XML, and JSON. The data is also available from the cord19-cdcs-nist GitHub repository and via the cv-py package which links with the cord19-cdcs-nist repository to provide the latest release via an API.

The COVID-19 Registry is a web application created in parallel to the COVID-19 CDCS Repository and collects resource descriptions focused on the COVID-19 R&D community. These resources can include repositories, databases, services, portals, websites, organizations, and other items of interest to this community. Research community members may contribute to the registry using either the web interface or the automated application programming interface (REST API). Resources may also be harvested from other registries. Due to the design of the CDCS registry as easily inter-connectable community infrastructure, this registry has the potential to develop into a global, comprehensive registry of COVID-19-related resources.

7.2. NIST Scientific Indexing Resource for COVID-19

The NIST Scientific Indexing Resource for COVID-19[6] is an online resource that provides a simple search interface into CORD-19 data using R&R to power its indexing. Similar to CDCS, it was possible to launch the NIST Scientific Indexing Resource relatively quickly after the announcement of CORD-19. The NIST Scientific Indexing Resource relies on the same core data as CDCS, and the two systems are integrated to make use of each others' workflows.

The NIST Scientific Indexing Resource is updated iteratively as follows, and operated on the latest data from CORD-19:

1. The latest CORD-19 dataset release is acquired from CDCS.
2. New data is normalized and document IDs are deduplicated, resulting in an index mapping linguistic phrases to the original text.

[6] https://randr19.nist.gov

3. The resulting database is exported in full form and in simplified form.
4. The simplified database is used to power the web portal and to provide keywords for the CDCS interface, while the full database is used for advanced research queries.

7.3. cv-py

The end result of the cv-py project is a low-friction interface to CORD-19 through an installable Python data-package, similar in principle to the way pre-trained NLP models are distributed in the various corresponding libraries (eg. SpaCy, gensim, flair, nltk). This is intended to provide smooth access to the collection of publications for researchers and analysts, along with validated tools for preliminary pre-processing and interface to various formats. Researchers and other CORD-19 stakeholders are, therefore, able to import curated data as easily as they would any Python package. The three main features of cv-py are:

1. Tasks-as-data, meaning the CORD-19 challenge task questions are accessible directly through code as data.
2. Public API helper-functions, using state-of-the-art neural question answering to search for relevant CORD19 passages with a single query to Korea University's covidAsk model [94].
3. Scalable, fast, versioned access to data by leveraging NIST's CDCS using Dask and a read-optimized Apache Parquet storage format. At time of publication, cv-py packages two data sets: the CORD-19 CDCS and the biomedical data sets included in scispaCy.

8. Impacts on our Research

We will now describe how our current research has been impacted by what we have learned from the CORD-19 infrastructure effort.

8.1. Data Curation

Processing the CORD-19 dataset has provided the team with a valuable use-case for interacting with high-volume, high-throughput data. CORD-19's size and complexity have given the development team a basis for studying and improving the scaling and performance of CDCS operations. Data loading and exploration were the main targets of the performance improvements.

Our role in the CORD-19 efforts was largely the introduction of additional data quality and augmentation into a high-value research dataset. CDCS provides researchers with an easily browsable dataset of linked records and data scientists with annotated data, helping to answer research questions. Initially, and throughout, we applied basic practices to address needs involved in syntax, encoding, linking, organization, integration, access, and distribution of the data for individuals involved in CORD-19-related research. However, the CORD-19 dataset contains papers that are not peer-reviewed, raising questions about verifying the quality of their content. To help users gain trust in the system built and the quality of its contents, additional quality metadata should be added. Going forward, we believe it will be beneficial to maximize user insight into our dataset changes by providing more in-depth information about the processing tasks that were performed on the original CORD-19 dataset.

Even as our systems grew and as we learned much in our collaborations, the real take-aways, in our minds, are that curation is about focusing on the basic pre-requisites necessary for drawing valid inferences no matter the level of scale. Thus, as we move foward, our ability to move toward that ideal would involve encouraging the use, growth, and interconnection of systems, resources, and efforts – in particular, in the development of CORD19-specific CDCS-communities - that could move users closer to the primary sources of data, while still encouraging relevant and rich interlinking of data as appropriate.

8.2. Extracting Semantics from Syntax

Exploring the use of R&R for COVID-19 has provided us with new insights into domain-specific linguistic data representation that has opened the door to new domains. The NIST Information Technology Laboratory (ITL) is now working on several projects that incorporate linguistic data representations and have benefited from our experiences with CORD-19.

One of the challenges we faced while working with CORD-19 was evaluation. When extending tools to new domains, we would like to know that the data representation is accurate and useful. With CORD-19, we primarily used qualitative evaluations based on the system's ability to help users answer questions and find documents. As we move to new domains, we plan to develop new quantitative methods for evaluating key phrase extraction in new domains.

Our experiences using different levels of representation continues to influence our research as well. In some cases, our work depends on taking advantage of linguistic and textual structure in documents. In other cases, we provide streamlined user interfaces with minimal learning curves.

8.3. Data Management, Data Operations, and cv-py

Data management at NIST's Engineering Laboratory (EL) has previously been a source of problems: namely, the ongoing use of shared drives to host data files poses issues. From the stand point of researchers, having to share CSV and JSON files through either a shared drive or through instant messaging systems can be tedious, impractical, and lead to confusion as to what versions of data have been used, and where these versions are located.

