# Outbreak.info Research Library: A standardized, searchable platform to discover and explore COVID-19 resources

Tsueng, Ginger; Mullen, Julia; Alkuzweny, Manar; Cano, Marco; Rush, Benjamin; Haag, Emily; Outbreak Curators; Latif, Alaa Abdel; Zhou, Xinghua; Qian, Zhongchao; Hufbauer Emory; Zeller, Mark; Andersen, Kristian G.; Wu, Chunlei; Su, Andrew I.; Gangavarapu, Karthik; Hughes, Laura D.

## Abstract

To combat the ongoing COVID-19 pandemic, scientists have been conducting research at breakneck speeds, producing over 52,000 peer reviewed articles within the first 12 months. In contrast, a little over 1,000 peer reviewed articles were published within the first 12 months of the SARS-CoV-1 pandemic starting in 2002. In addition to publications, there has also been an upsurge in clinical trials to develop vaccines and treatments, scientific protocols to study SARS-CoV-2, methodology for epidemiological modeling, and datasets spanning molecular studies to social science research. One of the largest challenges has been keeping track of the vast amounts of newly generated disparate data and research that exist in independent repositories. To address this issue, we developed outbreak.info, which provides a standardized, searchable interface of heterogeneous resources on COVID-19 and SARS-CoV-2. Unifying metadata from twelve repositories, we have assembled a collection of over 260,000 publications, clinical trials, datasets, protocols, and other resources as of March 2022. We used a rigorous schema to enforce a consistent format across different data sources and resource types, and linked related resources where possible. This enables users to quickly retrieve information across data repositories, regardless of resource type or repository location. Outbreak.info also combines the combined research library with spatiotemporal genomics data on SARS-CoV-2 variants. The web interface provides interactive visualizations and reports to explore the unified data and generate hypotheses. In addition to providing a web interface, we also publish the data we have assembled and standardized in a high performance public API and an R package. Finally, we discuss the challenges inherent in combining metadata from scattered and heterogeneous resources and provide recommendations to streamline this process to aid scientific research.

## Introduction

In early January 2020, SARS-CoV-2 was identified as the virus responsible for a series of pneumonia cases with unknown origin in Wuhan, China (World Health Organization, 2020). As the virus quickly spread all over the world, the global scientific community began to study the new

virus and disease, resulting in the rapid release of research outputs (such as publications, clinical trials, datasets, etc.) and resources (i.e. - research outputs, collections of research outputs, websites, portals and more). The concurrent and frequently uncoordinated generation and curation of resources by different types of resource generators (such as government agencies, NGOs, academic/medical/research institutes/associations, etc.) exacerbate four factors that make finding and using resources a challenge: volume, fragmentation, variety, and standardization (Figure 1). Only after addressing these four factors, can we consider addressing challenges in translating centralized, standardized resources into a user-friendly platform for resource discovery.

For example, the volume of peer-reviewed articles from a single resource (LitCovid) has grown from about 52,000 published within the first twelve months to over 240,000 as of January 2022. Since April 2020, over 1,000 different research outputs have been published on a weekly basis. The rapid proliferation of resources could be manageable if there were a centralized repository for finding them; however, there is no single centralized repository. In addition to research outputs like scientific literature, researchers, public health officials, media outlets, and concerned communities independently developed websites providing highly localized or specialized information on infection rates, prevention policies, and travel restrictions (Dong, Du, and Gardner, 2020)(Kaiser, 2020)(Noren, Neff, and Li, 2020) (Morris and citizen scientists, 2020) (James and citizen scientists, 2020) (Pogkas et al., 2020) resulting in a fragmented landscape of very different types of resources (Figure 1).

The volume and fragmentation issues were immediately obvious. Lacking alternate solutions for addressing these issues, individual and community efforts for curating these resources were created via shared Google spreadsheets to aid in discoverability (Joachimiak et al., 2020) (Skenderi et al., 2020) (Navarro and Capdarest-Arest, 2020). However, the sheets were not a scalable solution and usually lacked sufficient metadata for describing resources, with the exception of Navarro and Capdarest-Arest. Several projects have attempted to address the volume and fragmentation issues, but were most often focused on a single type of resource. For example, NIH's iSearch COVID-19 portfolio (NIH OPA, 2020) and the Kaggle COVID-19 Open Research Dataset Challenge (CORD-19) aggregate scholarly articles (Allen Institute For AI, 2020), but do not include clinical trials, datasets, or other types of resources.

Compounding search issues caused by the variety of resource types, there has been a long-standing lack of standardization even *within* a particular type of resource. Existing resource repositories which were able to pivot quickly and curate COVID-19 content from their collections utilized pre-existing metadata standards. For example, researchers involved in PubMed, which uses Medline citation standards, shifted quickly to create LitCovid which follows the same standard (Chen, Allot and Lu, 2020). Similarly, National Clinical Trials Registry has their own custom list of COVID-19 Clinical Trials which follows their own Protocol Registration and Results System (PRS) schema (ClinicalTrials.gov, 2018), but these conventions are not followed by the WHO International Clinical Trials Registry Platform. Zenodo (Fava et al., 2020) and Figshare (Hyndman, 2020), which both enable export to multiple open data formats including json schema/schema.org, do not completely agree on the marginality, cardinality, and selection of the

properties in profiles they use (European Organization For Nuclear Research, 2013)(Hahnel, 2018)(Schema.org, 2015).

Who contributes to resource proliferation during an outbreak?

| | Existing resources | Medical / research institute / association | Governmental Agencies / Trade associations / NGOs | News and Media Outlets | Citizen Scientists | Search engines and others |
|---|---|---|---|---|---|---|
| **Examples of resource generators** | PubMed, Figshare | JHU, Boston Children's Hospital | CDC, IATA, WHO, ACAPS, Dept. of Public Health | The New York Times, The Atlantic | Phildini, COVID-19 LST | google |
| **What kind of resources do they generate?** | Subsets of existing resources | Research Outputs, Websites, portals, tools | Research outputs, portals, lists, guides, policies, and more | Dashboards, Datasets and other Research outputs, websites, apps | Lists, collections, tools, websites, portals, dashboards | Dashboard-like, specialized search results |
| **What are some examples of resources they generate?** | LitCovid, Covid Figshare | JHU COVID cases and deaths data, VaccineFinder | WHO Reports, Local county dashboards, CDC, State Declarations | The Atlantic's Covid Tracker, New York Time's data | stayinghomeclub , Daily COVID-19 LST Reports | Google's covid19 search including maps with testing sites, etc. |
| **What is the volume and fragmentation of the resources they generate?** | Volume and fragmentation depend on resource | High collective volume but very fragmented | High collective volume, fragmentation dependent on data infrastructure | Varies—depends on syndication and shared infrastructure | Volume unknown due to high level of fragmentation | Low--single resources (search engines) aggregating metadata from other resources |
| **What is the level of standardization for the resource data or metadata?** | High within a resource, Low across different resources | Varies– can be high within a resource type or conform to repository requirements. Low across types /repositories | Varies depending on existing policies and infrastructure—lack of guidance/policy: less standardization | High within organization, low between organizations | Low across individual efforts, maybe higher in collective efforts | Prefers high–specialized views require standardized resources |

