ACHSM Asia-Pacific Journal of Health Management

RESEARCH ARTICLE

DATA LICENSING FOR PUBLIC INTEREST: A RETROSPECTIVE ANALYSIS OF THE COVID-19 OPEN DATASET

Wahltinez, O, Rincón, M*

Universidad Nacional de Educación a Distancia, Departamento de Inteligencia Artificial (UNED), Madrid, Spain

*Correspondence: owa@google.com

ABSTRACT

The COVID-19 pandemic highlighted the critical need for accessible public health data. However, navigating the complexities of data licensing remains a major challenge, complicating the aggregation, analysis, and sharing of such data. These difficulties are further compounded by diverse legal systems, each with its own copyright and intellectual property rules.

This study examines the licensing terms of 301 datasets within the COVID-19 Open Data metadataset, revealing key barriers to public health data accessibility. To address these challenges, we propose a decision flow diagram to guide users through license compatibility and republishing permissions, enabling ethical and legal aggregation under permissive licenses.

We also present practical guidelines for data producers in public health, aimed at improving the clarity and usability of shared data. These recommendations reflect common licensing issues and the needs of data users, promoting more effective data sharing.

By addressing these barriers, this work offers strategies to enhance the availability and impact of public health data, supporting better responses to current and future global health emergencies.

KEYWORDS

COVID-19, open dataset, data licensing

INTRODUCTION

The COVID-19 pandemic demonstrated the vital importance of data in managing public health crises [1]. In response, the COVID-19 Open Data metadataset [2] was created as a comprehensive repository, consolidating a wide array of COVID-19-related data. This dataset includes information on infection rates, mortality statistics, and various covariates such as mobility trends, weather conditions, non-pharmaceutical interventions, and vaccination progress. The initiative aimed to make this data functional for diverse stakeholders, including researchers, policymakers, and data scientists, fostering a data-driven approach to understanding the pandemic's complexities and informing decisions [3].

Distinguished by its scale and accessibility, the project aggregated data from 301 sources with frequent updates, providing tools like an application programming interface (API) and interactive visualizations to democratize data use. These resources enabled broad applications in research, policymaking, and analytics, supporting advancements in public

health as a science [4]. Data sources were selected based on public availability and minimal usage restrictions, typically requiring only attribution [5,6].

However, assembling such a comprehensive dataset revealed significant challenges, particularly regarding the legal and licensing frameworks governing data use [7]. The diversity of sources, each subject to different licensing agreements, created barriers to integration and dissemination [8]. These issues extend beyond the project itself, highlighting broader challenges in sharing and utilizing public health data during crises [9]. By addressing this often-overlooked issue, this introduction provides a foundation for critically examining the licensing landscape, which holds great potential for advancing pandemic responses yet remains fraught with legal complexities [10].

Data's journey from collection to end-use involves many stakeholders, each playing a pivotal role [11]. Data is initially collected by healthcare professionals or government personnel [12,13] and then curated and shared by health authorities [14,15] or data-sharing entities [16], sometimes through press releases [17]. The value of this data is enhanced when aggregated with other datasets or reformatted for broader accessibility, tasks often handled by intermediaries [18]. Ultimately, researchers, analysts, and policymakers use this data for analysis, visualization, and decision-making [19].

Government and health authority-led 'open data' portals represent a significant step toward democratizing data access. These platforms, ranging from local [20,21] to national authorities [22,23], promote transparency and informed decision-making. Yet, licensing variability—ranging from open licenses to restrictive terms or no license at all—creates challenges for potential users. Licensing ambiguities particularly affect entities like news organizations and researchers intending to republish data, forcing them to navigate the complexities of fair use [24,25], which varies across jurisdictions and depends on the relationship between the source, intermediary, and derived outputs. Recent controversies in artificial intelligence development have brought such issues into sharper focus [26].