From a broader scientific perspective, this confusion can pose issues of transparency, reliability and reproducibility. The lack of trace left by modifications and processing of the data makes it so research is at the risk of being "one and done", where the data used for a specific research project is immediately left to gather dust after the project has been completed.

At EL, the main impact of the cv-py project has thus been the way that data are treated within our teams: DVC (Data Version Control) [65] is now used routinely by the EL Information Testing and Modeling Group as part of a broader shift towards DataOps. This new way of treating data should enable better reproducibility and transparency of the analyses, and the models and insights the team generates. Additionally, researchers are able to write repeatable pipelines through DVC that can pre-process and analyze data, giving re-users a baseline to start. This feature was not used in cv-py, but is part of the broader use of tools to implement DataOps at NIST. On a broader institutional level, the set-up necessary to get this baseline data management system up and running allowed stakeholders including developers, researchers, and IT system administrators to discuss their needs and limitations more clearly. This first pass at having scalable data operations identified potential issues that would have prevented wider adoption from researchers. This allows a baseline analysis to be done through this automated pipeline, giving end-users a way to start gathering insight.

9. Toward a General System Architecture

In the previous sections, we laid out the various efforts that were brought together to address the needs of the COVID-19 community. In this section, we speculate on a future architecture based on the above experience. The main lessons of the experience are that we need an architecture that combines:

1. the ability to curate data and meta-data,
2. indexing at a multiple levels of granularity (sentences and paragraphs),
3. a variety of language processing tools such as language models, syntactic analysis, and named entity recognition, and
4. support for tool composition at the system integration level.

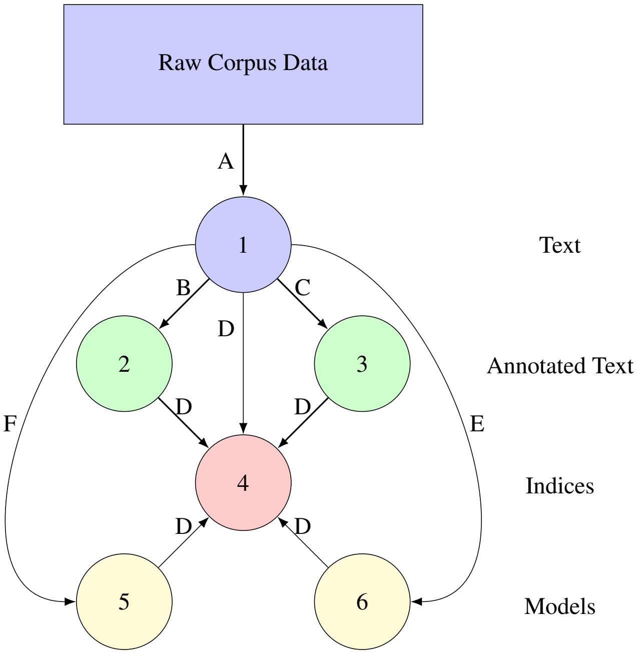
In order to provide research and practitioner communities with the computational functionality identified in Section 4.1, we propose a modular system architecture that encapsulates multiple workflows and handles many use-cases. Some of the potential data transformations and processing pipelines for such an architecture are described in Figure 1 below. Each node in the diagram represents a particular type of data that may be instantiated in many different ways by following different paths from the raw corpus data. The edges represent possible data transformations converting one type of data into another. This is not meant to be exhaustive; many other nodes and edges are possible, and each edge represents many possible implementations, which can be added to the system in a modular fashion.

The labeled edges of the diagram are explained in more detail below. There are four main layers of data representation in the diagram. The first is the raw corpus data, which is provided as a reference and the primary data source. The next is a semi-structured textual representation, which provides access to the text and metadata of the corpus in different formats. The annotated text layer builds on top of that by providing additional information about properties of the text, usually made explicit by human annotation or automatic classification using natural language processing. Indices provide a means of searching and organizing data based on different properties. Finally, the models layer provides representation of abstract properties of the corpus as a whole.

A (Preprocessing). Preprocessing converts the raw corpus data into one or more standardized formats (node 1). This may include JSON or XML representations that are designed to be easier to process, or pre-processed versions of the data that reduce noise or provide additional metadata. These formats may also be converted from one to the other, resulting in possible links between node 1 and itself. The preprocessing stage may also include an automatic filter or manual curation to ensure the quality and relevance of the data.

B (Labeling/Classification). Labeling and classification provides additional information that describes spans of text or entire documents. This includes common annotations such as named enti-

Fig. 1. State diagram of corpus data transformations and processing



1. Semi-structured Text with Metadata
2. Labeled language data
3. Structured language data
4. Indices
5. Language Models
6. Topic Models

ties, part-of-speech tags, and lemmas or stems. Labeled data makes implicit details of the original text explicit, allowing it to be used directly for other use-cases.

C (Structured Prediction). Structured prediction provides additional information that describes structural relationships between text or documents. For natural language, this typically involves syntactic parsing or discourse analysis. Like labeled data, structured data makes implicit details explicit and can support downstream tasks.

D (Indexing). There are many ways to index data, and many different kinds of information can be indexed, as shown by the different edges that all bear the same label. Indexing provides a relationship between data and its source, allowing efficient search. By indexing different kinds of data, not just the raw text, this architecture supports operations at different levels of representation. For example, it is possible to search not only distinct phrases, but their abstract linguistic features or broader patterns as well.