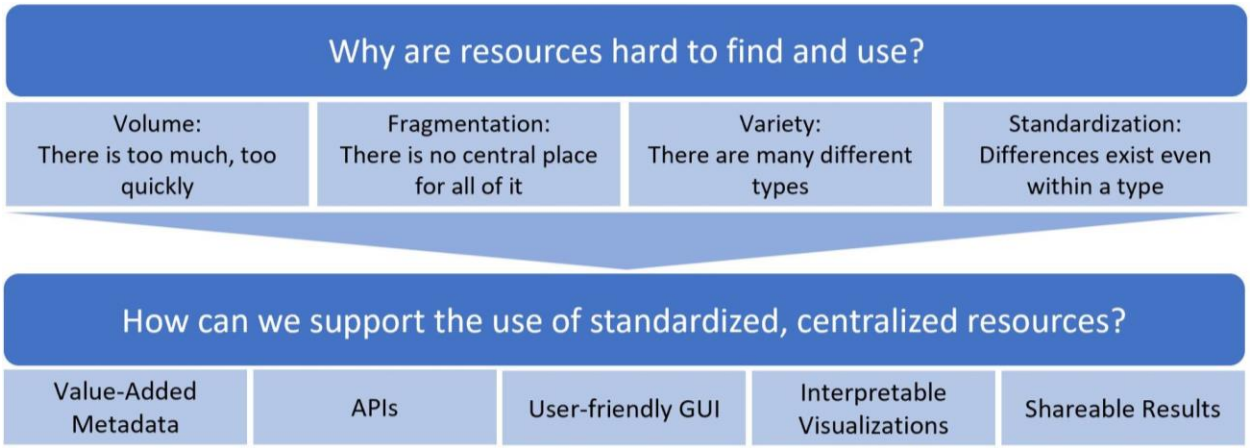



**Figure 1. What are resources, who contributes to the proliferation of resources, why are resources difficult to find and use, and how can we support their use?**

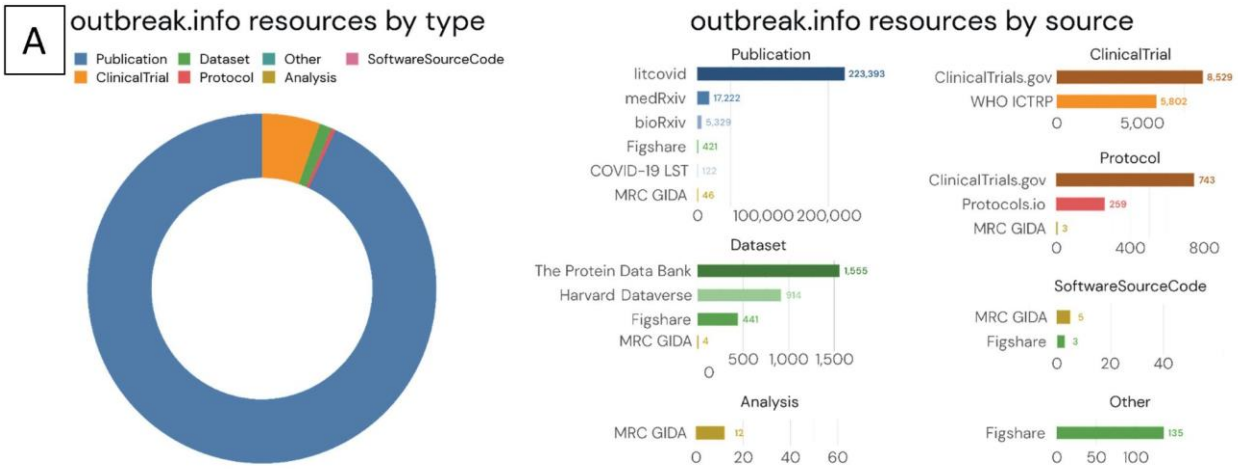
Once the issues of volume, fragmentation, variety, and standardization of resources are addressed, accessibility of the resulting resources for reuse must be addressed. Standardized, centralized resources are of no value if researchers are not able to leverage them. Researchers seeking to process information *en masse* will need an API, while researchers seeking to browse and explore will prefer a user-friendly interface. APIs themselves are less useful without a means of understanding the underlying metadata/data (such as documentation or a GUI), and a user-friendly search portal will be less useful without the inclusion of value-added metadata (such as ones supporting search/filter, linkage and exploration, or qualitative evaluation) for improving resource discovery and interpretation. Interpretability of metadata/data is influenced by the order in which information is presented. To address this challenge, the user interface must encourage exploration which gives users control over the information flow to suit their needs. Lastly, if a user has been able to successfully leverage the standardized, centralized resources, they should be able to easily save and share the results of their efforts.

Leveraging our experience from building the NIAID Data Portal (https://discovery.biothings.io/niaid/), we address the aforementioned challenges inherent in combining metadata from disparate and heterogeneous resources and making information more interpretable by building a website which integrates a searchable interface for a diverse, heterogeneous resources which we have collected and standardized (metadata) with surveillance reports on SARS-CoV-2 variants and mutants (data). Following implementation considerations for FAIRness (Jacobsen et al., 2020), our website includes programmatic access via APIs and a standardized metadata interface built off schema.org. Daily updates ensure that site users have up-to-date information, essential in the midst of a constantly changing research landscape. We will discuss issues with centralizing, standardizing, and returning resource metadata, epidemiological data, and supporting the use of the metadata/data by providing an explorable and interactive user-friendly interface with interpretable and shareable visualizations. Mutations, lineage, and variant reports are described elsewhere (Gangavarapu et al., 2022).