The inconsistency in licensing practices hampers data accessibility, complicates compliance for users, and stifles innovation by limiting data integration from diverse sources [27]. Addressing these challenges is essential to fostering an environment where open data fulfills its role in advancing public health, research, and policy [28]. Despite the critical need for accessible public health data underscored by the pandemic, there remains a lack of comprehensive frameworks or best practices for data producers. This gap results in inconsistent access across jurisdictions, limiting usability for both individuals with specialized accessibility needs, such as those relying on screen readers, and automated tools for aggregating datasets [34]. This inconsistency restricts opportunities for large-scale data applications, particularly in machine learning [35].

Selecting appropriate licenses is another challenge for data publishers, given the numerous options [29]. Some publishers opt for custom licenses to address specific needs, adding complexity and necessitating legal expertise [30]. Custom terms can create interpretative challenges, particularly around concepts like derivative works. The classification of a dataset as derivative works, pivotal in copyright law, depends on whether a dataset utilization constitutes significant incorporation of original content [31]. Like fair use, derivative works have faced scrutiny amid advancements in artificial intelligence [32].

Permissive licenses offer a straightforward solution, simplifying compliance by removing the need to evaluate derivative work status. This approach reduces barriers and encourages broader data use [33]. Yet, a unified framework for public health data licensing remains absent, leading to inconsistent access across jurisdictions and challenges for users.
This study aims to analyze datasets within the COVID-19 Open Data metadataset, focusing on metrics such as data type, licensing, and publication sources. By examining this comprehensive collection, we propose recommendations for public health data publishers to facilitate broader use of public health data, support automated aggregation, and enhance accessibility.

METHODS

DATA COLLECTION

This section outlines the methodology used to assess the licensure of data sources and determine their eligibility for inclusion in our metadataset. The process began with an exploratory search to identify potential datasets addressing gaps in our objectives, whether related to new data types or specific geographical or temporal coverage. This search utilized general search engines and targeted queries in governmental and public data repositories, yielding several potential datasets.

Datasets were evaluated using both technical and non-technical criteria. Licensing terms and conditions were a primary focus, with preference given to datasets under licenses compatible with our metadataset. These datasets were directly incorporated unless other terms introduced conflicts. Conversely, datasets bound by restrictive share-alike licenses, which impose conditions on the licensing of aggregated outputs, were typically excluded unless successful negotiations with the providers secured more permissive terms.

The feasibility of reliable data aggregation was another critical consideration. Challenges such as network instability and schema volatility sometimes required manual adjustments to the aggregation protocol. Additionally, datasets in machine-readable formats were prioritized for their potential to enable automation, over less accessible formats, such as images or charts.

After this technical evaluation, selected datasets underwent a rigorous legal review, described in detail in the following section, to ensure compliance with licensing requirements for their inclusion in the metadataset. This review involved structured steps to identify potential licensing issues and confirm adherence to relevant legal standards.

For exclusive datasets with incompatible licenses or barriers to automated aggregation, outreach efforts were made to engage with the publishers. These negotiations aimed to either re-license the data under more permissive terms or improve accessibility. While some discussions successfully improved access or licensing conditions, other publishers maintained restrictions to prevent potential misinterpretation of raw data.

LEGAL EVALUATION

Initial Review by Core Team

The legal review began with a preliminary examination by our core team to identify any obvious licensing conflicts or restrictions that could affect a dataset's inclusion in the metadataset. This step served as an initial filter to pinpoint clear legal barriers and streamline the selection process for detailed legal scrutiny.

From a non-technical perspective, datasets were assessed for the presence of personally identifiable information (PII) or data from which PII could be inferred. Aggregated datasets, especially those comprising sufficiently large samples at an administrative region level, were deemed to mitigate PII concerns, adhering to data protection frameworks such as the GDPR [36].

Compilation for Legal Counsel Review

Datasets that passed the initial screening were cataloged into a comprehensive list for further examination by our legal counsel. This documentation included detailed notes on potential legal considerations flagged by the core team, helping to guide the legal review process. The complete list of data sources and their associated licenses was later shared as part of our transparency efforts and for proper attribution.

