HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2023 October 01.

Published in final edited form as:

Nat Methods. 2023 April ; 20(4): 536–540. doi:10.1038/s41592-023-01770-w.

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

Ginger Tsueng[1], **Julia L. Mullen**[1], **Manar Alkuzweny**[2,3], **Marco Cano**[1], **Benjamin Rush**[4], **Emily Haag**[1], **Jason Lin**[1], **Dylan J. Welzel**[1], **Xinghua Zhou**[1], **Zhongchao Qian**[1], **Alaa Abdel Latif**[3], **Emory Hufbauer**[3], **Mark Zeller**[3], **Kristian G. Andersen**[3,5], **Chunlei Wu**[1,5,6], **Andrew I. Su**[1,5,6], **Karthik Gangavarapu**[3,7], **Laura D. Hughes**[1]

[1]Department of Integrative, Structural and Computational Biology, the Scripps Research Institute, La Jolla, CA, USA.

[2]Department of Biological Sciences,University of Notre Dame, Notre Dame, IN, USA.

[3]Department of Immunology and Microbiology, the Scripps Research Institute, La Jolla, CA, USA.

[4]Ocuvera, Lincoln, NE, USA.

[5]Scripps Research Translational Institute, La Jolla, CA, USA.

[6]Department of Molecular Medicine, the Scripps Research Institute, La Jolla, CA, USA.

[7]Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA.

Abstract

Outbreak.info Research Library is a standardized, searchable interface of coronavirus disease 2019 (COVID-19) and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) publications, clinical trials, datasets, protocols and other resources, built with a reusable framework. We developed a rigorous schema to enforce consistency across different sources and resource types and linked related resources. Researchers can quickly search the latest research across data repositories, regardless of resource type or repository location, via a search interface, public application programming interface (API) and R package.

Correspondence and requests for materials should be addressed to Ginger Tsueng or Laura D. Hughes. gtsueng@scripps.edu; lhughes@scripps.edu.

Author contributions

L.D.H., K.G., M.C., E. Haag, J.L.M., X.Z., Z.Q., E. Hufbauer, C.W., A.I.S., K.G.A., A.A.L., M.Z., G.T., J.L. and D.J.W. contributed to the design, construction and/or maintenance of the outbreak.info website and data pipelines. K.G., M.A. and L.D.H. designed and built the R outbreak. info package. M.A., A.A.L., K.G., E. Haag, E. Hufbauer, M.Z., K.G.A. and L.D.H. designed and linked the variant reports. L.D.H., J.L.M., G.T. and M.C. developed the schemas. E. Haag performed the usability studies. B.R. developed the curation app. L.D.H., G.T., E. Haag, K.G., M.Z. and J.L.M. contributed to writing and editing the manuscript.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592–023-01770-w.

Peer review information *Nature Methods* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Lin Tang, in collaboration with the *Nature Methods* team. Peer reviewer reports are available.

Competing interests

K.G.A. has received consulting fees and/or compensated expert testimony on SARS-CoV-2 and the COVID-19 pandemic. The remaining authors declare no competing interests.

In January 2020, SARS-CoV-2 was identified as the virus responsible for a series of pneumonia cases with unknown origin[1]. As the virus spread globally, the scientific community rapidly released research outputs (such as publications, clinical trials and datasets) and resources (websites, portals and more). The frequently uncoordinated generation and curation of resources exacerbated four challenges in finding and using them: volume, fragmentation, variety and standardization (Supplementary Fig. 1). While many specialized websites were developed independently[2–7], a centralized and standardized repository for finding COVID-19 research has limited researchers' ability to discover these resources and translate them into insights about the virus.

To address the fragmented research landscape, individual and community efforts created shared Google spreadsheets[8–10] to aid in discoverability, but these efforts were not scalable and often lacked metadata to promote findability (aside from Navarro and Capdarest-Arest[10]). Several projects attempted to address the volume and fragmentation issues through large-scale aggregation but failed to tackle variety, focusing on a single resource type such as publications[11,12]. Even within a particular type of resource, standardization issues abound. Repositories pivoted quickly to curate COVID-19 content from their collections using pre-existing metadata standards but were often not interoperable with other sources. For example, PubMed created LitCovid[13] based on their MEDLINE standards, and the National Clinical Trials Registry cataloged COVID-19 clinical trials using their schema[14], but the World Health Organization (WHO) International Clinical Trials Registry uses different conventions. Similarly, Zenodo[15] and Figshare[16] do not agree on the marginality, cardinality and property names[17,18], despite compatibility with the standards of https://schema.org.

We address issues in metadata volume, variety, standardization and fragmentation by creating a single searchable index of COVID-19 publications, clinical trials, datasets and more: the outbreak.info Research Library. To address variety and standardization, we developed a harmonized schema based on https://schema.org, a framework standardizing metadata across the internet. Using this schema, we harvested and harmonized metadata from 16 resources (Fig. 1a). Daily updates ensure that site users have up-to-date information, essential amid a constantly changing research landscape.

Next, to address volume and fragmentation, we developed a web-based search portal for researchers to browse across the centralized and standardized resources (https://outbreak.info/resources) and an API to access and analyze information en masse (https://api.outbreak.info). Within the search interface, users can search, filter and view related records and share the associated metadata to easily query across resource repositories and types. For instance, a single query (for example, 'Delta variant') to our API can return relevant publications, datasets, clinical trials and more (Fig. 1b), and the Research Library summarizes the search results in visualizations to promote exploration. For instance, the histogram in Fig. 1b indicates that the number of resources mentioning 'Delta variant' began growing in mid 2021 and declined in the summer of 2022, and the donut charts show that LitCovid is the dominant source. To ensure ease of use of our Research Library, we conducted usability studies and iteratively improved our site (Supplementary Fig. 2).