# Results

**Variety, Standardization, and Fragmentation: Developing a schema to harmonize a variety of resource types and standardizing the metadata.**

Schema.org provides a framework to standardize metadata for many different types of data found on the world wide web. However, these standards are not preserved across different types of data. For example, publication providers like PubMed typically use the 'author' property in their metadata, while dataset providers like Figshare and Zenodo are compliant with the DataCite schema and typically prefer 'creator'. Although both properties are valid for their respective schema.org classes, we normalized our schema to use 'author' for all 5 of our classes since we expected the volume of publications to dwarf all other classes of resources. We developed a schema that encompassed five types of resources based on their proliferation at the beginning of the pandemic and their importance to the research community: Publications, Datasets, Clinical Trials, Analysis, and Protocols. We added this schema to the Schema Registry of the Data Discovery Engine (Cano et al., 2021), a project to share and reuse schemas and register datasets according to a particular schema. Using this schema we ingested and harmonized metadata from an initial set of twelve key resources: LitCovid (Publications), bioRxiv/medRxiv (Publications), ClinicalTrials.gov (NCT) (ClinicalTrials), WHO International Clinical Trials Registry Platform (WHO ICTRP) (ClinicalTrials), Figshare (Datasets, Publications, and more), Zenodo (Datasets, Publications, and more), MRC Centre for Global Infectious Disease Analysis (Analyses, Publications, and More), Protocols.io (Protocols), PDB (Datasets), Harvard Dataverse (Datasets), COVID-19 Literature Surveillance Team (Publications), Immport (Datasets) (Figure 2B).



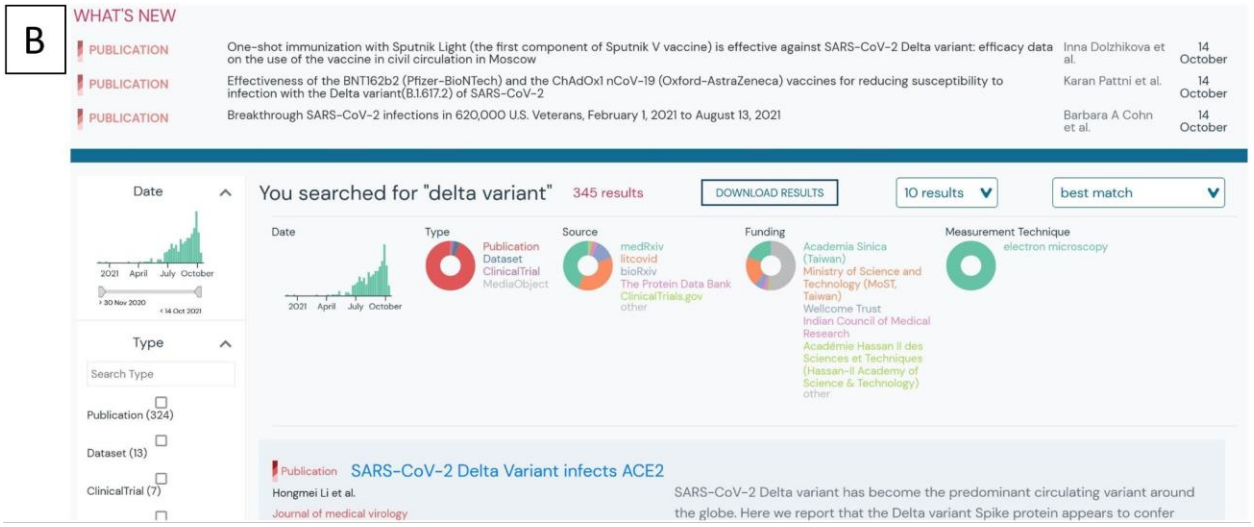



**Figure 2. Supporting resource centralization and standardization by developing a harmonizing schema. (A) Distribution of resources by resource type and source. Note that the x-axis in the bar graphs have different scales. (B) Heterogeneous and filterable resources (ie- publications, clinical trials, datasets, etc.) resulting from a single search of the phrase "Delta Variant"**

Sources of certain metadata did not map readily to existing schema.org classes. For example, clinical trials registries like NCT have one general schema for both observational and interventional studies, while schema.org provides separate classes for each of these types of studies. Since NCT was a primary source of clinical trials metadata for our research library, we tailored the Outbreak schema based on the combined general NCT schema. Fortunately, many dataset repositories offered schema.org-compliant metadata, even if the repositories differed in the metadata fields that were available. Datasets were imported from Zenodo, Figshare, PDB, and Harvard Datasets, while Protocols were imported from Protocols.io and NCT Protocols. Once our schema was developed, we created parsers (data plugins) to import compliant metadata from our initial set of resources. We assembled the data plugins into a single API via BioThings SDK (Lelong et al., 2021), and scheduled them to update on a daily basis to ensure up-to-date information. By leveraging the BioThings SDK, we had a technology stack that helped address the fragmentation issue by facilitating the integration of metadata from different pre-existing resources. LitCovid, a source of publications, provided the vast majority of entries into our resource library, followed by the preprint servers (bioRxiv and medRxiv). With a unified schema that could harmonize information across heterogeneous resource types, a single search (for example "delta variant") can return relevant publications, datasets, clinical trials, and more. (Figure 2B)

**Fragmentation and Standardization: Acquisition and integration of individual dataset metadata and collections of resource metadata via community curation**

At the start of the pandemic many curation efforts were neither coordinated, standardized, or easy to find; however, these efforts served an important role in organizing information early on.

Given the highly-fragmented, diffuse and frequently changing nature inherent to biomedical resources, we built outbreak.info with the idea that it should be expanded with the participation of the community. Not only is finding and adding resources to the collection an onerous process, it also requires us to know the full landscape of resources on the internet. Furthermore, many resources cannot be readily parsed for metadata useful for linkage, exploration, and evaluation. We enabled contributions of resource metadata in a variety of ways (Figure 3A). At the onset of a pandemic, time is of essence. Our contribution pipeline allows us to quickly and flexibly leverage and integrate the results of different kinds of resource generation pipelines such as those observed within the first month of the pandemic (Supplemental Figure 1).

