nature methods

Brief Communication

https://doi.org/10.1038/s41592-023-01770-w

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

In the format provided by the authors and unedited

What are resources?

- Websites, dashboards, portals
- Research outputs (Publications, Datasets, etc.)
- Lists and collections of other resources
- Guides, policies, recommend-dations
- Tools and code
- Other (learning materials, videos, etc.)

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

Why are resources hard to find and use?

- Volume: There is too much, too quickly
- Fragmentation: There is no central place for all of it
- Variety: There are many different types
- Standardization: Differences exist even within a type

How can we support the use of standardized, centralized resources?

- Value-Added Metadata
- APIs
- User-friendly GUI
- Interpretable Visualizations
- Shareable Results

Supplemental 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?

Usability Studies and Protocols

A. Usability Study Results

Date	n	Task Number	SCR	EFR	TPT	EOU	LTR
Nov 23-25, 2020	5	1	100%	60%	~1 min	96%	Highly likely
		2	100%	0%	~1 min	100%	Highly likely
Mar 4 - Apr 2 , 2021	5	3	100%	100%	40 sec	100%	Highly likely

SCR- Average Successful Completion Rate for the task, EFR - Average Error-Free Rate, TPT- Average Time Per Task, EoU- Average Ease of Use Rating, LTR - Average Likelihood To Return Rate

B. Usability Study Protocol

Audience: Biomedical researchers

Purpose: Test the ease by which biomedical researchers could use outbreak.info to access or visualize data and find up-to-date publications.

Format: Moderated usability study conducted via Zoom. Users were asked to perform one or more of the following tasks:

1. Please use the Research Library to find new datasets related to the COVID-19 reproduction number in the United States.
2. Please use the Research Library to find publications related to cytokine response storm in COVID-19 patients.
3. Please find all publications related to B.1.1.7.

Post-test questions asked:

- Overall, please rate how easy or difficult it is to use outbreak.info on a scale of 1-5, where 1 is very difficult and 5 is very easy.
- How likely are you to continue using outbreak.info to regularly access data or resources, on a scale of 1-5 where 1 is very unlikely and 5 is very likely?
- What do you like most about the outbreak.info app?
- What would you improve about the app? Or what would you add to the app?
- How would you compare outbreak.info to other sites you've used to access data & resources?