To further address fragmentation and maintenance issues, we use modular infrastructure, allowing easy addition of new data sources, including community contributions. Citizen scientists have played an active role in data collection[19] (https://covidsample.org/) and accessibility[12,20] throughout the pandemic. Given the highly fragmented, diffuse and frequently changing nature inherent to biomedical research, we built in three mechanisms to expand the Research Library through community participation (Supplementary Fig. 3a).

First, contributors can submit individual or multiple datasets via an online form that ensures that the curated metadata conform to our schema. Second, leveraging the benefits of human curation, the community-contributed metadata using the form can be exhaustively detailed (Supplementary Fig. 3b) and can further be augmented through pull requests on GitHub. Lastly, anyone with Python coding skills can submit collections of standardized datasets, publications and other resources to the Outbreak Resources API by contributing a resource parser. Our community-contribution pipeline allows us to integrate the uncoordinated data-curation efforts quickly and flexibly, particularly apparent at the start of the pandemic (Supplementary Fig. 4).

To support resource exploration and interpretation, we added properties (value-added metadata) to every class in our schema that would support searching, filtering and browsing (topicCategories, Supplementary Fig. 5a); linkage and exploration (correction, citedBy, isBasedOn, isRelatedTo; Supplementary Fig. 5b); and interpretation (qualitative evaluations) of resources. We selected these properties based on pre-existing citizen science- and resource-curation activities, suggesting their value in promoting discoverability. For example, citizen scientists categorized resources in their lists or collections by type (Dataset, ClinicalTrials, etc.) in their outputs[10] or area of research (epidemiological, prevention, etc.)[20] as they found these classifications helpful for searching, filtering and browsing their lists or collections. Given the ability of citizen scientists to perform information extraction[21] and their immense contributions to classification tasks[22], we incorporated citizen science contributions into the training data for classifying resources into topic categories. Citizen scientists also provided Oxford 2011 Levels of Evidence annotations to improve its interpretability (that is, understanding the credibility or quality of the resource)[20]. To further enable assessment of the quality of a resource, we leveraged Digital Science's Altmetric ratings[23].

Finally, we integrated resources with the analyses that we developed to track SARS-CoV-2 variants of concern (VOCs)[24], sets of mutations within the virus associated with increased transmissibility, virulence and/or immune evasion. Researchers can seamlessly traverse from a specific variant report such as Omicron to resources in the Research Library that help understand its behavior (Supplementary Fig. 5c), and variant searches are among our most commonly queried terms (Supplementary Table 1a). Without a centralized search interface with linked records such as outbreak.info, a similar attempt to explore resources would require extensive manual searching from multiple different sites (Supplementary Fig. 6), each with their own interfaces and corresponding search capabilities.

To demonstrate the unique features of the outbreak.info Research Library, we explored the dynamics of research into SARS-CoV-2 variants over time to address two key questions:

(1) how has the research community responded to the emergence of new variants and (2) how has that response changed over time? We extracted research related to variants in the Research Library using the query ‘variant OR lineage’, allowing us to query metadata from 16 sources of different research types simultaneously (Fig. 2a). Over 10,000 separate entries about variants are within the Library as of October 2022, including publications, datasets, clinical trials, protocols and more. Using filters and the quality metrics provided through Altmetric badges, we quickly identified which results have been recognized by the community via Altmetric scores, such as a quantitative PCR protocol with reverse transcription (RT–qPCR) to screen VOCs (Fig. 2b). Clearly, variants are an active area of research, but has this enthusiasm changed over time? Using the outbreak.info R package, we accessed the harmonized metadata to examine the proportion of research related to variants in the Research Library over time. We observed an increase in research on variants following the first identification of VOCs such as Alpha (B.1.1.7*) and Beta (B.1.351*) (Fig. 2c). This increase was even more prominent for the Omicron (B.1.1.529*) variant in late 2021; we hypothesize that this increase was due to the heightened awareness of the value in studying variants among the scientific community, and early indications that the variant could be of global concern (high growth rate of Omicron and the presence of many mutations in important sites). To examine how research differed by VOC over time, we constructed queries for each VOC, including its Pango lineage name and associated sublineages. With the three VOCs that became the dominant worldwide form of SARS-CoV-2 (Alpha, Delta and Omicron), we find that the increase in research on these VOCs mirrors the rise in worldwide prevalence for each variant, with the research output roughly proportional to global prevalence (Fig. 2d). With Alpha and Delta, there was a slight lag in research publications that was not observed with Omicron, and research on Omicron over the last 10 months has dwarfed that for the other VOCs. Lastly, research on previously circulating variants (Alpha, Beta, Gamma, Delta) continues, even though these variants are rarely detected presently, and focuses on retrospective analyses, fundamental studies on mechanisms of action, Omicron comparisons and studies of recombinant variants. In sum, the research community’s response to the emergence of new variants has been robust, has become a greater focus of overall research effort over the last year and quickly pivots to studying the dominant variant.

The outbreak.info Research Library and resources API have been widely used by the external community, including journalists, members of the medical and public health communities, students and biomedical researchers[25]. For instance, the RADx-Rad Data Coordination Center created the SearchOutbreak app (https://searchoutbreak.netlify.app), which uses the Outbreak API to collect articles for customized research digests for its partners[26]. On average, the Research Library receives nearly 3,000 pageviews per month, of which 85% are unique visitors (Supplementary Table 1b). The Research Library site has been used for over 11,000 unique searches, and the Research Library API receives an average of nearly 63,000 unique hits per month (including web traffic and programmatic access). Some limitations of the Research Library include incomplete or unstructured metadata descriptions provided by the sources and optimally querying these descriptions, which often include acronyms and synonyms. Future work will focus on augmenting the harvested metadata and optimizing search results to provide the most salient results to users.