E (Topic Modeling). Topic modeling algorithms represents abstract topics over the entire corpus. Topic models can be used in many different ways, including clustering or classifying documents.

F (Language Modeling). Language models provide an abstract representation of the statistical properties of the language in a corpus. Recently, language models have been applied to many different areas, including question answering and natural language inference.

Many of the tools that have already been built for the CORD-19 dataset can be described using this general system architecture. For example, the NIST COVID-19 Data Repository is the result of following a path from 1 to 2 to 4 to 7: the raw data are first converted into a standardized format (node 1), which is then processed to produce several useful annotations such as named-entities and keywords (node 2). These are then used to create indices for a search engine (node 4), which is exposed under the web-based user interface at https://covid19-data.nist.gov/. Similarly, the NIST Scientific Indexing Resource for CORD-19 can be described as a path from 1 to 3 to 4 to 7: the raw data are once again standardized (node 1), but rather than being given span-based annotations, they are given structural representations, which are normalized by the root- and rule-based system (node 3). These normalized structural representations can be used to create normalized indices (node 4), which are once again exposed under a web-based interface (https://randr19.nist.gov/). Other systems within the broader community can also be described as paths; SPIKE-CORD, for example, may be described as a path from 1 to 3 to 7, albeit using a different arrow from 1 to 3 than the NIST Scientific Indexing Resource, since SPIKE-CORD uses its own underlying linguistic representations rather than R&R.

The place of subsystems such as cv-py under this architecture is somewhat different. In its current form, cv-py could be conceived of as a user interface into several paths of the architecture. For example, it provides programmatic access to a representation of the corpus' raw text (node 1), as well as programmatic access to a number of natural language processing tools on top of the corpus (nodes 2 and 3). However, a loftier goal for tools such as this would be to provide an interface into the entire architecture, by allowing the user to specify which path they would like to follow, and providing a systematic way to process the original text to arrive at any point in the architecture following any path. This would provide a single tool that would allow a user to access the data that they need for their various use-cases.

Another important feature of this architecture is its modularity. Because it is possible to add new nodes and arrows to Figure 1, above, new developments in AI and NLP can be incorporated as new paths to create new representations of the data. Tools made to access this architecture will need to function not only as software, but as curation systems that can handle the introduction of new elements.

10. Conclusions and Future Work

COVID-19 brings with it an urgency to organize and access fast-moving information in the research community. We explored these issues by bringing together multiple efforts at NIST to create a prototype information infrastructure to investigate the facilitation of multidisciplinary research. Our work provided us with insights and the ability to articulate needs for a future community-centered information infrastructures for scientific and technological research.

In this paper, we also describe a general information architecture for infrastructure that handles the curation, querying, and exploration of COVID-19 data. The system that we have described is not limited to COVID-19, but can be applied to other domains. However, COVID-19 provides an interesting special case, in which the data were rapidly evolving at the time of this writing and a large number of data sets were being created, updated, and integrated into more advanced systems. The introduction of such a flexible but cohesive framework could help in the future by creating community-driven platforms for handling different levels of representation of the data.

Our approach is generic and modular, making it possible to extend and swap capabilities as needed. We have described how existing techniques have already been implemented, and how new techniques could be incorporated. For example, R&R is an emerging technique developed at NIST for representing the linguistic structure of text in a normalized form. This was incorporated into the infrastructure as a way of representing data, which can then be indexed or visualized using other techniques.

While this architecture is designed to be constantly evolving, it requires certain foundational components to be put in place initially, including tools for curation and the introduction of new functionality. While some of these components exist, (e.g., the Configurable Data Curation System (CDCS)), there are several under development and others that are the subject of research. Future research is not limited to addition and composition of new linguistic models, AI and NLP tools, but also concern how a user interface can be responsive to changing tools and their composition for different needs and roles. Perhaps most important for future research is how to evaluate platform performance, that is identifying and constructing metrics for measuring performance. We expect that the principles of this architectural framework can be followed and expanded upon in the future, in order to create more accessible, extensible, and modular information system architectures for science and technology communities.

Acknowledgments

The authors gratefully acknowledge Ravikanth Vishnubhotla, Asian Institute of Gastroenterology, for his feedback early in the development of this document.

References

[1] Materials genome initiative, https://www.mgi.gov/. Accessed: 2022-1-3.

[2] (2020) Call to action to the tech community on new machine readable COVID-19 dataset, https://trumpwhitehouse.archives.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/. Accessed: 2022-1-18.

[3] Lu Wang L, Lo K, Chandrasekhar Y, Reas R, Yang J, Eide D, Funk K, Kinney R, Liu Z, Merrill W, Mooney P, Murdick D, Rishi D, Sheehan J, Shen Z, Stilson B, Wade AD, Wang K, Wilhelm C, Xie B, Raymond D, Weld DS, Etzioni O, Kohlmeier S (2020) CORD-19: The covid-19 open research dataset. *ArXiv* Accessed: 2021-2-22 https://doi.org/10.48550/arXiv.2004.10706.

[4] Colavizza G, Costas R, Traag VA, van Eck NJ, van Leeuwen T, Waltman L (2021) A scientometric overview of CORD-19. *PLoS One* 16(1):e0244839. Accessed: 2021-2-22 https://doi.org/10.1371/journal.pone.0244839.

[5] Kaggle: Your machine learning and data science community, https://www.kaggle.com/. Accessed: 2022-1-18.