Consultation with Licensing Experts

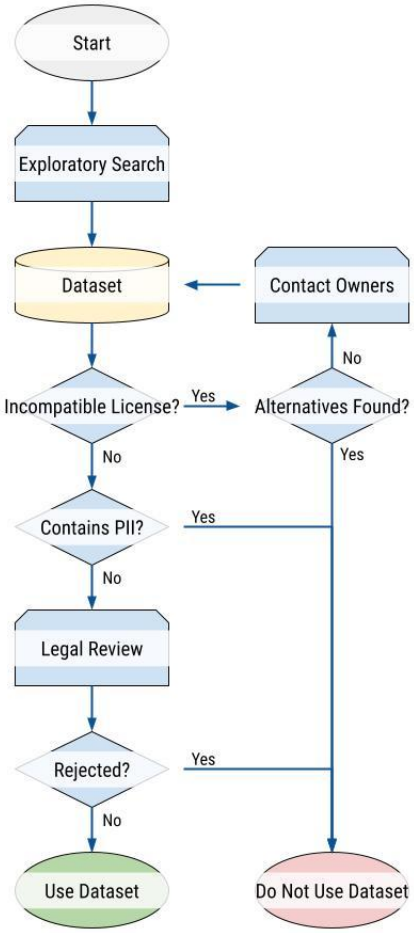
For datasets with ambiguous licensing terms or requiring specialized knowledge, our legal counsel consulted external data licensing experts. These experts, chosen for their expertise in relevant domains, provided insights into resolving complex licensing issues, ensuring compliance with jurisdictional and copyright requirements.

Final Approval and Inclusion

Datasets that underwent this layered review process and received endorsement from our legal counsel were approved for inclusion in the metadataset. This endorsement confirmed that each dataset met the legal standards for use and redistribution, aligning with our commitment to fostering open and accessible public health data.

This structured approach to legal review, depicted in the decision flow diagram in Figure 1, demonstrates our dedication to upholding high standards of legal and ethical data use. By integrating internal assessments with expert consultations, we ensured that the metadataset not only serves as a reliable resource for public health analysis but also complies with the intricate legal landscape surrounding data licensing and usage.

FIGURE 1: FLOWCHART DESCRIBING THE DATASET INGESTION PROCESS.



DATA PUBLICATION

Republishing the aggregated data within the COVID-19 Open Data metadataset required a carefully considered licensing framework. Given the diverse origins of the datasets, each with unique licenses and terms, the chosen strategy needed to respect these individual stipulations while ensuring broad usability. To address this complexity, we selected the Creative Commons Attribution (CC-BY) license for the metadataset. This permissive licensing scheme was chosen for its compatibility with the varied licenses of the included datasets, requiring only that appropriate attribution be provided to the original data sources. This requirement is aligned with the common denominator of terms across the constituent datasets and reflects our commitment to promoting data accessibility and supporting applications in research, policy development, and other fields.

Our approach was guided by the principle that the metadataset's license should neither impose fewer restrictions than the original datasets' licenses nor introduce unnecessary limitations. For example, adopting a license such as Creative Commons Zero (CC0) or placing the metadataset in the public domain would conflict with the attribution requirements mandated by certain source datasets, thereby violating their licensors' terms. Conversely, adopting a more restrictive license, such as one including a ShareAlike condition, was considered but ultimately rejected. While legally permissible, such a choice could hinder the dataset's usability by requiring derivative works to adopt the same licensing terms. This would conflict with our goal of maximizing the metadataset's accessibility and utility for a wide range of users.

RESULTS

The COVID-19 Open Data metadataset aggregates information from 260 distinct sources, comprising a total of 301 datasets, as some sources publish multiple datasets. To enhance the analysis of trends, patterns, and anomalies, we categorize these data sources into various groups. Classification is based on multiple criteria derived from metadata, including data type, publishing source, and license category.

DATA TYPE

Data type refers to the contents of the dataset, including the variables and whether they describe outcomes or covariates. As seen in Table 1, 200 of the data sources contain outcome data, and 60 for related covariates.