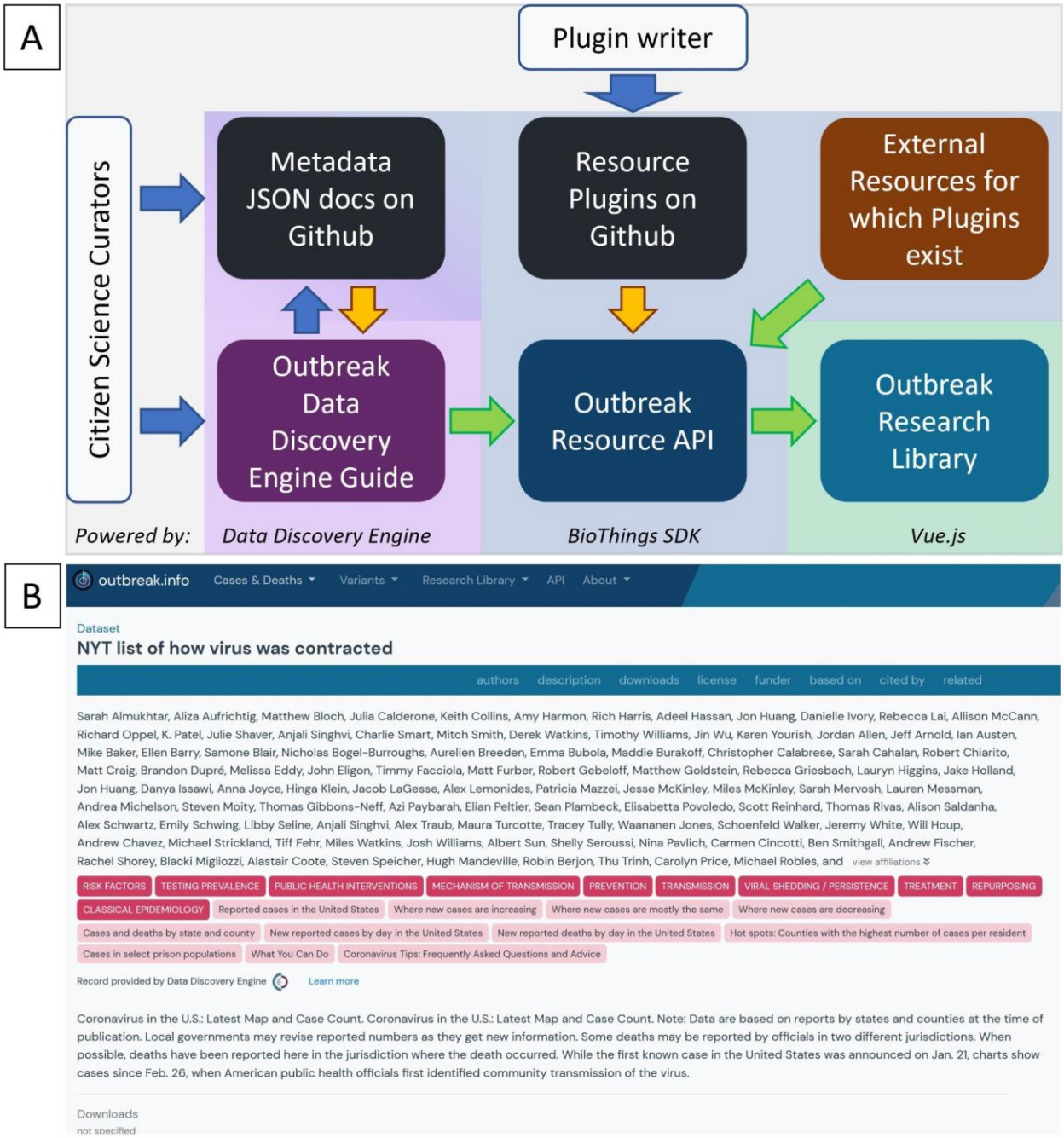


**Figure 3. Aggregating resource metadata by leveraging community contributions. (A) The community contribution pipeline and technology stack for Outbreak's Research Library. Curators may submit dataset metadata using the DDE built-in guide or from Github via the DDE/BioThings SDK. Python-savvy contributors can create parsers to contribute even more metadata via the BioThings SDK plugin architecture. A resource plugin allows the site to automatically ingest and update metadata from the corresponding external resource. Blue arrows indicate manual steps, yellow arrows indicate automatable steps after an initial set up, green arrows indicate completely automated steps. (B) An example of a detailed metadata record manually-curated by volunteers as it appears in the Research Library.**

For single datasets, contributors can submit the metadata via Outbreak's dataset submission guide on the Data Discovery Engine, which ensures that the curated metadata conforms to our schema. From there, it can be saved to GitHub, where it can be improved by other contributors via forking and pull requests. The DDE will automatically pass the information to the Outbreak Resources API where it is made discoverable with the Outbreak Research Library. We demonstrated its utility by asking two volunteers to annotate metadata from thirty different individual resources from across the internet and submitted the metadata for integration via the DDE. As seen in Figure 3B, community-contributed metadata using the DDE is standardized and can be exhaustively detailed. Although both of our volunteers provided values for many of the available metadata properties (name, description, topicCategories, keywords, etc.), one provided an extensive list of authors. Using the BioThings SDK in conjunction with the DDE allows us to centralize and leverage individualized curation efforts that often occur at the start of a pandemic. Additionally, collections of standardized datasets, publications and other resources can be submitted to the Outbreak Resources API by contributing a resource plugin. Resource plugins are BioThings-compatible Python scripts to harvest metadata from a source and standardize it to our schema; these parsers can be submitted by anyone with Python coding skills (Lelong et al., 2021).

**Supporting resource exploration after standardization and centralization by improved searching, linkage and evaluation of resources via community curation and machine learning**

Centralizing and standardizing the resources does not automatically make the resources explorable to a user. While centralizing and standardizing allows for search, aggregation and some filtering; additional metadata and a user-friendly interface is needed to allow thematic browsing/filtering and to enable iterative traversal from query to search result to refined query and vice versa. To support resource exploration and interpretation, we added properties (for value-added metadata) to every class in our schema that would support searching/filtering/browsing (topicCategories), linkage/exploration (correction, citedBy, isBasedOn, isRelatedTo), and interpretation (qualitative evaluations) of resources. We identified these properties to be value-added based on pre-existing citizen science and resource curation activities which we could leverage.

For example, citizen scientists categorized resources in their lists/collections by type (Dataset, Clinical Trials, etc.) in their outputs (Navarro, and Capdarest-Arest, 2020) or area of research (Epidemiological, Prevention, etc.)(Rah et al., 2020) as they found these classifications helpful for searching, filtering, and browsing their lists/collections. They also evaluated the level of evidence provided by these resources in order to improve its interpretability (i.e.- understanding of the credibility/quality)(Rah et al., 2020). Existing resources, too, were engaging in organizational efforts to enhance browsability, but the outputs of these organizational activities by existing resources and citizen scientists were often less explicit (i.e. - organized on the site or in unstructured text, but not actually in the metadata)--making their capture and integration more challenging. For instance, LitCovid organized publications into eight research areas such as Treatments or Prevention, but these classifications are not available in the actual metadata records for each publication.

To classify resources by topicCategory and improve search/browse/filtering capabilities in our user interface, we used a combination of existing work (LitCovid) and human curation to augment that categorization to provide higher specificity of topics and to extend to new types of data (datasets, clinical trials). We applied out-the-box logistic regression, multinomial naive bayes, and random forest algorithms to create models for classifying each resource as belonging or not belonging to each topic. These three algorithms were found to perform best on this binary classification task using out-the-box tests. For example, if a user wants to browse for all resources (or filter down search results) related to the prevention of COVID-19, they can select the appropriate topicCategory in the search/search results view of the resources (Figure 4A). Users can also easily traverse from a view of a resource record to start a new search by clicking on a topicCategory of interest (Figure 4B). We further enable exploration by populating the linkage properties (corrections, citedBy, isBasedOn, isRelatedTo) from citation metadata (whenever possible), corrections metadata (from LitCovid, when available), and via an algorithm for matching peer-reviewed papers in LitCovid with their corresponding preprints in bioRxiv/medRxiv. Together with the corrections metadata from LitCovid, the algorithm has matched over 2600 peer-reviewed articles with their corresponding preprints--making it possible for users to follow from a publication record from LitCovid to a publication record from bioRxiv/MedRxiv (Figure 4B).