[6] Vreeman A (2020) Call to action to the tech community on new machine readable COVID-19 dataset - center for security and emerging technology, https://cset.georgetown.edu/article/call-to-action-to-the-tech-community-on-new-machine-readable-covid-19-dataset/. Accessed: 2021-10-19.

[7] Wang LL, Lo K (2021) Text mining approaches for dealing with the rapidly expanding literature on COVID-19. *Brief Bioinform* 22(2):781–799. Accessed: 2022-5-11 https://doi.org/10.1093/bib/bbaa296.

[8] Bouras S (2020) How import data, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/153046. Accessed: 2022-5-11.

[9] “jdj8af” (2020) Package for easily loading data, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/136781. Accessed: 2022-5-11.

[10] Houston L (2020) Convert JSON to TXT and search multiple keywords, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/139264. Accessed: 2022-5-11.

[11] “Feraldo L” (2020) JSON to CSV?, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/139297. Accessed: 2022-5-11.

[12] “popoye_1996” (2020) Converting the data into word embeddings using infersent and glove pretrained model, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/138050. Accessed: 2022-5-11.

[13] Bonde Y (2020) Incorrect JSON data!, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/136935. Accessed: 2022-5-18.

[14] Burke M (2020) SHA values in metadata.csv aren’t unique, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/137921. Accessed: 2022-5-18.

[15] Joslyn C (2020) Value of metadata.csv vs. JSON for bibliometrics?, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/139215. Accessed: 2022-6-17.

[16] “Julian” (2020) Tables referenced in the json files?, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/138697. Accessed: 2022-6-17.

[17] Mooney P (2020) Questions about table formats and column definitions, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/152615. Accessed: 2022-6-17.

[18] Reid S (2020) epi notes on under-represented topics in the data, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/137452. Accessed: 2022-6-16.

[19] da Silva LM (2020) Divergences between bib_entries, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/145907. Accessed: 2022-6-16.

[20] Sinha S (2020) Use NLP to answer key questions from the scientific literature, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/155077. Accessed: 2022-6-17.

[21] “pratik@semandex” (2020) Publish dates are not available in JSON files, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/139634. Accessed: 2022-5-18.

[22] “lenaschmidt0493” (2020) publish_time in metadata.csv bug: publication dates december 2020?, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/141752. Accessed: 2022-5-18.

[23] Dubey A (2020) Data size is too large, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/143503. Accessed: 2022-5-12.

[24] “DAVIDAK” (2020) 5143 of 128025 files are corrupted, https://www.kaggle.com/code/davidak/5143-of-128025-files-are-corrupted/notebook. Accessed: 2022-6-17.

[25] Harmsen J (2020) Duplicate body_text.text entries within articles?, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/137151. Accessed: 2022-5-18.

[26] Miller K (2020) An AI challenge with AI2, CZI, MSR, georgetown, NIH & the white house,

https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/138872. Accessed: 2022-6-17.

[27] Cuix Y (2021) Feel confused about the data, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/232948. Accessed: 2022-5-12.

[28] "Kacey" (2020) Parsing errors in the JSON data?, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/137995. Accessed: 2022-5-18.

[29] "Simon" (2020) How data originally parsed to json, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/138237. Accessed: 2022-6-16.

[30] "ODI6s" (2020) How do I link the JSON files to the metadata?, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/138849. Accessed: 2022-6-16.

[31] "heibufan" (2020) Parsed pdf tool used by CORD-19 project, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/144434. Accessed: 2022-6-16.

[32] Boumedine MS (2020) Reproduce cord-19 from scratch, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/168254. Accessed: 2022-5-12.

[33] Dannelly Z (2020) Metadata enrichment: Fixing journal abbreviations and references, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/140678. Accessed: 2022-5-12.

[34] Marble A (2020) Dataset evaluation, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/141577. Accessed: 2022-6-17.

[35] Allen Institute For AI (2020) Constellation: Helping researchers find relevant reads using text mining, ML, and visualization, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/146021. Accessed: 2022-6-17.

[36] Nafoosi M (2020) g6gtech, inc. offers artificial intelligence resource to aid COVID-19 research and the CORD-19 challenge, https://www.kaggle.com/discussions/general/139211. Accessed: 2022-6-17.

[37] Trainar G (2020) Free cloud-based GPU and storage, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/137655. Accessed: 2022-6-17.

[38] Reid S (2020) Epidemiologist available to hand-code records for training sets, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/137027. Accessed: 2022-6-17.

[39] Besomi J (2020) Universal toolkit for fast development, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/148389. Accessed: 2022-6-17.

[40] Covid-19 kaggle community contributions, https://www.kaggle.com/covid-19-contributions. Accessed: 2022-6-22.

[41] Golabek P (2020) Challenges come and go, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/141672. Accessed: 2022-6-17.

[42] Guéret C (2020) Sub-notebooks?, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/143808. Accessed: 2022-6-22.

[43] Marble A (2020) Dataset evaluation, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/141577. Accessed: 2022-5-12.

[44] Mezzetti D (2020) Kaggle notebook job framework, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/156398. Accessed: 2022-6-17.

[45] Mikhalev A (2020) Metrics or method to measure quality of the tokenisation - spacy/scispacy,

https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/144012. Accessed: 2022-6-17.

[46] Miller K (2020) Self-Organizing scientific research understanding system, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/147427. Accessed: 2022-6-22.

[47] "dublancas"' (2020) Open source base pipeline, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/138377. Accessed: 2022-6-17.