While the unprecedented amount of research on COVID-19 offers new opportunities to accelerate the pace of research, the difficulty in finding research amid this 'infodemic' remains a fundamental challenge. In the outbreak.info Research Library, we address many of these challenges to assemble a collection of heterogeneous research outputs and data from distributed data sources into a searchable platform. Our metadata-processing platform is modular, allowing easy extension to add new metadata sources including contributions from the community, allowing the Research Library to grow with the pandemic as research changes. To enable further analysis, we enable programmatic access to the standardized library. Lastly, with the embrace of open science stored in decentralized sources, quickly finding information will be critical for the next pandemic. Our approach to unify metadata across repositories will serve as a template for rapidly creating a unified search interface to aggregate research outputs for any pathogen or any research domain.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592–023-01770-w.

Methods

Schema development

The development of the schema for standardizing our collection of resources is as previously described[27]. Briefly, we prioritized six classes of resources that had seen a rapid expansion at the start of the pandemic due to their importance to the research community: publications, datasets, clinical trials, analyses, protocols and computational tools. We identified the most closely related classes from https://schema.org and mapped their properties to available metadata from two to five of the most prolific sources. Additionally, we identified subclasses that were needed to support the aforementioned six classes and standardized the properties within each class. In addition to standardizing ready-to-harvest metadata, we created new properties that 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. For example, publication providers such as PubMed typically use the 'author' property in their metadata, while dataset providers such as Figshare and Zenodo are compliant with the DataCite schema and typically prefer 'creator'. Although both properties are valid for their respective https://schema.org classes, we normalized our schema to use 'author' for all six of our classes (Dataset, ClinicalTrial, Analysis, Protocol, Publication, Computational Tool), because we expected the volume of publications to dwarf that of all other classes of resources. We added this schema to the Schema Registry of the Data Discovery Engine (DDE)[27], a project to share and reuse schemas and register datasets according to a particular schema. 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 (Supplementary Fig. 7). First, metadata can be ingested from other resource repositories or collections using the BioThings SDK[28] data plugins. By leveraging the BioThings SDK, we developed a technology stack that addresses the fragmentation issue by easily integrating metadata from different pre-existing resources. For each resource repository or collection, a parser or data plugin enables automated import and updates from that resource. To import the data, the metadata is harvested from the source using API calls (if available), HTML web scraping or .CSV or .TXT tables of metadata. All structured metadata provided by the sources is compiled and mapped to our schema using custom Python scripts. The harmonized metadata is dumped into a JSON output. Supplementary Fig. 8 shows the completeness of each metadata property within our schema, broken down by resource type (data are provided in Supplementary Table 2). Data plugin code for the sources is available at https://github.com/outbreak-info (Code availability).

In the second mechanism, metadata for individually curated resources can be submitted via an online form through the DDE Metadata Registry[27]. 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 (Supplementary Table 3) included generalist open data repositories, biomedical-specific data projects including those recommended by the NIH[29] and the NSF[30] to house open data and individual websites that we came across through search engines and other COVID-19 publications. Prioritizing those resources that had a large number of resources related to COVID-19, we selected an initial set of two to three 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[31]). An analysis resource is defined as a frequently updated, web-based, data visualization, data interpretation and/or data analysis resource.

Creation of the Research Library API and query interface

To accommodate a large number of heterogeneous data sources, each of which is independently harvested, we used the BioThings SDK framework to combine the data sources into a combined, public searchable index (Supplementary Fig. 7). The JSON outputs of our data plugins are ingested by the BioThings framework and merged into an intermediary MongoDB database, and the processed data are indexed in an Elasticsearch index that can be accessed through our public API (api.outbreak.info). The BioThings SDK plugin architecture handles errors in individual parsers without affecting the availability of the API itself. Errors thrown by individual parsers may result in a lack of updates of an individual resource until the error is resolved, but the API will serve the latest version of data from the broken parser and up-to-date data from all functional parsers, which will continue to be updated independently. Using the plugin architecture also allows the creation and maintenance of individual resource parsers to be crowdsourced to anyone with basic Python knowledge 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 that are not located in large repositories. As it is not feasible for one team to locate, identify and collect standardized metadata from these

individual datasets and research outputs, we leveraged the DDE[27] to enable crowdsourcing and citizen science participation in the curation of individual resource metadata.

A Tornado server is used to create an API endpoint, api.outbreak.info/resources, that leverages the search capabilities of Elasticsearch to efficiently query data. Within the search results, Elasticsearch sorts them by relevance based on Lucene's Practical Scoring Function[32], which prioritizes the query normalization factor, coordination factor, term frequency, inverse document frequency and any custom query-boosting fields selected by the user[33]. To adjust this behavior based on common search patterns, we upweighted queries for which the search term occurs in the name field and/or the name of a clinical trial therapeutic intervention (for example, 'remdesivir') with the following parameters: weight of 4 for 'name' and 3 for 'interventions.name'. We continue to monitor common query patterns using our analytics to refine the scoring algorithm to improve the list of results for the user. Within the web interface, the user has the option to sort by the best match-relevance score, update date for the document or alphabetically by name. Within search queries, terms are automatically combined by 'AND'. For instance, the search 'long COVID' will be interpreted as 'long AND COVID'. This search will find resources containing both terms, although not necessarily together; the Elasticsearch default scoring function will first list resources that contain both words together and that frequently mention the terms. Exact phrases can be explicitly declared by encapsulating the terms in quotes (for example, 'long COVID' to search only for the phrase 'long COVID'). Additionally, terms can be combined by the term 'OR' (for example, (Moderna OR Pfizer) AND ('side effects' OR 'adverse effects')). Further details on advanced searching behavior are provided in our guide to the outbreak.info R package at https://outbreak-info.github.io/R-outbreak-info/articles/researchlibrary. html#some-notes-on-constructing-queries. Further optimization will be the subject of future work, based on continuing analysis of analytic patterns for the most common search queries and filters to promote user-driven design. Additional work will also focus on creating an advanced query builder to make it easier to combine terms by any combination of 'AND', 'OR' and 'NOT' and to help the user search for exact phrases.