Once a user has found a record of interest, they might wonder about the credibility of the resource. To populate resource evaluations so that users can assess the quality of a resource and tailor their interpretation accordingly, we leveraged the Oxford 2011 Levels of Evidence annotations generated by the COVID-19 Literature Surveillance (COVID-19 LST) team (Rah et al., 2020) as well as Digital Science's Altmetrics (Digital Science, n.d.). These evaluations are currently visible for records that have them, but in the future, we will enable users to further filter or sort search results by some measurement of quality (ie- Altmetrics: degree of access, or COVID-19 LST: level of evidence). Lastly, the integration of resources with other data on our site allows users to seamlessly traverse from a location-based variant report (Figure 4C left) to a specific variant report (Figure 4C middle), to resources on that specific variant (Figure 4C right). Details about exploring variant reports can be found in (Gangavarapu et al., 2022). In contrast, a similar attempt to explore resources outside the outbreak.info portal would require extensive manual searching

from multiple different sites (Supplemental Figure 1), each with their own interfaces and corresponding search/filtering nuances/capabilities.

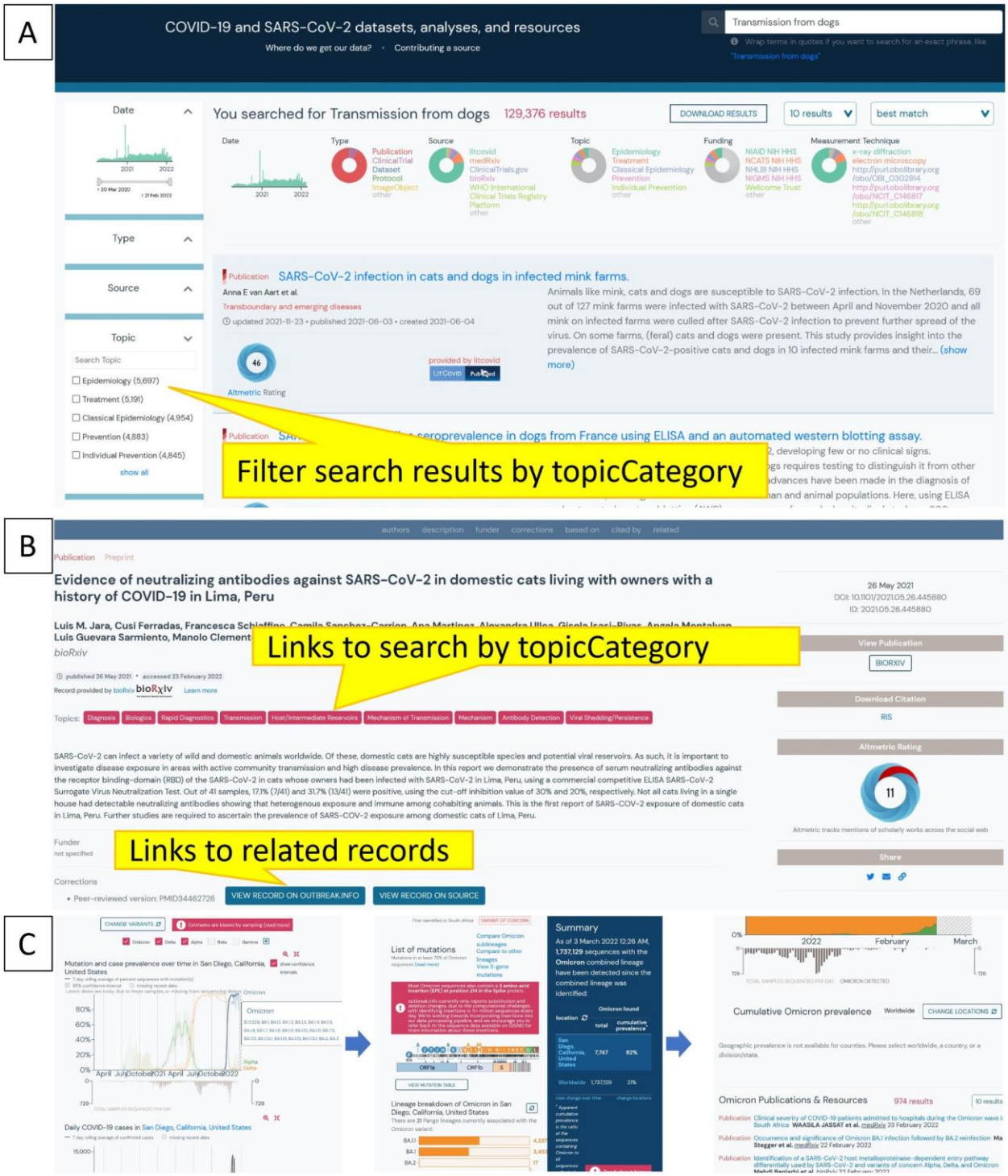



**Figure 4. Enabling exploration of the resources. (A) Selectable options for filtering results by topic category or other facets enhance searchability and exploration from the search results view. (B) Links to other records or to additional potential searches of interest**

**enabling further exploration from a record view. (C) Links from San Diego Variant Report (left) to Omicron Variant report (middle) to related resources (right)**

## Discussion

Over the course of the COVID-19 outbreak, researchers have shared the results of their work at unprecedented levels – exacerbating existing issues in resource volume, fragmentation, variety, and standardization. These issues make it challenging to assemble/aggregate, traverse, and maintain up-to-date resources. Further, the urgency of a pandemic requires that these issues be addressed quickly, and in a scalable manner to be able to accommodate more data flexibly. We launched outbreak.info within 2 months of the start of the COVID-19 pandemic to address these issues and to highlight barriers in rapidly sharing research outputs in the midst of a pandemic.

To address the structure and standardization issue, we developed a schema and standardized metadata across different resources and ingested them into an accessible API and user-friendly search-and-filter, web-based interface. In addition to issues with standardizing and mapping metadata between resources, it is challenging to maintain a resource library that imports metadata from so many sources, particularly when the metadata updates daily and is prone to change structure. Any changes to an external API or the way in which metadata is offered by an external site will necessitate a change in the parser. The resource API utilizes the BioThings SDK plugin architecture to handle errors in individual parsers without affecting the availability of the API itself. Using the plugin architecture allows the introduction and maintenance of the individual resource parsers to be crowdsourced to anyone with some basic knowledge or Python coding and a GitHub account. Although resource plugins allow outbreak.info to ingest large amounts of standardized metadata, there are still many individual datasets and research outputs scattered throughout the web. Since it is not feasible for one team to locate, identify, and collect standardized metadata from these individual datasets and research outputs, we leveraged the Data Discovery Engine to enable crowdsourcing and citizen science participation in the curation of individual resource metadata.