[48] madhucharis (2020) Topic author towards understanding "Learnable/Recognizable" patterns in dataset, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/138398. Accessed: 2022-6-22.

[49] bluebalam (2020) User stories and validation with MDs and health researchers?, https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/143339. Accessed: 2022-6-22.

[50] Lo K, Wang LL, Neumann M, Kinney R, Weld D (2020) S2ORC: The semantic scholar open research corpus. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Online), pp 4969–4983.

[51] Allen AI (2022) cord19: Get started with CORD-19, https://github.com/allenai/cord19. Accessed: 2022-6-22.

[52] Cord-19 covid-19 open research dataset, https://semanticscholar.org/cord19. Accessed: 2022-4-25.

[53] Allen Institute For AI COVID-19 open research dataset challenge (CORD-19).

[54] Wang Ll (2022) Sunsetting CORD-19, https://blog.allenai.org/sunsetting-cord-19-239fb2f9ff4a. Accessed: 2022-6-22.

[55] Tiwari P, Kaur H (2020) The flood of COVID-19 publications: a word of caution. *SN Compr Clin Med* :1–3.

[56] Academic papers about covid-19 primer from mar 24, 2022 til jun 22, 2022, https://covid19-science.primer.ai/5322aaf6-974b-46da-92e0-ac0a0f754a21. Accessed: 2022-6-22.

[57] (2016) Online summarizing tool, https://www.scholarcy.com/. Accessed: 2022-6-22.

[58] Donnelly M, Jones S (2009) Dcc data management plan content checklist.

[59] Lyon L (2007) Dealing with data: Roles, rights, responsibilities and relationships (University of Bath),

[60] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, et al. (2016) The fair guiding principles for scientific data management and stewardship. *Scientific data* 3(1):1–9.

[61] Nitecki DA, Alter A (2021) Leading fair adoption across the institution: A collaboration between an academic library and a technology provider. *Data Science Journal* 20(1).

[62] Continuum IO Intake. Available at https://github.com/intake/intake.

[63] DoltHub, Inc Dolt. Available at https://www.dolthub.com/.

[64] Quilt Data, Inc Quilt. Available at https://quiltdata.com.

[65] iterativeai Data version control. Available at https://dvc.org.

[66] Ereth J (2018) Dataops-towards a definition. *LWDA* 2191:104–112. Available at https://api.semanticscholar.org/CorpusID:52164356.

[67] Google COVID-19 research explorer, https://covid19-research-explorer.appspot.com/. Accessed: 2021-2-22.

[68] Bhatia P, Arumae K, Pourdamghani N, others (2020) AWS CORD19-search: a scientific lit-

erature search engine for COVID-19. *arXiv* https://doi.org/10.48550/arXiv.2007.09186.

[69] Zhang E, Gupta N, Nogueira R, Cho K, Lin J (2020) Rapidly deploying a neural search engine for the COVID-19 open research dataset: Preliminary thoughts and lessons learned. *arXiv* https://doi.org/10.48550/arXiv.2004.05125.

[70] Verspoor K, Šuster S, Otmakhova Y, Mendis S, Zhai Z, Fang B, Lau JH, Baldwin T, Yepes AJ, Martinez D (2020) COVID-SEE: Scientific evidence explorer for COVID-19 related research. *arXiv* https://doi.org/10.48550/arXiv.2008.07880.

[71] Esteva A, Kale A, Paulus R, Hashimoto K, Yin W, Radev D, Socher R (2020) CO-Search: COVID-19 information retrieval with semantic search, question answering, and abstractive summarization. *arXiv* https://doi.org/10.48550/arXiv.2006.09595.

[72] Taub-Tabib H, Shlain M, Sadde S, Lahav D, Eyal M, Cohen Y, Goldberg Y (2020) Interactive extractive search over biomedical corpora. *Proceedings of the BioNLP 2020 workshop* (Association for Computational Linguistics), pp 28–37.

[73] Voorhees E, Alam T, Bedrick S, Demner-Fushman D, Hersh WR, Lo K, Roberts K, Soboroff I, Wang LL (2020) TREC-COVID: Constructing a pandemic information retrieval test collection. *arXiv* https://doi.org/10.48550/arXiv.2005.04474.

[74] Lima LC, Hansen C, Hansen C, Wang D, Maistro M, Larsen B, Simonsen JG, Lioma C (2020) Denmark's participation in the search engine TREC COVID-19 challenge: Lessons learned about searching for precise biomedical scientific information on COVID-19. *arXiv* https://doi.org/10.48550/arXiv.2011.12684.

[75] Heaton CT, Mitra P (2020) Repurposing TREC-COVID annotations to answer the key questions of CORD-19. *arXiv* https://doi.org/10.48550/arXiv.2008.12353.

[76] Tomokiyo T, Hurst M (2003) A language model approach to keyphrase extraction.

[77] Chuang J, Manning CD, Heer J (2012) Without the clutter of unimportant words: Descriptive keyphrases for text visualization. *ACM Transactions on Computer-Human Interaction* 19(3).

[78] Mihalcea R, Tarau P (2004) TextRank: Bringing order into text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Barcelona, Spain), pp 404–411. Available at https://aclanthology.org/W04-3252.

[79] Wadden D, Wennberg U, Luan Y, Hajishirzi H (2019) Entity, relation, and event extraction with contextualized span representations. *ArXiv* abs/1909.03546.