To update the API with new data provided by the data sources, the BioThings Hub schedules daily updates to pull data upstream and add them to the existing index. The BioThings Hub independently maintains each data source, enabling independence if an individual data source pipeline breaks, and maintains historical data by default, creating automated backups. The code for the server-side application is available at https://github.com/outbreak-info/outbreak.api (https://doi.org/10.5281/zenodo.7343503).

outbreak.info Research Library web application and metadata access

The web application was built using Vue.js, a model–view–viewmodel JavaScript framework that enables the two-way binding of user interface elements and the underlying data allowing the user interface to reflect any changes in underlying data and vice versa. The client-side application uses the high-performance API to interactively perform operations on the database. To iteratively improve the interface, we conducted usability studies as described in Supplementary Fig. 2. The code for the client-side application is available at https://github.com/outbreak-info/outbreak.info (https://doi.org/10.5281/zenodo.7343497).

To enable programmatic access to all our harmonized metadata collection, all data are available in our API (api.outbreak.info) and can be accessed through an R package as described by Gangavarapu et al.[24] (package website, https://outbreak-info.github.io/R-outbreak-info/; code, https://github.com/outbreak-info/R-outbreak-info, https://doi.org/10.5281/zenodo.7343501).

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[28] and the DDE[27] so that individual resource collections can be added by writing BioThings plugins that conform to our schema. Expanding available classes of resources can be easily carried out by extending other classes from https://schema.org via the DDE Schema Playground at https://discovery.biothings.io/schema-playground. Community contributions of resource plugins can be carried out via GitHub. In addition to contributing resource plugins for collections or repositories of metadata, users can enter metadata for individual resources via the automatic guides created by the DDE. 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 DDE's Outbreak Data Portal Guide at https://discovery.biothings.io/guide/outbreak/dataset. Although limited by the original submission form (Google forms), the raw and merged responses illustrating the thoroughness of the submissions from the two volunteers can be found at https://docs.google.com/spreadsheets/d/1q1c400UFIOyXedFf2L81zROVkXi3BWBhU46Ic0cMYsI/edit?usp=sharing. 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. 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[13] with input from the infectious disease research community and volunteer curators. The list consists of 11 broad categories and 24 specific child categories. 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 obtain these classifications from LitCovid, subsetted exports of identifiers were downloaded from LitCovid and then mapped to the metadata records from PubMed.

Whenever possible, sources with thematic categories were mapped to our list of categories 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 using 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 ClinicalTrials metadata, the aforementioned term search iteration and citizen science curation of Zenodo and Figshare datasets. Details on the logical mapping from NCT ClinicalTrials metadata can be found at https://github.com/gtsueng/outbreak_CT_classifier (https://doi.org/10.5281/zenodo.7442988). The keyword mapping from LitCovid can be found at https://github.com/outbreak-info/topic_classifier/tree/main/data/keyword and https://github.com/outbreak-info/topic_classifier/tree/main/data/subtopics/keywords.

While positive categorical data were identified via the aforementioned methods, negative controls were generated by randomly selecting from alternative topics and ensuring no overlap. The categorical data were randomly split into training (80%) and test (20%) sets per test, and five tests were performed per topic by applying out-of-the-box logistic regression and multinomial naive Bayes and random forest algorithms from scikit-learn. These three algorithms were found to perform best on this binary classification task using out-of-the-box tests. Topics were only added to the record if all three methods agreed on the classification. The set size and test results using default tests from scikit-learn for each algorithm for each topic and subtopic for each of the five test runs can be found at https://github.com/outbreak-info/topic_classifier/blob/main/results/in_depth_classifier_test.tsv.

The efforts of our two volunteers suggested that non-experts were capable of thematically categorizing datasets; therefore, 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 five topics by at least three different citizen scientists to ensure quality of the results. Citizen scientists were asked to prioritize specific topic categories over broader ones. Ninety citizen scientists recruited via either participation in the Mark2Cure project[34] or a Scripps Research summer program participated in classifying 530 datasets pulled from Figshare and Zenodo, increasing the likelihood of quality submissions and decreasing the likelihood of abuse and false information. 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. The citizen science classifications can be found at https://github.com/outbreak-info/topic_classifier/blob/main/data/subtopics/curated_training_df.pickle. To evaluate the quality of the citizen scientist classifications, we first filtered classifications where at least two or three of three to five curators agreed on the topic category. We then compared the results of their classification with predictions by an out-of-the-box algorithm that was trained on LitCovid-classified abstracts. A total of 186 of 530 classifications did not agree and were manually inspected; only about 10% of the categorization (54) was worse with citizen scientists over the predictions, and, in many cases, the curators provided more precise categorization. Full details of the evaluation are available at https://github.com/gtsueng/curate_outbreak_data (https://doi.org/10.5281/zenodo.7442949). 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-of-the-box logistic regression and multinomial naive Bayes and random forest algorithms available from scikit-learn. The topic classifier can be found at https://github.com/outbreak-info/topic_classifier (https://doi.org/10.5281/zenodo.7439573).

In addition to community curation of topic categorizations, we identified a citizen science effort, the COVID-19 Literature Surveillance Team (COVID-19 LST), that was evaluating the quality of COVID-19 related literature. The COVID-19 LST consists of medical students (many of which were in their third or fourth year), practitioners and researchers who evaluate publications on COVID-19 based on the Oxford Levels of Evidence criteria and write bottom line, up front summaries[20]. With their permission, we integrated their outputs (daily reports or summaries and Levels of Evidence evaluations) into our collection. Although the project has since ended, the valuable work by this team was integrated without further evaluation due to their background and training.