At the onset of our data harvesting and harmonization efforts, we focused on creating a unified search interface backed by a common schema.org-based schema. With an extendable pipeline in place, we focused next on augmenting the existing metadata by adding properties to help researchers find information more quickly: topic categorization to group related research, resource linking to connect related entities along the data lifecycle from data generation through publication, and integrating external evaluations of the research trustworthiness using a combination of human curation and automated methods.

Citizen scientists have played an active role in data collection (Birkin, Vasileiou, and Stagg 2021) (Rohwer Lab at San Diego State University, n.d.) and making information more accessible (Rah et al., 2020)(Allen Institute For AI, 2020) throughout the current pandemic. Given their ability to

perform information extraction (Tsueng et al., 2020) and their immense contributions to classification tasks (Blickhan, Trouille, and Lintott, 2018), we incorporated citizen science contributions into the training data for classifying resources into our topic categories. Some resource aggregators have used clustering algorithms to categorize the entries in their resource libraries--though many only aggregate resources of a single type (ie- publications). We employed a different approach due to the heterogeneity of our resources, but our API is openly accessible, so anyone is welcome to apply clustering approaches to classify the entries. The research library and resources API has been used by others to monitor available research on the immune escape of variants, or create custom research digests. The Radx-Rad Data Coordination Center (https://www.radxrad.org) is utilizing the Outbreak API to collect articles for customized research digests for its partners. Using SearchOutbreak ( https://searchoutbreak.netlify.app ), users select topics based on information submitted from partners. These are turned into queries for the Outbreak API. These are stored in a database (airtable). Weekly the Outbreak API is queried, and new articles are added to the digests which are available at the website. A workflow sends an email to subscribed users. While these digests are not currently available to the public, they are expected to be available to the public in the future (Valentine D and Radx. 2021).

In addition to generating metadata values for improved searching and filtering, we enabled linkages between resources in our schema. For instance, ideally a publication about a clinical trial would link to its clinical trial record, protocols used to collect the data, datasets used in their analyses, and software code underlying the analyses to enable a more meaningful understanding of this trial. However, these connections rarely exist within the metadata; as a result, we have generated linkages between preprints and peer-reviewed publications, and hope to create more linkages between other resource types. Challenges to include these linkages included: the lack of unique identifiers, inconsistent use of citation metadata fields between resources, and the lack of structured linkage metadata. Many Datasets and Analysis often don't have unique identifiers, or aren't referred to by their unique identifiers if they do have them. For example, the ONS Deaths Analysis does not have a unique identifier as assigned by Imperial College London, lacks any citation metadata fields, and instead mentions a potential linkage to an Imperial College London report in its mention of limitations (Imperial College COVID-19 Response Team, n.d.). Although preprints from bioRxiv and medRxiv may have links to the corresponding peer-reviewed manuscript on the bioRxiv site, this information is not accessible via their API necessitating the use of algorithms to generate these links.

While we have developed a framework for addressing resource volume, fragmentation, and variety that can be applicable to future pandemics, our efforts during this framework exposed additional limitations in how data/metadata is currently collected and shared. Researchers have embraced pre-publications, but resources (especially datasets and computational tools) needed to replicate/extend research results are not linked in ways that are discoverable. Although many journals and funders have embraced dataset/source code submission requirements, the result is that the publication of datasets and software code are still heavily based in publications instead of in community repositories with well-described metadata to promote discoverability and reuse. In the outbreak.info Research Library, the largest research output by far is publications, while dataset submission lags in standardized repositories encouraged by the NIH such as ImmPort,

Figshare, and Zenodo. We hypothesize that this disparity between pre-print and data sharing reflects the existing incentive structure, where researchers are rewarded for writing papers and less for providing good, reusable datasets. Ongoing efforts to improve metadata standardization and encourage schema adoption (such as the efforts in the Bioschemas community) will help make resources more discoverable in the future – provided researchers adopt and use them. For this uptake to happen, fundamental changes in the incentive structure for sharing research outputs may be necessary.

# Conclusions

Within the eighteen months since SARS-CoV-2 was first identified as the infectious agent of the COVID-19 pandemic, there have been over 170 million cases and nearly 4 million deaths. As those numbers continue to grow, so too does the research and understanding of the causes and consequences of the spread of this virus. Given that there will be other pandemics in the future, we demonstrate how we built and launched a standardized, searchable platform for exploring COVID-19 research outputs and genomics data within two months of the pandemic. We address many of the challenges faced when assembling a collection of heterogeneous research outputs and data into a searchable platform. Our platform, outbreak.info, seeks to make COVID-19 data more findable, accessible, interoperable, reusable and interpretable by addressing many data management issues exposed by an urgent and frequently-changing situation. Our site is used by a wide variety of professionals including journalists, members of the medical and public health communities, students, and biomedical researchers (blog.outbreak.info). On average, the site receives over a thousand hits per day and its visualizations are shared frequently across social media platforms like Twitter. This platform is also easily extensible to add new metadata sources, allowing the Research Library to grow with the pandemic as research changes.

# Methods

**Schema development**

The development of the schema for standardizing our collection of resources is as previously described (Cano et al., 2021). Briefly, we prioritized 5 classes of resources which had seen a rapid expansion at the start of the pandemic due to their importance to the research community: Publications, Datasets, Clinical Trials, Analysis, and Protocols. We identified the most closely related classes from schema.org and mapped their properties to available metadata from 2-5 of the most prolific sources. Additionally, we identified subclasses which were needed to support our main 5 classes and standardized the properties within each class. In addition to standardizing ready-to-harvest metadata, we created new properties which would support the linkage, exploration, and evaluation of our resources. Our schema was then refined as we iterated through the available metadata when assembling COVID-19 resources. The Outbreak schema is available at https://discovery.biothings.io/view/outbreak.