[80] Tyagin I, Kulshrestha A, Sybrandt J, Matta K, Shtutman M, Safro I (2021) Accelerating COVID-19 research with graph mining and transformer-based learning. *bioRxiv* https://doi.org/10.48550/arXiv.2102.07631.

[81] Cernile G, Heritage T, Sebire NJ, Gordon B, Schwering T, Kazemlou S, Borecki Y (2021) Network graph representation of COVID-19 scientific publications to aid knowledge discovery. *BMJ Health Care Inform* 28(1).

[82] Reddy RG, Iyer B, Sultan MA, Zhang R, Sil A, Castelli V, Florian R, Roukos S (2020) End-to-End QA on COVID-19: Domain adaptation with synthetic training. *arXiv* https://doi.org/10.48550/arXiv.2012.01414.

[83] Bhat TN, Subrahmanian E, Kattner U, Elliott J, Campbell C, Monarch I, Collard J (2018) Generating domain ontologies using root- and rule-based terms. *Journal of the Washington Academy of Science* .

[84] Wen A, Fu S, Moon S, El Wazir M, Rosenbaum A, Kaggal VC, Liu S, Sohn S, Liu H, Fan J (2019) Desiderata for delivering nlp to accelerate healthcare ai advancement and a mayo

clinic nlp-as-a-service implementation. *npj Digital Medicine* 2(1):130. https://doi.org/10.1038/s41746-019-0208-8. Available at https://doi.org/10.1038/s41746-019-0208-8

[85] Configurable data curation system, https://cdcs.nist.gov/. Accessed: 2022-1-3.

[86] Dima A, Bhaskarla S, Becker C, Brady M, Campbell C, Dessauw P, Hanisch R, Kattner U, Kroenlein K, Newrock M, Peskin A, Plante R, Li SY, Rigodiat PF, Amaral GS, Trautt Z, Schmitt X, Warren J, Youssef S (2016) Informatics infrastructure for the materials genome initiative. *JOM* 68(8).

[87] Smart manufacturing systems (SMS) test bed, https://www.nist.gov/laboratories/tools-instruments/smart-manufacturing-systems-sms-test-bed. Accessed: 2021-11-18.

[88] Helu M, Hedberg T (2020) Recommendations for collecting, curating, and re-using manufacturing data. https://doi.org/https://doi.org/10.6028/NIST.AMS.300-11.

[89] About large files on github, https://docs.github.com/en/repositories/working-with-files/managing-large-files/about-large-files-on-github. Accessed: 2022-3-9.

[90] About git large file storage, https://docs.github.com/en/repositories/working-with-files/managing-large-files/about-git-large-file-storage. Accessed: 2022-3-9.

[91] Comparison with related technologies, https://dvc.org/doc/user-guide/related-technologies. Accessed: 2022-3-9.

[92] So you want git for data?, https://www.dolthub.com/blog/2020-03-06-so-you-want-git-for-data/. Accessed: 2022-3-9.

[93] Ruiz-Real JL, Nievas-Soriano BJ, Uribe-Toril J (2020) Has covid-19 gone viral? an overview of research by subject area. *Health Education & Behavior* 47(6):861–869. https://doi.org/10.1177/1090198120958368. PMID: 32886013 Available at https://doi.org/10.1177/1090198120958368

[94] Lee J, Yi SS, Jeong M, Sung M, Yoon W, Choi Y, Ko M, Kang J (2020) Answering questions on COVID-19 in real-time. *CoRR* abs/2006.15830. 2006.15830 Available at https://arxiv.org/abs/2006.15830.

Appendix A. CORD-19 Infrastructure Uses Cases

We provide here a listing of the set of high-level use-cases exercised for the CORD-19 data set. Each use-case highlights how it fits into the general architecture described in this paper.

Appendix A.1. Curation of Published Data

ID: CURATE-1

Title: Curation of published data

Description: A data publisher makes a COVID-19 related data set publicly available. A curator discovers this data set, reformats the data and enhances the metadata and then makes the curated data set available via the publicly accessible repository.

Actor: Publisher, Curator

Architecture Path: Raw Data → Semi-structured text with metadata

Appendix A.2. Infrastructure for Publishing: CORD-19 CDCS

Infrastructure Components: CORD-19 CDCS

ID: PUBLISH-1

Title: Register published data

Description: A data publisher wishes to make their publicly available COVID-19 data set known to the community. The publisher creates a record in the COVID-19 registry that describes the data set and gives its URL.

Actor: Publisher

Architecture Path: Semi-structured text with metadata → raw data

Appendix A.3. Infrastructure for Search: COVID-19 Registry

Infrastructure Components: COVID-19 Registry

ID: SEARCH-1

Title: Researcher seeks and downloads data set from the repository

Description: A researcher seeking a COVID-19 related data set goes to the repository and searches for the appropriate one. The researcher then selects and downloads a data set.

Actor: Researcher

Architecture Path: Raw Data

Appendix A.4. Components for Search: CORD-19 CDCS

Infrastructure Components: CORD-19 CDCS

ID: SEARCH-2

Title: Researcher searches for a CORD-19-related data set.

Description: A researcher is seeking a COVID-19 related data set goes to the COVID-19 registry and searches for the appropriate one. The researcher then selects a data set and browses the metadata. If this data set is of interest, the researcher then follows the registered URL to the published location and downloads the data set.