Outcome data describes medical outcomes, such as positive test cases and hospitalizations. The datasets with outcome data can be subclassified into core epidemiology (cases, deaths and tests), hospitalizations (general admissions, intensive care, and assisted by ventilator) and vaccinations (single dose, two or more doses and doses by vaccine type). From the datasets classified as outcome data, the epidemiology subclassification had the largest number of data sources, totaling up to 100. This excludes the sources of epidemiological data that provide outcome data stratified by age and sex.

Covariates are variables that relate to outcome data, whether they have a direct relationship (such as non-pharmaceutical interventions) or are considered to be loosely related to behaviors that have a second-order effect on outcome variables (for example, mobility data being an indicator of social behavior). Similarly, the covariates datasets can be subclassified into each of the covered categories. From the datasets classified as covariates, the vaccinations subclassification had the largest number of data sources, totaling up to 35.

TABLE 1: BREAKDOWN OF DATASETS BY CLASSIFICATION (OUTCOME VS COVARIATES), SUBCLASSIFICATION, DESCRIPTION, AND COUNT OF DATA SOURCES (INCLUDING AUTHORITATIVE, JOURNALISTIC AND CROWDSOURCED BREAKDOWNS).

Table Name	Classification	Subclassification	Content	Data Sources
by_age	outcome	By Age	Epidemiology and hospitalizations data stratified by age.	33
by_sex	outcome	By Sex	Epidemiology and hospitalizations data stratified by sex.	28
demographics	covariates	Demographics	Various population statistics.	7
economy	covariates	Economy	Various economic indicators.	3
epidemiology	outcome	Epidemiology	COVID-19 cases, deaths, recoveries and tests.	100

geography	covariates	Geography	Geographical information about the region.	3
google_search_trends	covariates	Search Trends	Trends in symptom search volumes due to COVID-19.	1
health	covariates	Health	Health indicators for the region.	5
hospitalizations	outcome	Hospitalizations	Information related to patients of COVID-19 and hospitals.	39
lawatlas_emergency_declarations	covariates	Emergency Declarations	Government emergency declarations and mitigation policies.	1
mobility	covariates	Mobility	Various metrics related to the movement of people.	1
oxford_government_response	covariates	Government Response	Government interventions and their relative stringency.	1
vaccination_search_insights	covariates	Vaccination Search	Trends in Google searches for COVID-19 vaccination information.	1
vaccinations	covariates	Vaccinations	Trends in persons vaccinated regarding various Covid-19 vaccines.	35
weather	covariates	Weather	Dated meteorological information for each region.	1
worldbank	covariates	WorldBank	Latest record for each indicator from WorldBank.	1

PUBLISHING SOURCE

The publishing source refers to where the information originates from. Although the COVID-19 Open Data project attempted to use information as close to the original source as possible, sometimes data was not available directly from the relevant health authority or it was very challenging to process it. This is the case in developing nations, areas subject to local conflicts unrelated to the COVID-19 pandemic, or with data publishers that preferred to keep a tighter control over how the data could be used. Only the source that published the data used in the COVID-19 Open Data project is considered in this analysis, even if that data originates from yet another data source. The subclassification of publishing sources into authoritative, journalistic, and crowd-sourced can be found in Table 2.

A publishing data source is authoritative when the data is coming directly from a health authority, such as a department of health; or government organization, such as the United Kingdom's National Health Service (NHS) or the United States' Centers for Disease and Control (CDC). Authoritative data sources represent 77.69% of the data.

TABLE 2: BREAKDOWN OF DATASETS BY PUBLISHING SOURCE (AUTHORITATIVE, JOURNALISTIC OR CROWDSOURCED).