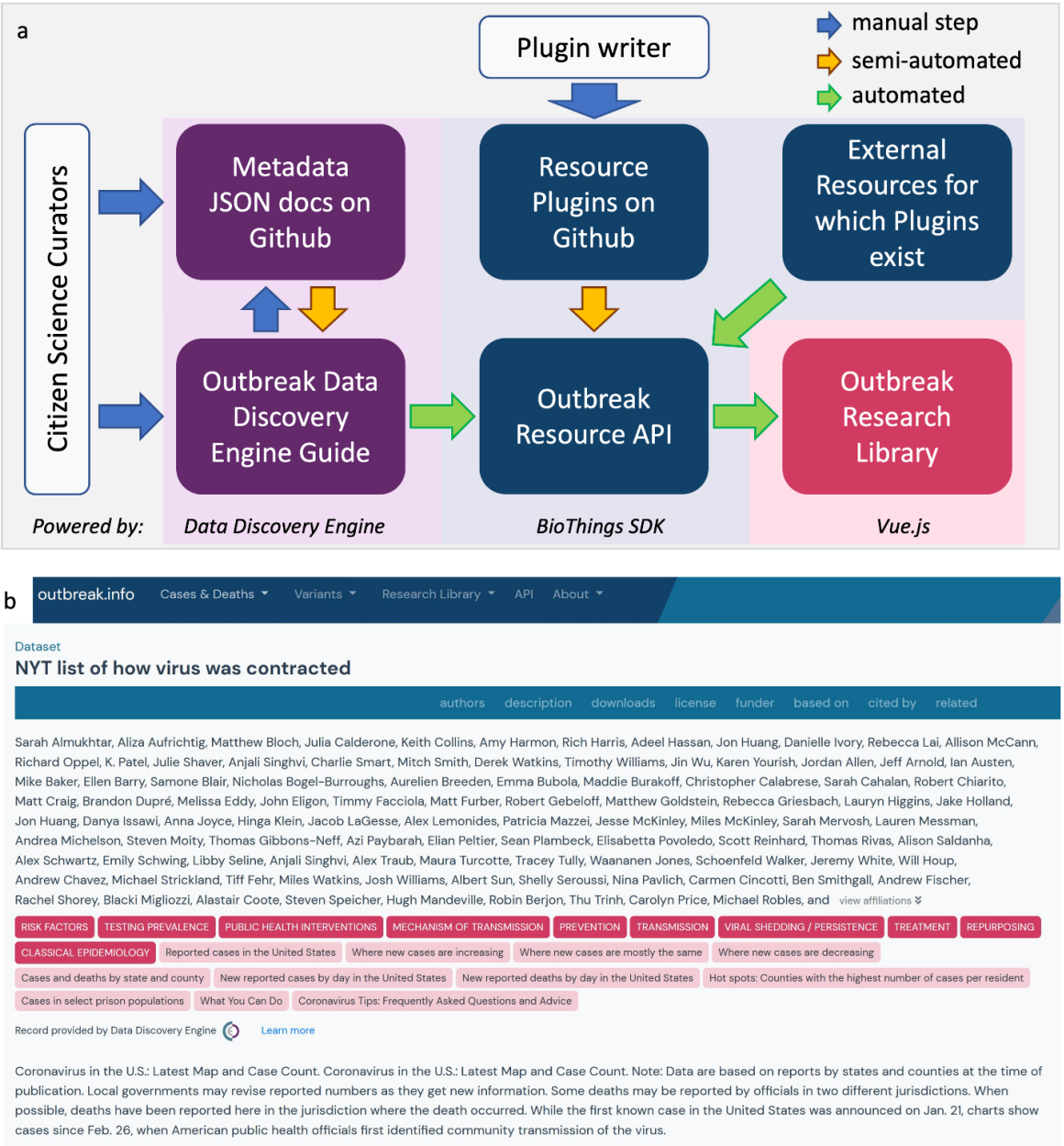
Metrics/Measurements recorded:

- Ease and satisfaction about each task (5-point Likert scale),
- Time on each task,
- Number of successful task completions and errors,
- Overall ease and satisfaction (5-point Likert scale),
- Likelihood to use (5-point Likert scale),
- Suggestions for improvement (likes, dislikes, recommendations),
- Error-free rate,
- And noted observations about the users' process.