Actor: Researcher

Architecture Path: Semi-structure text with metadata → Annotated spans → Indexes → User Interfaces

Appendix A.5. Components for Search: COVID-19 Registry

Infrastructure Components: COVID-19 Registry

ID: SEARCH-3

Title: Researcher searches CORD-19 for individual research papers

Description: A researcher is seeking relevant papers from the CORD-19 data set to answer questions about COVID-19. The researcher identifies keywords for each question and then researcher performs a keyword search. To find individual papers to read, the researcher, scans search results and clicks on an item of interest. The researcher reads the item summary and if the paper is relevant, clicks on the publication link.

Actors: Researcher

Architecture Path: Semi-structure text with metadata → Annotated spans → Indexes → User Interfaces **or** Semi-structured text with metadata → Indexes → User Interfaces

Appendix A.6. Components for Search: CORD-19 CDCS, R&R

Infrastructure Components: CORD-19 CDCS, R&R

ID: SEARCH-4

Title: Researcher searches CORD-19 as part of a systematic literature review

Description: A researcher is seeking all papers in the CORD-19 data set as part of a systematic literature review related to COVID-19 whose protocol has defined sets of keywords as part of the inclusion criteria. For each set of keywords, the researcher selects all of the papers returned by the search and downloads them.

Actors: Researcher

Architecture Path: Semi-structured text with metadata → Linguistic annotations → Indexes → User Interfaces

Appendix A.7. Components for Search: CORD-19 CDCS

Infrastructure Components: CORD-19 CDCS

ID: SEARCH-5

Title: Researcher seeking to answer questions with the most relevant papers

Description: A researcher is seeking to answer a specific COVID-19 question using the CORD-19 data set by iteratively searching for the most relevant papers. The researcher begins with an initial set of keywords which is refined during the search process. The search starts with the most relevant top-level paper. After reviewing the paper, and taking note of relevant information, the researcher reviews related papers suggested by the search engine and updates the notes and refines the set of keywords for later searches. The researcher then continues to the next most relevant paper from the initial search and repeats the process until all of the relevant papers have been examined. A new search is done with the refined keywords. This process continues iteratively until no new information is found.

Actors: Researcher

Architecture Path: Semi-structure text with metadata → Annotated spans → Indexes → User Interfaces

Appendix A.8. Components for Taxonomy: CORD-19 CDCS, R&R

Infrastructure Components: CORD-19 CDCS, R&R

ID: TAXONOMY-1

Title: Researcher seeking concepts and terminology related to a key phrase

Description: A researcher is seeking to find concepts and terms related to a particular key phrase, in order to expand an initial set of keywords (see SEARCH-5) or to determine which concepts have been studied in the COVID-19 literature. The researcher begins with a single key phrase, which is input into the R&R terminology explorer. The researcher may then select additional related key phrases to explore the taxonomy.

Actors: Researcher

Architecture Path: Semi-structured text with metadata → Linguistic annotations → User Interfaces

Appendix A.9. Components for Temporal Queries: R&R

Infrastructure Components: R&R

ID: TEMPORAL-1

Title: Researcher seeks to answer meta-scientific questions about the historical usage of key phrases

Description: A researcher is seeking to answer a meta-scientific question about how certain key phrases have been used over time in the CORD-19 literature, which includes a broad set of literature on different coronaviruses. The researcher enters key phrases into the R&R system and then enters the frequency graph interface. The researcher may elect to see absolute or relative frequencies of different terms over time, in order to observe how the use of each term has changed over a period of time.

Actors: Researcher

Architecture Path: Semi-structured text with metadata → Linguistic annotations → Indexes → User Interfaces

Appendix A.10. Components for Topic Modeling: R&R

Infrastructure Components: R&R

ID: TOPIC-MODELING-1

Title: Student or Researcher seeking to understand the key topics in all or portions of the COVID-19 literature

Description: A student or researcher is seeking to find significant topics in the COVID-19 literature. Topic modeling applications transform a document term matrix into a topic term matrix, reducing the dimensions of the matrix by three for four orders of magnitude. Each topic consists of single and multi-word terms of descending relevance to the topic. If the student or researcher has a particular data set in mind, they search for it using Search-1 or Search-2 techniques. If they do not, they browse the available data sets in either or both the repository or COVID-19 registry to find ones that seem interesting to them. They can then apply the topic modeling techniques to those data sets. If they are unsure which would be of interest, not wanting to miss something, they could apply the topic modeling techniques to the whole repository. Students or researchers can specify how many topics they want to investigate or allow the topic modeling application to find an optical one. Depending on the topics identified being either too general or too specific, or the number of topics brought back being unmanageable, the student or researcher can change the number of topics as they see fit. Each topic run is saved so that they can compare the results of the different topic modeling runs. When they find topics that interest them, they can bring back documents in the order of relevance to one or more topics. They can also use the terms (single word or multi-word) to perform Searches 3-5.

Actors: Student or Researcher

Architecture Path: Semi-structured text with metadata → Linguistic annotations → Topic Models → Indexes → User Interfaces

Appendix A.11. Components for Temporal Queries: R&R

Infrastructure Components: R&R

ID: TEMPORAL-2

Title: Student or researcher seeks to answer meta-scientific questions about the historical evolution of topics that interest them

Description: A researcher is seeking to answer a meta-scientific question such as how certain key topics have evolved over time and what the trends are in the CORD-19 literature, as represented in the repository. The researcher enters topics of interest into the R&R system and uses a graphic interface to view graphs showing when a topic, how it evolves, morphs into others or disappears. This makes use of TOPIC-MODELING-1.