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 and/or text and authors for preprint (bioRxiv or medRxiv)[35] versus LitCovid publications. We identified thresholds with high precision and low sensitivity and binned the matches into two groups: matched preprint or peer-reviewed publication versus 'needs review'. We also leveraged NLM's pilot preprint program to identify and incorporate additional matches. The code used for the preprint matching and the .XLSX file detailing the semi-automated and manual inspection of a sample of 1,500 matches from the results can be found at https://github.com/outbreak-info/outbreak_preprint_matcher (https://doi.org/10.5281/zenodo.7439581). Briefly, a subsample of 1,500 preprint or peer-reviewed matches were inspected and confirmed to match via the preprint listed within the PubMed record in the correction field (1,158 matches); manual inspection of preprint records, which listed the peer-reviewed publication (290 matches); and manual inspection of preprint and the corresponding PubMed record and publication content (52 matches). The inspection confirmed that our threshold cutoff for preprint matching ensured the inclusion of a limited number of the most accurate matches at the cost of many more potential but lower-quality matches. Expected matches were linked via the correction property in our schema.

Case study on variant research

To identify research about variants, we used the keyword phrase 'variant OR lineage' in the Research Library and within the R package outbreakinfo. For Fig. 2a, resources were counted by @type (Publication, Dataset, ComputationalTool, ClinicalTrial, Protocol, Analysis). The number of resources was aggregated to the weekly level by the date of the latest update and normalized to all resources within the Library for that week, creating a proportion of the Library for that week (Fig. 2c). For variant-specific queries, the WHO-designated name was combined with its Pango lineage[36] plus all descendants, as specified by the Pango team in October 2022 (https://raw.githubusercontent.com/cov-lineages/lineages-website/master/data/lineages.yml). To decrease the likelihood of a spurious hit for the resource (for instance, a publication mentioning Alpha in the description but focusing only on Omicron), we used fielded queries to only search

by the name of the resource. For instance, for Gamma, the following query was used: name:Gamma OR name:'P.1' OR name:'P.1.2'. Code to replicate the analysis and visualizations is available at https://github.com/outbreak-info/outbreak-resources-paper/blob/main/Figure%204%20-%20Variant%20analysis.R.

Harmonization and integration of resources and genomic data

The integration of genomic data from GISAID is discussed by Gangavarapu et al.[24]. We built separate API endpoints for our resources (metadata resource API) and genomics (genomic data API) using the BioThings SDK[28]. Data are available via our API at http://api.outbreak.info and through our R package as described by Gangavarapu et al.[24].

Limitations

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 and metadata are currently collected and shared. Researchers have embraced preprints, but resources (especially datasets and computational tools) needed to replicate and extend research results are not linked in ways that are discoverable. Although many journals and funders have embraced dataset and source code submission requirements, the result is that the publication of datasets and software code is 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 preprint and data sharing reflects the existing incentive structure, in which 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. As with many web-based, open-source resource sites, bugs and browser-compatibility issues may arise without notice for less-popular browsers. Users can bring these issues to our attention by submitting them to our issue tracker on GitHub (https://github.com/outbreak-info/outbreak.info/issues).

Comparison of the outbreak.info Research Library with other resources

To illustrate how our resource fits into the COVID-19 resource landscape, we compare features from our Research Library with other COVID-19 multisource aggregation efforts (Supplementary Table 4) and provide a list of terms and features in Supplementary Table 5. We provide the most commonly searched sources (that is, filter by source) and resource types (that is, filter by resource type) (Supplementary Table 1a). Usage statistics for record views and filtering by source are available in Supplementary Table 1b. Filtering was the most popular feature added to the Library, with over a quarter of all queries using some sort of filtering (Supplementary Table 1c). Users were most likely to filter results by resource type, followed by keywords and source.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All metadata harvested and harmonized in the outbreak.info Research Library is freely available through an API (http://api.outbreak.info/) and in an associated R package (https://outbreak-info.github.io/R-outbreak-info/).

Code availability

All code used to generate the outbreak.info Research Library is freely available on GitHub (https://github.com/outbreak-info) under open-source licenses. The outbreak.info web application is available at https://github.com/outbreak-info/outbreak.info (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7343497). The outbreak.info R package to access all the genomics and epidemiology data and Research Library metadata compiled and standardized on outbreak.info is available at https://github.com/outbreak-info/R-outbreak-info (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7343501). The code to create the API (https://api.outbreak.info) to access Research Library metadata and case and death data is available at https://github.com/outbreak-info/outbreak.api (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7343503). The harvester of bioRxiv and medRxiv preprint publications is available at https://github.com/outbreak-info/biorxiv (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7439483). The harvester of clinical trials from https://clinicaltrials.gov is available at https://github.com/outbreak-info/clinical_trials (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7439505). The harvester of COVID-19 LST level of evidence ratings is available at https://github.com/outbreak-info/covid19_LST_reports (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7439527). The COVID-19 LST annotations code is available at https://github.com/outbreak-info/covid19_LST_annotations (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7439515).The COVID-19 LST report data are available at https://github.com/outbreak-info/covid19_LST_report_data (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7439521). The harvester for manually curated metadata from the DDE is available at https://github.com/biothings/discovery-app/blob/master/scripts/outbreak.py) (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7439590). The harvester from Figshare COVID-19 is available at https://github.com/outbreak-info/covid_figshare (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7439543). The harvester for COVID-19 collection of Harvard Dataverse is available at https://github.com/outbreak-info/dataverses (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7439563). The harvester for analyses by Imperial College London is available at https://github.com/outbreak-info/covid_imperial_college (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7439545). The LitCovid publication harvester is available at https://github.com/outbreak-info/litcovid (version of