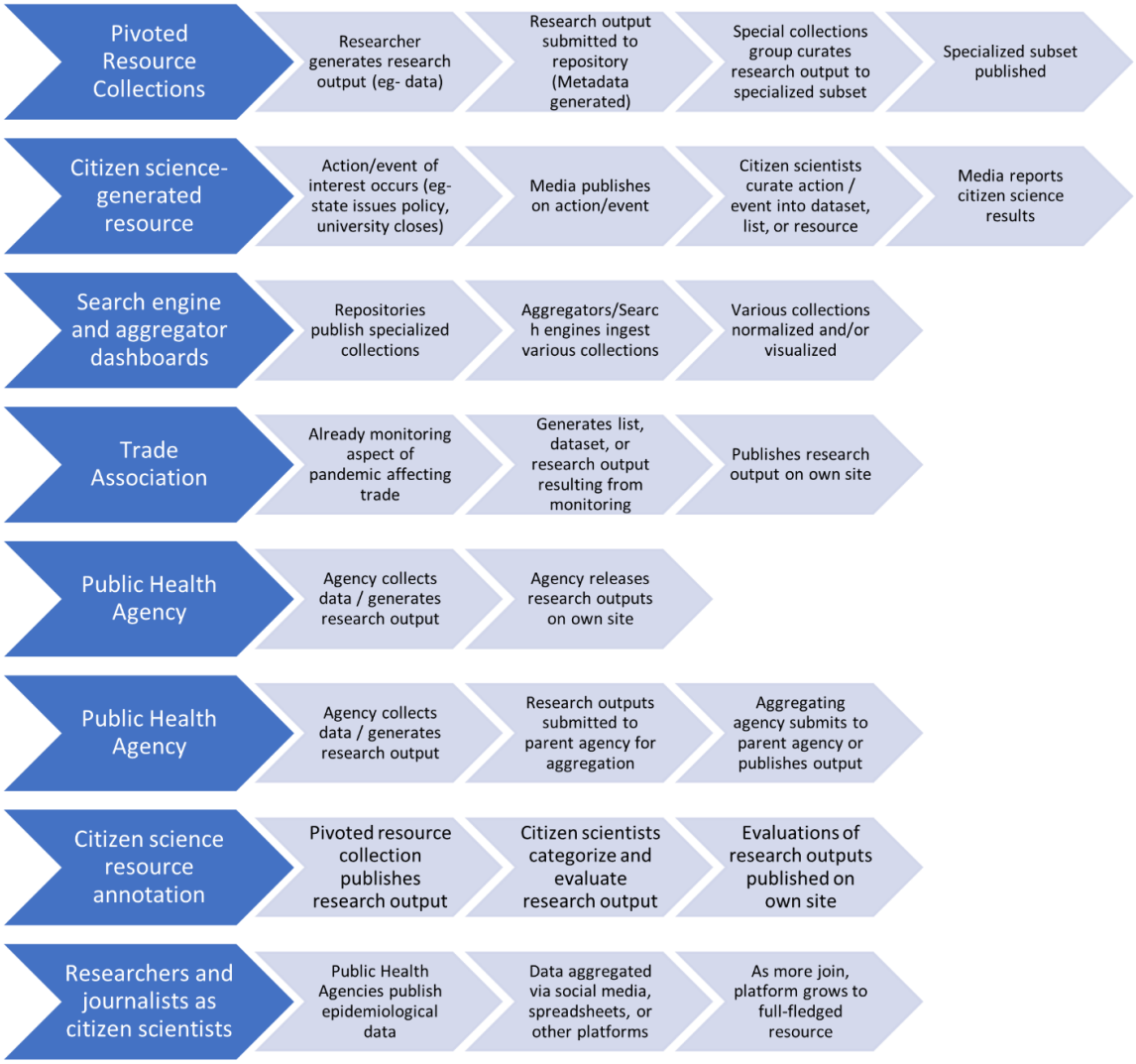
C. Changes ensure usability of Research Library

Usability Study	Changes Implemented
Nov 23-25, 2020	• Redesigned homepage to present outbreak.info's main features as three actionable categories near the top • Added an introductory video • Revised how other pages are linked from the homepage • Collapsed many of the site's details into expandable cards • Renamed "Resources" to "Research Library."
Mar 4 - Apr 2 , 2021	• No changes were made to the Research Library