Table Name	Authoritative Data Sources	Journalistic Data Sources	Crowdsourced Data Sources
by_age	31	-	2
by_sex	27	-	1
demographics	3	-	4

economy	2	-	1
epidemiology	71	3	26
geography	1	-	-
google_search_trends	1	-	-
health	4	-	1
hospitalizations	34	1	4
lawatlas_emergency_declarations	1	-	-
mobility	1	-	-
oxford_government_response	-	1	-
vaccination_search_insights	1	-	-
vaccinations	23	-	12
weather	1	-	-
worldbank	1	-	-

Journalistic data sources are those where the data is aggregated by a reputable third party source, such as a newspaper. In many cases, the main role of the journalistic data source is to aggregate a number of distinct data sources and validate whether they are correct, choosing different sources for different data points depending on some criteria. For example, the New York Times published case counts for USA counties [37] that were significantly simpler to aggregate than those published by the relevant health authority.

Crowd-sourced data sources are those aggregated by a non-journalistic third party source, such as a nonprofit organization. Especially at the beginning of the pandemic, before authoritative data sources established a data publishing pipeline, several groups of individuals collectively gathered data from wherever it was available (including newspaper reports, government press releases, hospital announcements) and made it available for reuse. For instance, the non-profit FinMango republished epidemiological data from multiple health authorities for the countries of Kenya, Sierra Leone and South Africa in an aggregated, machine-readable format. The original data was sometimes scattered across multiple websites or even available exclusively in the legend of a map provided in image format.

LICENSE CATEGORY

The License category is the broad classification of the data license type or, in the case of a lacking, specific license, the applicable conditions to use the data. As seen in Table 3, here are 20 distinct licenses in the various data sources of the COVID-19 Open Data metadataset, which are categorized in the following subgroups.

Permissive data sources include, for example, CC0 [38] (public domain) and CC-BY [6] (attribution required). These are considered permissive because, other than attribution, they impose no restrictions on the users of the data. Crucially, the permissive family of data licenses do not impose any requirements for the license that derivative work (such as data aggregations) need to be published as. The COVID-19 Open Data repository was published under the CC-BY license, in order to allow for unrestricted use of the data by downstream users of the dataset. 85.17% of the data sources used a permissive license.

Some data providers, such as the World Health Organization (WHO), use a custom license that can be considered permissive since none of the terms conflict with other permissive data licenses and aggregation and republication is permitted.

TABLE 3: BREAKDOWN OF DATASETS BY DATA LICENSING (PERMISSIVE VS OTHER).

Table Name	Permissive Licensing Data Sources	Other Licensing Data Sources
by_age	7	26
by_sex	6	22
demographics	6	1
economy	3	-
epidemiology	31	69
geography	3	-
google_search_trends	-	1
health	3	2
hospitalizations	16	23
lawatlas_emergency_declarations	-	1
mobility	-	1
oxford_government_response	1	-
vaccination_search_insights	-	1
vaccinations	25	10
weather	-	1
worldbank	1	-

Share-Alike licenses include CC-BY-SA, although some datasets were evaluated which used a similar but software-specific type of license such as GPL [39]. Share-alike licenses do not simply apply to data aggregations, but to any form of derivative work. This would include, for example, any visualizations made with the data such as charts or interactive tools. The share-alike license terms require any such derivative work to be published under the same license.

Since the data was aggregated by the COVID-19 Open Data project to be used by third parties with unknown requirements as well as interactive visualizations that made use of proprietary technology that could not be made public, the nature of share-alike licenses precluded the use of data sources with that license in the metadataset.

When a data source provided a license that could not be categorized as permissive or share-alike, or lacked a clear license altogether, it was classified as "other." These data sources required careful evaluation by legal experts specializing in data licensing. In most cases, the data did not include an explicit license, necessitating reliance on the custom terms of service of the relevant online portal, which were assessed for legal implications. In some instances, data was considered to fall under the fair use doctrine, even when the licensing terms were unclear or contradictory.

For example, the National Oceanic and Atmospheric Administration (NOAA) publishes the Global Summary of the Day (GSOD) and the Global Historical Climatology Network Daily (GHCN Daily) datasets. However, the licensing terms for these datasets are unclear. At the time of aggregation, the NOAA data search platform displayed a note indicating that the data was subject to the conditions outlined in Resolution 40 by the World Meteorological Organization (WMO) [40]. This resolution refers to an agreement among WMO members but does not specify usage restrictions. Additionally, contradictory information exists regarding usage restrictions, with some sources requiring citation of a specific publication [41]. On platforms such as the Google Cloud Platform Dataset Marketplace, these same datasets are linked to terms that describe them as available "free and without restriction" or even as CC0-licensed [42].