Actors: Student and Researcher

Architecture Path: Semi-structured text with metadata → Linguistic annotations → Topic Models → Indexes → User Interfaces

Appendix A.12. Components for Semantic Maps: R&R

Infrastructure Components: R&R

ID: KG-1

Title: Student or Researcher is guided by terminological maps representing the concepts in a corpus to help understand the concepts and topics that are dispersed and expressed in documents of the corpus

Description: A student or researcher is seeking to understand the conceptual landscape of a corpus like the COVID-19 repository, or portions of it, and to do so without having to read all the documents in such collections. Knowledge graphs represent the basic objects and relations expressed in a corpus. The graphs consist of labeled links between labeled objects. The knowledge graph interface allows the network thus represented to be navigated and the sections of the documents containing these objects and relationships to be read. While the multi-word terms gathered together in topics by Topic Modeling-1 techniques does explicitly represent relationships among objects, they do not provide navigable visualizations of these relationships. Moreover, these navigation facilities can be used to find represented relationships among the topics identified. While topic modeling techniques attempt to identify relations among topics graphically, these relationships are not labeled. The knowledge graph interface in conjunction with topics identified by topic modeling makes it possible for human users to label and understand the relationships among topics. So, knowledge graphs and topic modeling techniques mutually support each other and together support Search-1-5.

Actors: Student or Researcher

Infrastructure Components: R&R

Architecture Path: Semi-structured text with metadata → Linguistic annotations → Topic Models → Knowledge Graphs → User Interfaces. Knowledge Graphs is a new node not shown in Figure 1 at the modeling layer, which provides information about the relationships between concepts in a corpus.

Appendix B. Queries and Results

In Section 1, we pose three questions that CORD-19 should help researchers answer. To demonstrate how our work can help answer these questions, we convert them into queries which can be run against the CORD-19 dataset using the R&R system. The R&R system can convert questions into queries by analyzing the syntactic structure of the question and extracting key phrases to search for in the dataset. Below is given a description of how each of the questions posed in Section 1 is converted into a query and the results of this query as of the January 3, 2022 release of CORD-19. Currently, these queries cannot be performed as-is on the public R&R website, but must be performed directly against the database.

Appendix B.1. Question 1

Question: What mutations have been introduced into a coronavirus that enable amino acids to affect virus assembly?

Query: virus mutations; amino acids; virus assembly

Results

- https://arxiv.org/pdf/2006.03915.pdf
- https://doi.org/10.1016/j.medidd.2021.100099

- https://doi.org/10.1016/b978-0-12-384939-7.10007-9
- https://doi.org/10.1007/978-1-4899-7448-8_10
- https://doi.org/10.1038/nri2623
- https://doi.org/10.1007/s11756-021-00866-y
- https://doi.org/10.3390/v13061120
- https://doi.org/10.1002/jmv.26597
- https://doi.org/10.1007/978-3-030-33946-3_1
- https://doi.org/10.1007/s11262-016-1296-z
- https://doi.org/10.3389/fvets.2021.744276
- https://doi.org/10.1055/s-0040-1715790
- https://doi.org/10.1016/b978-0-12-373741-0.50004-0
- https://doi.org/10.3390/s20154289
- https://doi.org/10.1007/978-3-319-47426-7_11
- https://doi.org/10.1007/s43538-021-00058-x
- https://doi.org/10.1016/b978-0-12-816422-8.00011-8

Appendix B.2. Question 2

Question: Are there any small molecule drugs that have been specifically shown to affect virus assembly inside a cell?

Query: small molecule drugs; virus assembly

Results

- https://doi.org/10.1152/ajpendo.00298.2020
- https://doi.org/10.1371/journal.ppat.1008542
- https://doi.org/10.1016/bs.apcsb.2015.12.003
- https://doi.org/10.1007/978-90-481-8622-8_3
- https://doi.org/10.3390/pharmaceutics12100945
- https://doi.org/10.1101/2020.12.01.406637
- https://doi.org/10.3389/fmed.2020.00444
- https://doi.org/10.1038/s41422-021-00519-4
- https://doi.org/10.3390/v12111289
- https://doi.org/10.1038/s41594-020-00536-8

- https://doi.org/10.1016/j.cimid.2007.05.009
- https://doi.org/10.3390/ijms22041737
- https://arxiv.org/pdf/2009.12817.pdf

Appendix B.3. Question 3

Question: How stable is virus RNA in sewage?

Query: RNA stability; sewage

- https://doi.org/10.1016/j.scitotenv.2021.148834
- https://doi.org/10.1101/2021.05.06.21256753
- https://doi.org/10.1016/s2666-5247(21)00059-8
- https://doi.org/10.1016/b978-0-12-407698-3.00002-8
- https://doi.org/10.1016/j.watres.2021.117090
- https://doi.org/10.3390/microorganisms9061149
- https://doi.org/10.1002/jmv.26170
- https://doi.org/10.1016/j.scitotenv.2021.149562
- https://doi.org/10.1016/j.ijheh.2020.113621
- https://doi.org/10.1101/2020.09.01.20185280
- https://doi.org/10.1101/2020.12.19.20248508
- https://doi.org/10.1016/j.watres.2021.117183
- https://doi.org/10.1101/2020.12.01.20242131
- https://doi.org/10.1021/acsestwater.0c00216
- https://doi.org/10.1021/acsenvironau.1c00015
- https://doi.org/10.1016/j.watres.2021.117252