the code used in this paper is available at https://doi.org/10.5281/zenodo.7439565). The harvester of metadata for SARS-CoV-2 structures from the Protein Data Bank is available at https://github.com/outbreak-info/covid_pdb_datasets (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7439549). The harvester of protocol metadata from protocols.io is available at https://github.com/outbreak-info/protocolsio (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7439579). The harvester of clinical trials from WHO ICTR is available at https://github.com/outbreak-info/covid_who_clinical_trials/blob/master/parser.py (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7439553). The reusable Research Library schemas for publications, datasets, clinical trials, protocols and analyses and associated data mappings are available at https://github.com/outbreak-info/outbreak.info-resources (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7439569). The reusable Research Library tools for parsers are available at https://github.com/outbreak-info/outbreak_parser_tools (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7439577). The code to look up Altmetric ratings for outbreak.info resources is available at https://github.com/outbreak-info/covid_altmetrics (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7439533). The code to match preprints to their peer-reviewed publications is available at https://github.com/outbreak-info/outbreak_preprint_matcher (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7439581). The machine learning topic classification of categories within the Research Library is available at https://github.com/outbreak-info/topic_classifier (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7439573). The mapping logic used to classify clinical trial records using clinical trial-specific metadata is available at https://github.com/gtsueng/outbreak_CT_classifier (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7442988). The evaluation of citizen scientist efforts is available at https://github.com/gtsueng/curate_outbreak_data (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7442949). The code to generate the figures within this text, including for the case study, is available at https://github.com/outbreak-info/outbreak-resources-paper (version of the code used in this paper is available at https://doi.org/10.5281/zenodo.7439567).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank J. Rah, B.J. Enright, J. Doroshenko, T. Nishath and the rest of the COVID-19 LST for allowing us to share their work. We thank T. Adams and C. Lazarchick for their work in identifying metadata from various individual datasets and their extensive feedback. We thank S. Andarmani for her suggestions and feedback on dataset categories. We thank all Outbreak Curators contributors found at https://blog.outbreak.info/dataset-topic-category-contributors for taking the time to categorize datasets. We thank S. Ul-Hasan for their feedback on the R package. We thank D. Valentine for sharing details about his netlify app as part of the RADx-Rad Data Coordination Center, which is funded by the NIH (U24LM013755). Work on outbreak.info was supported by the National Institute for Allergy and Infectious Diseases (5 U19 AI135995: G.T., J.L.M., M.A., M.C., E. Haag, A.A.L., E. Hufbauer, M.Z., K.G.A., C.W., A.I.S., K.G., L.D.H.; 3 U19 AI135995-04S3: G.T., J.L.M., E. Haag, E. Hufbauer, K.G.A., C.W., A.I.S., K.G., L.D.H.; 3 U19 AI135995-03S2: G.T., J.L.M., E. Haag, E. Hufbauer, K.G.A., C.W., A.I.S., K.G., L.D.H.; 75N91019D00024: G.T., E. Haag, J.L., D.J.W., C.W., A.I.S., L.D.H.), the

National Center for Advancing Translational Sciences (5 U24 TR002306: G.T., J.L.M., M.C., C.W., A.I.S., L.D.H.), the Centers for Disease Control and Prevention (75D30120C09795: M.A., A.A.L., M.Z., K.G.A., K.G.) and the National Institute of General Medical Sciences (R01GM083924: G.T., M.C., X.Z., Z.Q., C.W., A.I.S.).

References

1. Novel Coronavirus (2019-nCoV): Situation Report, 1 (WHO, 2020); https://apps.who.int/iris/handle/10665/330760
2. Dong E.et al. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect. Dis 20, 533–534 (2020). [PubMed: 32087114]
3. Kaiser J.'Every day is a new surprise.' Inside the effort to produce the world's most popular coronavirus tracker. Science 10.1126/science.abc1085 (2020).
4. Noren LE et al. Institutional Response to COVID https://docs.google.com/spreadsheets/d/1IbF_wlmldVssG5spcmNE82nR9btcbF7rUlEqtcXW03o/edit#gid=0 (2020).
5. Morris A.& citizen scientists. USA COVID-19 K-12 School Closures, Quarantines, and/or Deaths https://docs.google.com/spreadsheets/d/e/2PACX-1vQSD9mm5HTXhxAiHabZA6BPUByWBlP5HZ2jfOPEeGZkMB0ZFsmFBL5orqjIq22mjFNZ7n-11ObCylGn/pubhtml?fbclid=IwAR2tJ8yDVehGpxoP97Cco5HYAxoN014opwwm6uYt4s3E2xDr_8u9KF_LlgI#(2020).
6. James P.& citizen scientists. Staying home club. GitHub https://github.com/phildini/stayinghomeclub (2020).
7. Pogkas D.et al. The airlines halting flights as virus outbreak spreads. Bloomberg https://www.bloomberg.com/graphics/2020-china-coronavirus-airlines-business-effects/ (2020).
8. Joachimiak M.et al. SARS-COV-2 and COVID-19 Datasets https://docs.google.com/spreadsheets/d/1eMhot7MjusyM7_2IBnzqi7RlzWWoYnfheWhMgDlPToQ/edit#gid=0(2020).
9. Skenderi J.et al. COVID-19 Resource Library https://docs.google.com/spreadsheets/u/2/d/1cqxDAg4jMHXI6gHOnoV8HqDdRHnmxEJRl-bhhpe1HEo/htmlview#(2020).
10. Navarro C.& Capdarest-Arest N.COVID-19 Open Dataset Sources https://docs.google.com/spreadsheets/d/10t3vtULr3nTz7mrlKj0rldUys47wsIfOVReHnx3Xu18/edit#gid=0(2020).
11. NIH OPA. iSearch COVID-19 Portfolio (NIH, 2020); https://icite.od.nih.gov/covid19/search
12. Allen Institute for AI. COVID-19 Open Research Dataset Challenge (CORD-19). Kaggle https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge (2020).
13. Chen Q.et al. LitCovid: an open database of COVID-19 literature. Nucleic Acids Res. 49, D1534–D1540 (2020).
14. ClinicalTrials.gov. Protocol Record Schema—XML Schema for Electronic Transfer of Protocol Information into the ClinicalTrials.gov Protocol Registration System (National Library of Medicine, 2018) https://prsinfo.clinicaltrials.gov/ProtocolRecordSchema.xsd
15. Fava I.et al. Coronavirus disease research community—COVID-19. Zenodo https://zenodo.org/communities/covid19/?page=1&size=20 (2020).
16. Hyndman A.A Figshare COVID-19 research publishing portal. Figshare https://figshare.com/blog/A_Figshare_COVID-19_Research_Publishing_Portal/558 (2020).
17. European Organization for Nuclear Research. Zenodo FAIR principles. Zenodo https://about.zenodo.org/principles/(2013).
18. Hahnel M.What Google dataset search means for academia. Figshare https://figshare.com/blog/What_Google_Dataset_Search_means_for_academia/422 (2018).
19. Birkin LJ et al. Citizen science in the time of COVID-19. Thorax 76, 636–637 (2021). [PubMed: 33653934]
20. Rah J.et al. COVID-19 Literature Surveillance Team. Internet Archive https://web.archive.org/web/20211020140102;https://www.covid19lst.org/copy-of-about (2020).
21. Tsueng G.et al. Applying citizen science to gene, drug and disease relationship extraction from biomedical abstracts. Bioinformatics 36, 1226–1233 (2020). [PubMed: 31504205]
22. Blickhan S.et al. Transforming research (and public engagement) through citizen science. Proc. Int. Astron. Union 14, 518–523 (2018).