CROSS-ANALYSIS OF DATASET BREAKDOWNS

Cross-analysis of the different breakdowns against data licensing was also performed, with the intent of shedding some light on potential correlation between the different classification categories. As seen in Table 4, data sources with other licensing are the dominant group across all breakdowns, accounting for 60.77% of all data sources; however, they are disproportionately represented in the outcome data type (70%) and authoritative data source (65.84%). In contrast, 30% and 43.10% of the data sources had other licensing for the covariates data type and the non-authoritative data publishers, respectively.

TABLE 4: CROSS-ANALYSIS OF LICENSING BY DATASET CATEGORY, COMPARING DATA TYPE AND PUBLISHER SUBCATEGORIES.

Dataset Breakdown	Permissive Licensing Data Sources	Other Licensing Data Sources
Data Type: Outcome	60 (30%)	140 (70%)
Data Type: Covariates	42 (70%)	18 (30%)
Data Publisher: Authoritative	69 (34.2%)	133 (65.8%)
Data Publisher: Journalistic	3 (60%)	2 (40%)
Data Publisher: Crowdsourced	30 (56.7%)	23 (43.3%)

DISCUSSION

LICENSING CHALLENGES AND IMPLICATIONS

A recurring challenge in data aggregation is determining copyright ownership, particularly when data is repurposed or republished by intermediaries. While facts cannot be copyrighted, the collection or representation of these facts (e.g., under sui generis database rights) introduces complexities in determining the applicable licensing terms—whether those of the original publisher, the aggregator, or both.

Our comprehensive examination of numerous data sources highlights a strong preference among data consumers for clear and compatible licensing, emphasizing the critical need for accessibility and utility in public health data.

In this context, a distinct preference for permissive licensing emerges. This inclination is driven by the simplicity of fulfilling attribution requirements which facilitates broader use of data by enabling diverse users—from non-governmental organizations and private research entities to vaccine manufacturers—to integrate the data into their work with fewer barriers.

Our analysis also revealed that many datasets suffered from licensing ambiguities or relied on bespoke licenses. While such datasets might nominally be protected under the fair use doctrine, the lack of explicit licensing places an undue burden on data aggregators and republishers. These entities are not only required to provide attribution but must also detail the unique licenses or terms of service, complicating compliance and risking misinterpretation by downstream users. Adopting an explicit, widely recognized permissive license could address these challenges by offering data aggregators a consistent framework, eliminating the need for end-users to navigate bespoke licensing agreements. The UK Parliamentary Office of Science and Technology has reinforced this view, noting that international data sharing can be improved through the adoption of common data standards and sharing frameworks [43].

Encouragingly, some data publishers have recognized the limitations of restrictive licensing and transitioned to more open models. For instance, The COVID Tracking Project shifted from a share-alike CC-BY-SA license to a simpler attribution-based CC-BY license [44], demonstrating a commitment to fostering more accessible data practices.

Despite potential selection bias in our dataset compilation—due to the exclusion of sources with incompatible licenses or accessibility barriers—the incidence of datasets for which no suitable alternatives could be identified was minimal.

In conclusion, this analysis underscores the critical need for public health data to be easily aggregable through automated processes and governed by clear, user-friendly licensing. By illuminating the challenges and preferences inherent in the current data licensing landscape, this study aims to inform and influence future practices among data publishers, fostering a more open, accessible, and efficient ecosystem for public health data utilization

THE ROLE OF AI AND MACHINE LEARNING

Artificial intelligence (AI) and machine learning (ML) have the potential to revolutionize public health by enabling faster and more accurate analysis of complex datasets. These tools can identify patterns, predict disease outbreaks, and evaluate the impact of public health interventions. However, their success depends heavily on having access to high-quality, clearly licensed data.