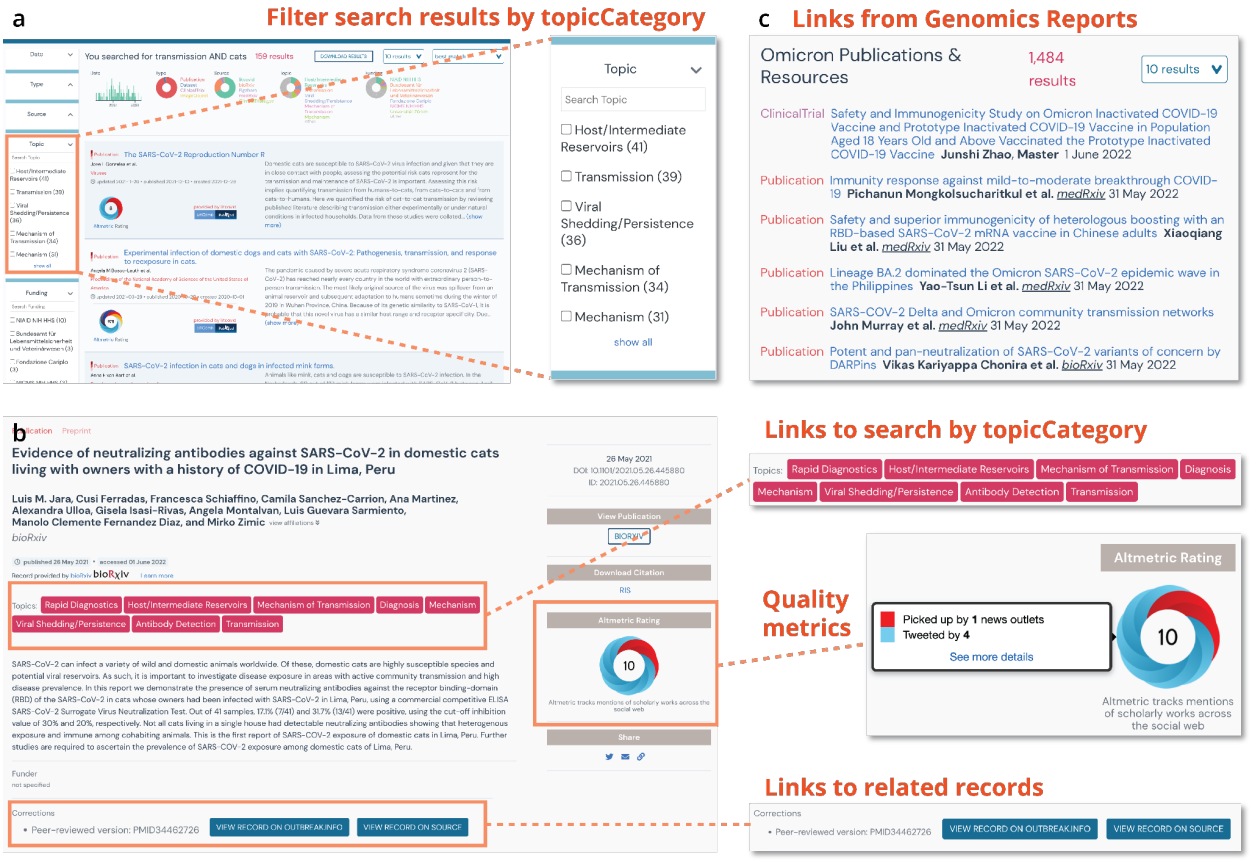
Supplemental Figure 2. Usability studies for iterative design improvements of the Outbreak.info Research Library. **A**. Remote moderated usability test of outbreak.info was conducted between November 23-25, 2020 and between March 4, 2021 and April 2, 2021 over Zoom. Five biomedical researchers participated in each round of tests which consisted of asking users to perform the tasks (and recording metrics) in accordance with protocol (**B**). There was a successful completion rate of 100% for both tasks and zero critical errors, but some non-critical errors resulted in an error-free rate of 60% for the first task and 0% for the second task. The average amount of time spent on each task was around 1 minute. **C.** Improvements to the site and Research Library were made with accordance to the results of the usability studies in November of 2020, but follow-up usability studies in March-April of 2021, indicated no additional changes were needed in order to improve the user performance on the tested tasks.