23. Science Digital. About us. Altmetric https://www.altmetric.com/about-us/(2022).
24. Gangavarapu K.et al. outbreak.info: real-time surveillance of SARS-CoV-2 mutations and variants. Nat. Methods 10.1038/s41592-023-01769-3 (2023).
25. Haag E.User stories outbreak.info blog. Sulab https://blog.outbreak.info/?tag=user_stories (2022).
26. Valentine D.& RADx. SearchOutbreak. Radical data coordination center. Netlify https://searchoutbreak.netlify.app (2021).
27. Cano M.et al. Schema Playground: a tool for authoring, extending, and using metadata schemas to improve FAIRness of biomedical data. Preprint at bioRxiv 10.1101/2021.09.02.458726(2021).
28. Lelong S.et al. BioThings SDK: a toolkit for building high-performance data APIs in biomedical research. Bioinformatics 38, 2077–2079 (2021).
29. BioMedical Informatics Coordinating Committee. Data Sharing Resources (NIH, 2020) https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html
30. Open Data at NSF (National Science Foundation, 2013); https://www.nsf.gov/data/
31. Imperial College COVID-19 Response Team. ONS Excess Deaths (Imperial College London, 2021); http://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/covid-19-reports/
32. Controlling Relevance. Elasticsearch: the Definitive Guide [2.x] (Elasticsearch B.V., 2023) https://www.elastic.co/guide/en/elasticsearch/guide/current/controlling-relevance.html
33. Lucene's Practical Scoring Function. Elasticsearch: the Definitive Guide [2.x] (Elasticsearch B.V., 2023); https://www.elastic.co/guide/en/elasticsearch/guide/current/practical-scoring-function. html
34. Tsueng G.et al. Citizen science for mining the biomedical literature. Citiz. Sci 1, 14 (2016). [PubMed: 30416754]
35. COVID-19 SARS-CoV-2 (medRxiv and bioRxiv, 2021); https://connect.biorxiv.org/relate/content/181
36. Rambaut A.et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol 5, 1403–1407 (2020). [PubMed: 32669681]