One major challenge is the fragmented and restrictive licensing of public health data. Unclear ownership and overly restrictive terms make it harder to build and use large, interoperable datasets for AI/ML research [45]. Without these datasets, it is difficult to train models that are reliable and useful across different public health contexts.
The permissive licensing frameworks proposed in this study, such as Creative Commons Attribution (CC-BY), are particularly suited to AI/ML needs. These licenses allow data to be easily shared and combined while minimizing legal complexities. On the other hand, restrictive licenses or unclear terms can limit innovation, as researchers may face barriers when combining or reusing data for AI/ML applications.

Another challenge is understanding how licenses apply to derived datasets. For instance, some licenses may require legal expertise to determine if combining datasets creates a "derivative work." This uncertainty slows down research and discourages collaboration. Clear, simple licensing practices could help overcome these challenges and make it easier to integrate public health data into AI/ML workflows.

Ethical considerations also play an important role. The decline of open data resources risks ignoring critical issues like privacy, consent, and fairness [46]. Public health data publishers need to find a balance between making data widely available and protecting individual privacy while ensuring that AI/ML technologies benefit all communities.
By adopting clear, permissive licensing and user-friendly data formats, public health organizations can support AI/ML innovation. These steps will make it easier to use data for developing new tools and strategies, ultimately helping to improve responses to public health challenges.

GUIDELINES FOR DATA PUBLISHERS

To enhance the accessibility and utility of public health data, we propose a concise set of guidelines tailored for data publishers, including government agencies and health authorities. These guidelines, detailed on the COVID-19 Open Data project page, are designed to address specific challenges identified during the project's development, complementing existing resources like the Open Data Handbook [47,48].

Machine-Readable Formats

Recommendation: Publish data in electronic, machine-readable formats (e.g., CSV) rather than in binary formats (PDF) or as images and charts. This ensures compatibility with accessibility tools like screen readers and diminishes the need for third-party data reformatting.
Example: Utilize CSV for datasets instead of PDFs for reports or summaries, facilitating direct use and analysis.

Stable Access and APIs

Recommendation: Provide stable URLs or APIs for regularly updated data to support automatic downloading. For historical data preservation, maintain versioned datasets accessible through distinct links.
Example: Implement an API for daily case counts that allows users to query current and past data without manual searching.

Aggregated Time-Series Format

Recommendation: For data subject to revisions, publish in an aggregated time-series format to avoid the necessity for users to redownload past versions when updates are made.
Example: Combine daily epidemiological data into a single, continuously updated file rather than separate daily files.

Timeliness and Regularity

Recommendation: Ensure data, especially on topics of general interest like global pandemic epidemiology, is updated in a timely and consistent manner. Establish and adhere to a publishing cadence that balances informational value and feasibility.
Example: Decide on a weekly update schedule for epidemiological data, maintaining this cadence to ensure consistency for users.

Consistency in Data Formatting and Metrics

Recommendation: Maintain consistent data formatting and reporting metrics over time to allow for comparative analysis across different periods.
Example: Use the same data structure and metrics for case counts, ensuring comparability from one update to the next.

Confidentiality and Privacy

Recommendation: Apply techniques to preserve the confidentiality of human subjects, such as redacting identifiable information, aggregating data to obscure individual identities, and implementing differential privacy measures where appropriate.
Example: Aggregate case data to the level of municipalities or districts, applying differential privacy to small counts to prevent identification.

These guidelines are intended to serve as a foundation for data publishers seeking to improve the accessibility, reliability, and ethical use of public health data. By adhering to these principles, publishers can significantly contribute to the global effort of managing public health crises through informed, data-driven strategies.

AVAILABILITY OF DATA AND MATERIALS:

All the data used for this analysis is available in the COVID-19 Open Data repository, hosted on the GitHub platform at the following URL: https://github.com/GoogleCloudPlatform/covid-19-open-data.

FUNDING AND COMPETING INTERESTS:

The authors did not receive any external funding for this work, and have no competing interests to declare.