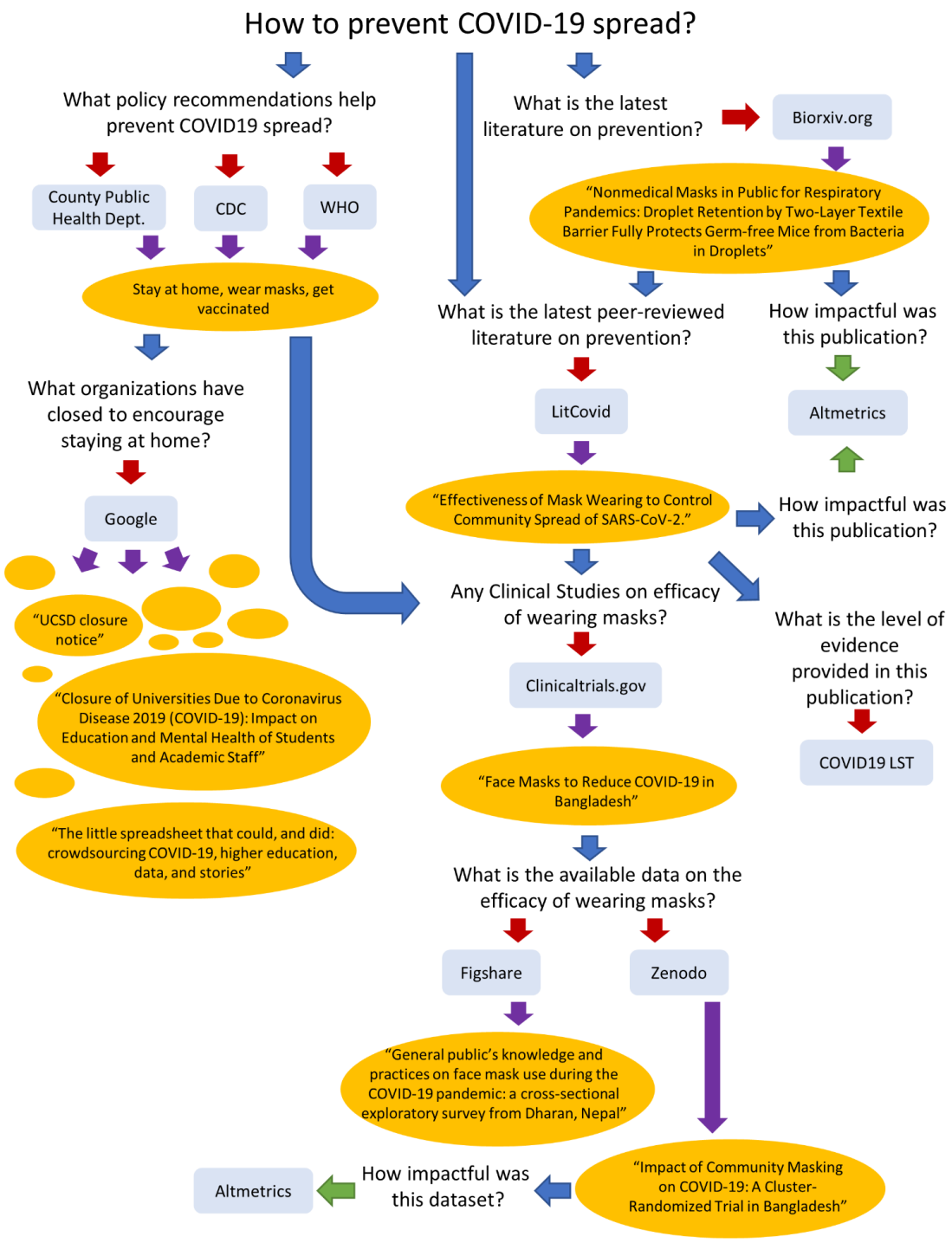
Supplemental Figure 3. Aggregating resource metadata by leveraging community contributions. **a,** The community contribution pipeline and technology stack for outbreak.info's Research Library. Curators may submit dataset metadata using the DDE built-in guide or from GitHub via the DDE/BioThings SDK. 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.



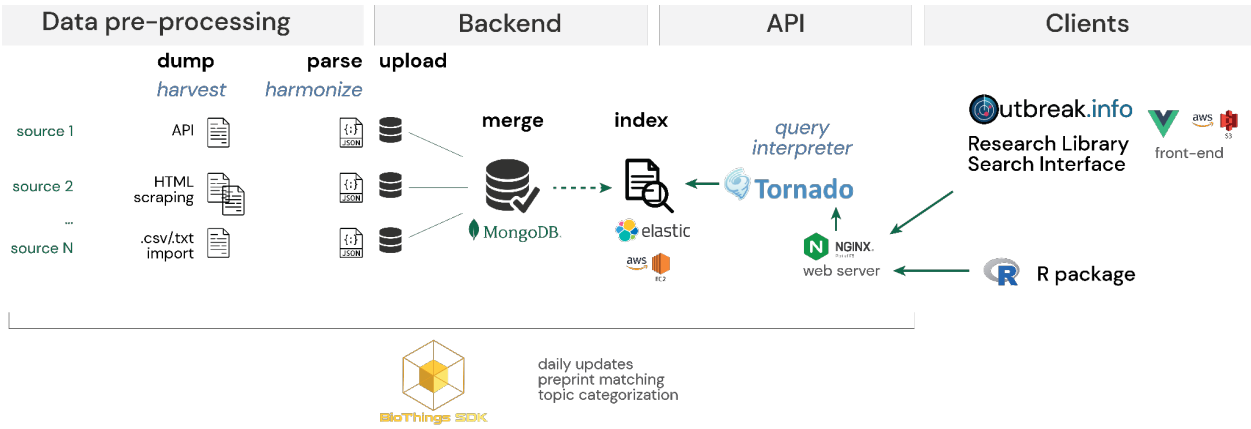
Supplemental Figure 4. Example resource-generation pipelines observed within the first month of the pandemic, resulting in a highly fragmented, non-standardized resource landscape that is difficult to search, centralize and leverage.