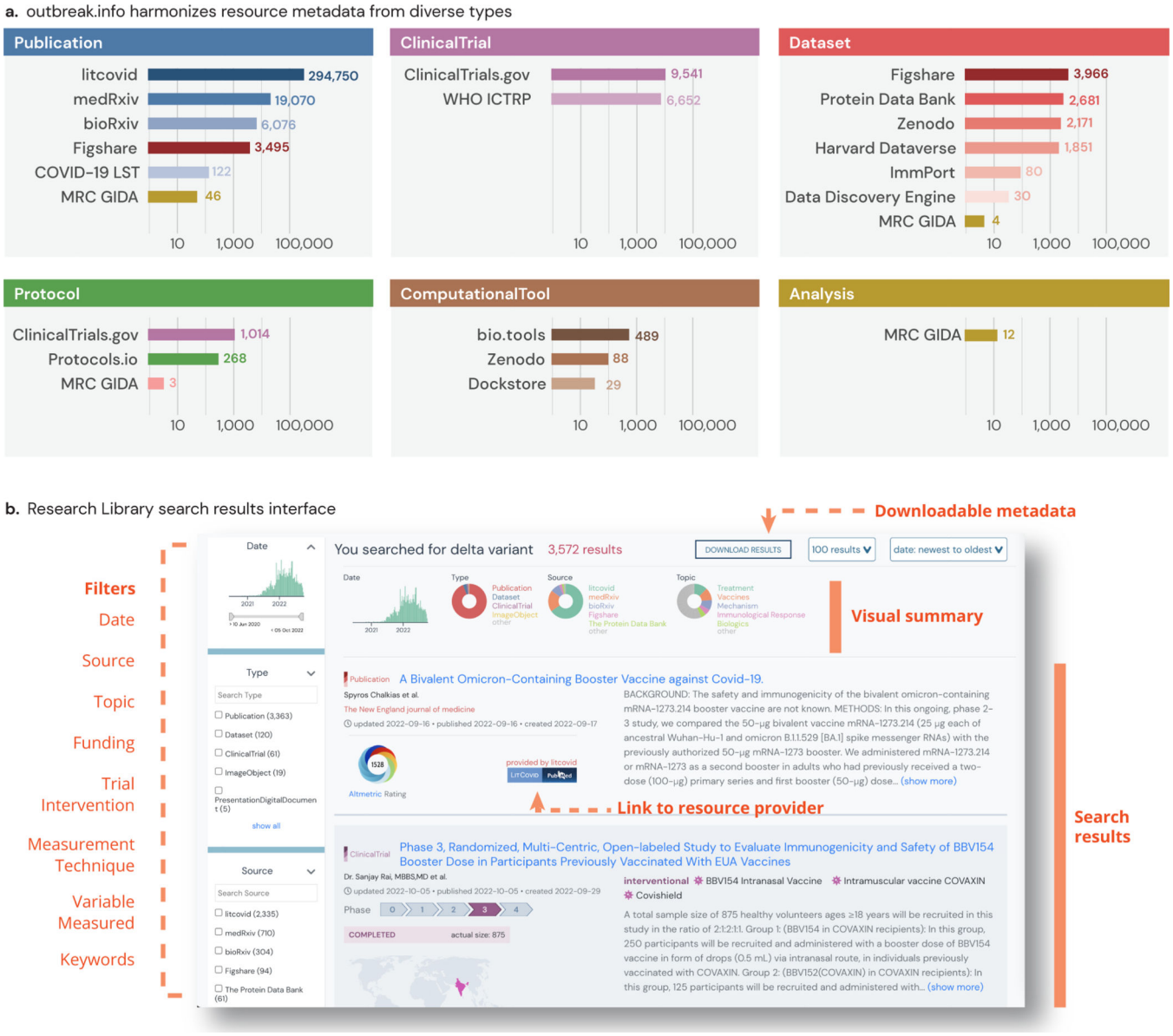

Fig. 1 |. outbreak.info centralizes and standardizes resources into a single searchable index.

a, Distribution of resources by resource type and source. ICTRP, International Clinical Trials Registry Platform; MRC GIDA, MRC Centre for Global Infectious Disease Analysis. **b**, Searching for ‘Delta variant’ finds heterogeneous resources.

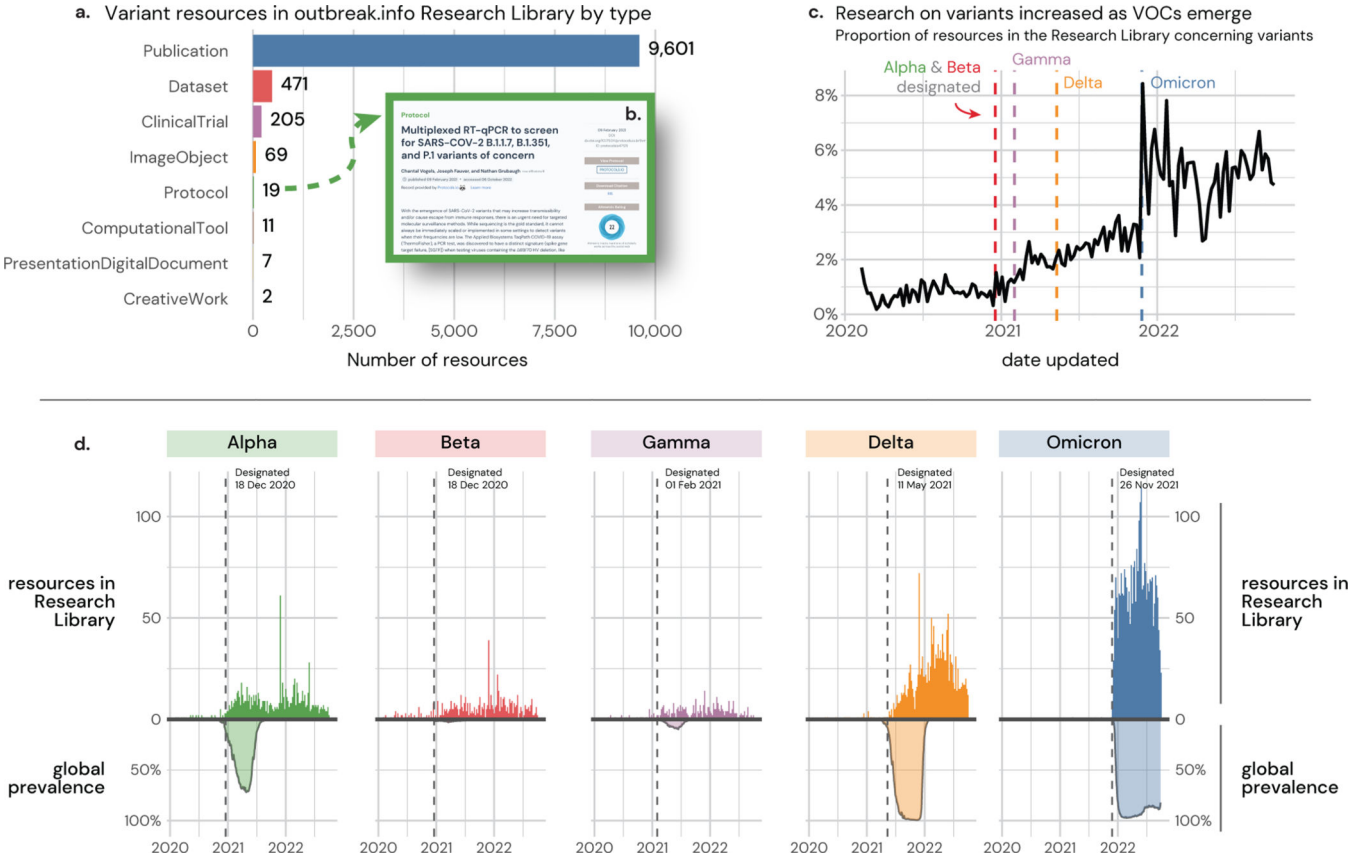

Fig. 2 |. Resources concerning variants within the outbreak.info Research Library.

a, Standardized searching across resource types, including Publications, Datasets, ClinicalTrials and more. **b**, A variant protocol discovered within the Library. **c**, As VOCs were designated, the proportion of research in the Library focused on variants increased. **d**, The increase in research on each VOC mirrored its worldwide prevalence, with research on the transmissibility, virulence and/ or immune evasion supporting their VOC designation by public health agencies, and these designations encouraging further research.