**Assembly of COVID-19 resources**

The resource metadata pipeline for outbreak.info includes two ways to ingest metadata. First, metadata can be ingested from other resource repositories or collections using the BioThings SDK data plugins. For each resource repository/collection, a parser/data plugin enables automated import and updates from that resource. Second, metadata for individual resources can be submitted via an online form. To assemble the outbreak.info collection of resources, we collected a list of over a hundred separate resources on COVID-19 and SARS-CoV-2. This list (Supplemental Table X) included generalist open data repositories, biomedical-specific data projects including those recommended by the NIH (BioMedical Informatics Coordinating Committee, 2020) and NSF (National Science Foundation, 2013) to house open data, and individual websites we came across through search engines and other COVID-19 publications. Prioritizing those resources which had a large number of resources related to COVID-19, we selected an initial set of 2-3 sources per resource type to import into our collection. Given the lack of widespread repositories for Analysis Resources, only one source would be included in our initial import (Imperial College London). An Analysis resource is defined as a frequently-updated, web-based, data visualization/interpretation/analysis resource.

**Community curation of resource metadata**

Resource plugins such as those used in the assembly of COVID-19 resources do not necessarily have to be built by our own team. We used the BioThings SDK (Lelong et al., 2021) and the Data Discovery Engine (Cano et al., 2021) so that individual resource collections can be added by writing BioThings plugins that conform to our schema. Expanding available classes of resources can be done easily by extending other schema.org classes via the DDE Schema Playground at https://discovery.biothings.io/schema-playground. Community contributions of resource plugins can be done via GitHub. In addition to contributing resource plugins for collections/repositories of metadata, users can enter metadata for individual resources via the automatic guides created by the Data Discovery Engine. To investigate potential areas of community contribution, we asked two volunteers to inspect 30 individual datasets sprinkled around the web and collect the metadata for these datasets. We compared the results between the two volunteers and their combined results were subsequently submitted into the collection via the Data Discovery Engine's Outbreak Data Portal Guide at https://discovery.biothings.io/guide/outbreak/dataset. Improvements or updates for manually curated metadata can be submitted via GitHub pull requests.

**Community curation of searching, linkage, and evaluation metadata and scaling with machine learning**

In an effort to enable improved searching and filtering, we developed a nested list of thematic or topic-based categories based on an initial list developed by LitCovid (Chen, Allot and Lu, 2020) with input from the infectious disease research community and volunteer curators. The list consists of 11 broad categories and 24 specific child categories. Whenever possible, sources with thematic categories were mapped to our list of categories in order to develop a training set for basic binary (in group/out group) classifications of required metadata fields such as (title, abstract and/or description). If an already-curated training set could not be found for a broad category, it would be created via an iterative process involving term/phrase searching on LitCovid, evaluating

the specificity of the results, identifying new search terms by keyword frequency, and repeating the process. To generate training data for classifying resources into specific topic categories, the results from several approaches were combined. These approaches include direct mapping from LitCovid research areas, keyword mapping from LitCovid, logical mapping from NCT Clinical Trials metadata, the aforementioned terms search iteration, and citizen science curation of Zenodo and Figshare datasets.

The efforts of our two volunteers suggested that non-experts were capable of thematically categorizing datasets, so we built a simple interface to allow citizen scientists to thematically classify the datasets that were available in our collection at that point in time. Each dataset was assigned up to 5 topics by at least three different citizen scientists. Citizen scientists were asked to prioritize specific topic categories over broader ones. 90 citizen scientists participated in classifying 500 datasets pulled from Figshare and Zenodo. The citizen science curation site was originally hosted at https://curate.outbreak.info, the code for the site can be found at https://github.com/outbreak-info/outbreak.info-resources/tree/master/citsciclassify and the citizen science classifications at https://github.com/outbreak-info/topic_classifier/blob/main/data/subtopics/curated_training_df.pickle. These classifications have been incorporated into the appropriate datasets in our collection, and have been used to build our models for topic categorization. Basic in-group/out-group classification models were developed for each category using out-the-box logistic regression, multinomial naive bayes, and random forest algorithms available from SciKitLearn. The topic classifier can be found at https://github.com/outbreak-info/topic_classifier.

In addition to community curation of topic categorizations, we identified a citizen science effort, the COVID-19 Literature Surveillance Team (COVID19-LST), that was evaluating the quality of COVID-19 related literature. The COVID-19 LST consists of medical students, practitioners and researchers who evaluate publications on COVID-19 based on the Oxford Levels of Evidence criteria and write Bottom Line, Up Front summaries (Rah et al., 2020). With their permission, we integrated their outputs (daily reports/summaries, and evaluations) into our collection.

We further integrated our publications by adding structured linkage metadata, connecting preprints and their peer-reviewed versions. We performed separate Jaccard's similarity calculations on the title/text and authors for preprint vs LitCovid Publications. We identified thresholds with high precision, low sensitivity and binned the matches into (expected match vs needs review). We also leveraged NLM's pilot preprint program to identify and incorporate additional matches. The code used for the preprint-matching can be found at https://github.com/outbreak-info/outbreak_preprint_matcher. Expected matches were linked via the `correction` property in our schema.

**Harmonization and integration of resources and genomics data**

The integration of genomics data from GISAID is discussed elsewhere (Gangavarapu et al., 2022). We built separate API endpoints for our resources (metadata resources API) and genomics (genomics data API) using the BioThings SDK (Lelong et al., 2021). Data is available via our API at api.outbreak.info and through our R package, as described in Gangavarapu et al..

## Acknowledgements

We would like to thank Jasmine Rah, Brennan Joseph Enright, Justin Doroshenko, Thamanna Nishath and the rest of the COVID-19 Literature Surveillance Team for their incredible work and allowing us to share their work. We would like to thank Tom Adams and Craig Lazarchick for their work in identifying metadata from various individual datasets and their extensive feedback. We thank Sue Andarmani for her suggestions and feedback on dataset categories. We thank all of the contributors found at (https://blog.outbreak.info/dataset-topic-category-contributors) for taking the time to categorize datasets. We thank David Valentine for sharing details about his netlify app as part of the Radx- Radical Data Coordination Center which is funded by NIH (U24LM013755).

## Funding

Work on outbreak.info was supported by National Institute for Allergy and Infectious Diseases (5 U19 AI135995-02), the National Center For Advancing Translational Sciences (5 U24 TR002306) and Centers for Disease Control and Prevention (75D30120C09795).

## References

Allen Institute For AI, 2020. COVID-19 Open Research Dataset Challenge (CORD-19). [online] Available at: <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge> [Accessed 24 August 2021].

bioRxiv, 2021. COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv. [online] Available at: <https://connect.biorxiv.org/relate/content/181> [Accessed 17 August 2021].

Birkin, L. J., Vasileiou, E., and Stagg, H. R. 2021. Citizen science in the time of COVID-19. *Thorax*, *76*(7), 636-637. 10.1136/thoraxjnl-2020-216673

Blickhan, S., Trouille, L. and Lintott, C. J. 2018. Transforming research (and public engagement) through citizen science, Proceedings of the International Astronomical Union. Cambridge University Press, 14(A30), pp. 518–523. doi: 10.1017/S174392131900526X.