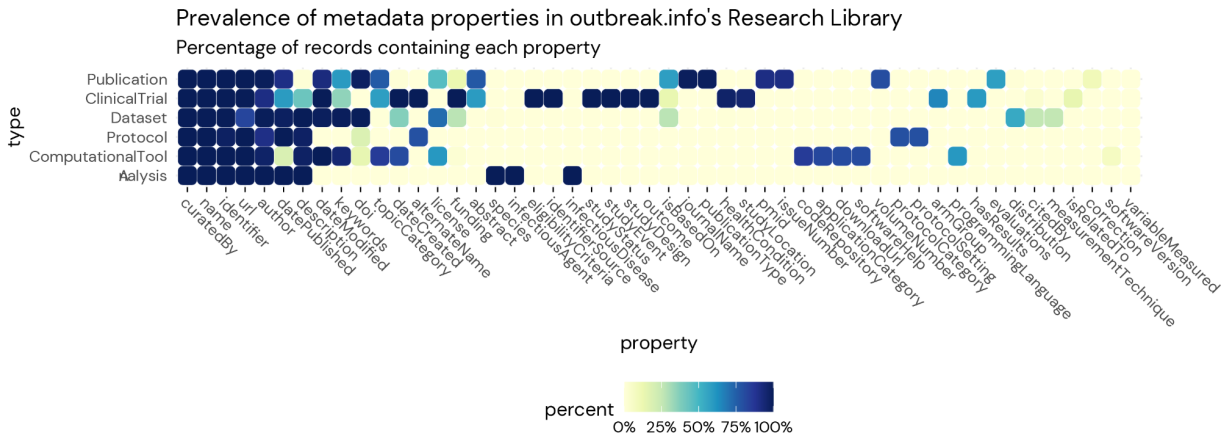
Supplemental Figure 5. Enabling exploration of the resources. **a,** Selectable options for filtering results by topic category or other facets enhance searchability and exploration from the search results view. **b,** Links to other records or to additional potential searches of interest enabling further exploration from a record view. **c,** Links from the Omicron Variant report to related resources.



Supplemental Figure 6. Example resource exploration pathway for addressing the question, "How to prevent COVID-19 spread" early on in the pandemic. The question leads to an iterative process of refining the question (unbound text), searching a resource/website (blue rounded box), finding a result of interest (orange oval). Blue arrows indicate a question refinement step. Purple arrows indicate a result review step (*i.e.*- identifying a potential result of interest). Red arrows indicate a manual search and filter step. Green arrows indicate linked search (i.e.- manual input to a resource not needed for result).



Supplemental Figure 7. Systems architecture for the outbreak.info Research Library metadata infrastructure. Individual data sources are harvested either through API-based calls, reading a tabular file or scraping-based HTML crawling, harmonized to our schemas through custom parser scripts, and uploaded, merged, and indexed to an Elasticsearch index using BioThings SDK. Since each source is harvested and harmonized independently, errors in a given data source do not affect updates in the other sources or the overall status of the API. Queries from the Research Library front-end or the R package are routed through a Tornado-based interpreter which accesses the underlying Elasticsearch index to return results from the API call.



Supplemental Figure 8. Prevalence of each primary metadata field in our schemas by resource type. Raw data is provided in **Supplemental Table 2**.