Cano, M., Tsueng, G., Zhou, X., Hughes, L., Mullen, J., Xin, J., Su, A. and Wu, C., 2021. Schema Playground: A tool for authoring, extending, and using metadata schemas to improve FAIRness of biomedical data. bioRxiv.

Chen, Q., Allot, A. and Lu, Z., 2020. LitCovid: an open database of COVID-19 literature. Nucleic Acids Research, 49(D1), pp.D1534-D1540.

ClinicalTrials.gov. 2018. Protocol Record Schema - XML Schema for electronic transfer of protocol information into the ClinicalTrials.gov Protocol Registration System. [online] Available at: <https://prsinfo.clinicaltrials.gov/ProtocolRecordSchema.xsd> [Accessed 17 August 2021].

Digital Science (n.d.) About Us. Altmetric. [online] Available at: <https://www.altmetric.com/about-us/> [Accessed 07 March 2022].

Gangavarapu, K., Latif, A. A., Mullen, J. L., Alkuzweny, M., Hufbauer, E., Tsueng, G., Haag, E., Sattler, R., Cano, M. A., Zhou, X., Qian, Z., Zeller, M., Matteson, N., Levy, J., Suchard, M. A., Wu, C., Anderson, K. G., Su, A. I., and Hughes, L. D. 2022. Outbreak.info: Real-time surveillance of SARS-CoV-2 mutations and variants. medRxiv. doi: 10.1101/2022.01.27.22269965

Hyndman, A., 2020. A Figshare COVID-19 Research Publishing Portal. [online] https://COVID19.figshare.com/. Available at: <https://figshare.com/blog/A_Figshare_COVID-19_Research_Publishing_Portal/558> [Accessed 17 August 2021].

Hahnel, M., 2018. What Google Dataset Search means for academia. [online] Figshare Blog. Available at: <https://figshare.com/blog/What_Google_Dataset_Search_means_for_academia/422> [Accessed 17 August 2021].

Imperial College COVID-19 Response Team. (n.d.). *ONS excess deaths*. Imperial College London. Retrieved 10-21-2021, from http://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/covid-19-reports/

Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C., Goble, C., Guizzardi, G., Hansen, K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R., Imming, M., Jeffery, K., Kaliyaperumal, R., Kersloot, M., Kirkpatrick, C., Kuhn, T., Labastida, I., Magagna, B., McQuilton, P., Meyers, N., Montesanti, A., van Reisen, M., Rocca-Serra, P., Pergl, R., Sansone, S., da Silva Santos, L., Schneider, J., Strawn, G., Thompson, M., Waagmeester, A., Weigel, T., Wilkinson, M., Willighagen, E., Wittenburg, P., Roos, M., Mons, B. and Schultes, E., 2020. FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence, [online] 2(1-2), pp.10-29. Available at: <https://doi.org/10.1162/dint_r_00024> [Accessed 4 March 2022].

James, Phillip, and citizen scientists. 2020. Staying Home Club. Github. https://github.com/phildini/stayinghomeclub

Joachimiak, Marcin, Nomi Harris, Justin Reese, and Citizen Scientists. 2020. SARS-COV-2 and

COVID-19 datasets. Google Spreadsheets. https://docs.google.com/spreadsheets/d/1eMhot7MjusyM7_2IBnzqi7RlzWWoYnfheWhMgDlPToQ/edit#gid=0

Kaiser, J. 2020. 'Every day is a new surprise.' Inside the effort to produce the world's most popular coronavirus tracker. *Science*. 10.1126/science.abc1085

Lelong, S., Zhou, X., Afrasiabi, C., Qian, Z., Cano, M. A., Tsueng, G., Xin, J., Mullen, J., Yao, Y., Avila, R., Taylor, G. K., Su, A. I., and Wu, C. 2021. BioThings SDK: a toolkit for building high-performance data APIs in biomedical research. *Bioinformatics*, 10.1093/bioinformatics/btac017

National Science Foundation. 2013. Open Data at NSF. National Science Foundation. https://www.nsf.gov/data/

Navarro, Christy, and Capdarest-Arest, Nicole. 2020. COVID-19 Open Dataset Sources. Google Spreadsheets. https://docs.google.com/spreadsheets/d/10t3vtULr3nTz7mrlKj0rldUys47wsIfOVReHnx3Xu18/edit#gid=0

NIH OPA, 2020. iSearch COVID-19 Portfolio [online] Available at: <https://icite.od.nih.gov/covid19/search> [Accessed 24 August 2021].

Noren, Laura E., Gina Neff, and Zhaofeng Li. 2020. Institutional Response to Covid. Google Spreadsheets. https://docs.google.com/spreadsheets/d/1IbF_wlmldVssG5spcmNE82nR9btcbF7rUlEqtcXW03o/edit#gid=0

Rah, J., Doroshenko, J., Nishath, T., Gordon, W., Enright, B., Archuleta, C., Nelson, E., Taylor, S., Parulian, N., Woodfield, S., Maloney, C., Knoll, A., Thevuthasan, S., Hansen, A., Ellingsen, K., Woodfield, M., Richards, C., Shelton, E., Nguyen, D., Ghering, J., Baker, J., Kumarasan, K., Meckl, R., Gallagher, T., Awan, Z., Graham, V., Bellardi, K., Dharmendran, A., Reidy, E., Kal, S. and Zaidi, S., Smith, W., 2020. COVID-19 Literature Surveillance Team. [online] Covid19lst.org. Available at: <https://www.covid19lst.org/copy-of-about> [Accessed 17 August 2021].

Rohwer Lab at San Diego State University. n.d. #swab4corona. CoVID-19 Citizen Science. Accessed 09 17, 2021. https://covidsample.org/.

Schema.org. 2015. About Schema.org. Accessed 10 18, 2021 at https://schema.org/docs/about.html.

Skenderi, J., Costello, M., Simon-Kennedy, A., and citizen scientists. 2020. "COVID-19 Resource Library." Google Spreadsheets. https://docs.google.com/spreadsheets/u/2/d/1cqxDAg4jMHXI6gHOnoV8HqDdRHnmxEJRl-

bhhpe1HEo/htmlview#

Tsueng G, Nanis M, Fouquier JT, Mayers M, Good BM, Su AI. Applying citizen science to gene, drug and disease relationship extraction from biomedical abstracts. Bioinformatics. 2020 Feb 15;36(4):1226-1233. doi: 10.1093/bioinformatics/btz678. PMID: 31504205; PMCID: PMC8104067.

Valentine, D and Radx. 2021. SearchOutbreak. Radical Data Coordination Center. Accessed 10 13, 2021 https://searchoutbreak.netlify.app

World Health Organization, 2020. Novel Coronavirus (2019-nCoV): situation report, 1. World Health Organization. https://apps.who.int/iris/handle/10665/330760