AUTHORS' CONTRIBUTIONS:

O.W. performed the original research, analyzed the data, drafted and revised the paper. M.R., drafted and revised the paper.

References

1. Checchi F, Warsame A, Treacy-Wong V, Polonsky J, van Ommeren M, Prudhon C. Public health information in crisis-affected populations: a review of methods and their use for advocacy and action. Lancet. 2017 Nov 18;390(10109):2297–313.
2. Wahltinez O, Cheung A, Alcantara R, Cheung D, Daswani M, Erlinger A, et al. COVID-19 Open-Data a global-scale spatially granular meta-dataset for coronavirus disease. Scientific Data. 2022 Apr 12;9(1):162.
3. Teresa M. Harrison University at Albany, SUNY, Albany, NY, Theresa A. Pardo University at Albany, SUNY, Albany, NY. Data, Politics and Public Health. Digital Government: Research and Practice [Internet]. 2020 Dec 3 [cited 2024 Mar 28]; Available from: https://dl.acm.org/doi/10.1145/3428123
4. Nagaraj A, Shears E, de Vaan M. Improving data access democratizes and diversifies science. Proc Natl Acad Sci U S A. 2020 Sep 22;117(38):23490–8.

5. Hansen J, Wilson P, Verhoeven E, Kroneman M, Kirwan M, Verheij R, et al. Assessment of the EU Member States' rules on health data in the light of GDPR. European Union; 2021.
6. Creative Commons — Attribution 4.0 International — CC BY 4.0 [Internet]. [cited 2023 Apr 21]. Available from: https://creativecommons.org/licenses/by/4.0/
7. Mockus M, Palmirani M. Open Government Data Licensing Framework. Electronic Government and the Information Systems Perspective. 2015;287–301.
8. Grabus S, Greenberg J. The Landscape of Rights and Licensing Initiatives for Data Sharing. CODATA. 2019 Jul 4;18:29–29.
9. Wang H, Cleary PD, Little J, Auffray C. Communicating in a public health crisis. Lancet Digit Health. 2020 Oct;2(10):e503.
10. Khayyat M, Bannister F. Open data licensing: More than meets the eye. Information Polity. 2015 Jan 1;20(4):231–52.
11. The multiple meanings of open government data: Understanding different stakeholders and their perspectives. Gov Inf Q. 2015 Oct 1;32(4):441–52.
12. Gianfredi V, Pennisi F, Lume A, Ricciardi GE, Minerva M, Riccò M, et al. Challenges and Opportunities of Mass Vaccination Centers in COVID-19 Times: A Rapid Review of Literature. Vaccines. 2021 Jun 1;9(6):574.
13. O'Doherty KC, Christofides E, Yen J, Bentzen HB, Burke W, Hallowell N, et al. If you build it, they will come: unintended future uses of organised health data collections. BMC Med Ethics. 2016 Sep 6;17(1):1–16.
14. COVID-19 vaccination coverage among hospital-based healthcare personnel reported through the Department of Health and Human Services Unified Hospital Data Surveillance System, United States, January 20, 2021-September 15, 2021. Am J Infect Control. 2021 Dec 1;49(12):1554–7.
15. Reis-Santos B. Health Information Systems: how much progress are we making? Epidemiol Serv Saúde. 2023 Aug 21;32(2):e2022433.
16. McBride K, Olesk M, Kütt A, Shysh D. Systemic change, open data ecosystem performance improvements, and empirical insights from Estonia: A country-level action research study. Information Polity. 2020 Jan 1;25(3):377–402.
17. Analysis of crisis communication by the Prime Minister of Australia during the COVID-19 pandemic. International Journal of Disaster Risk Reduction. 2021 Aug 1;62:102375.
18. Wahltinez O, Glasgow S, Cheung A, Glasgow JF, Noguera M, Glasgow JW, et al. The Mango Model: Best Practices in the Creation of a COVID-19 Open Data Project Through a Partnership with Google Health and the Non-Profit FinMango. Am J Health Educ [Internet]. 2023 Jul 4 [cited 2024 Mar 28]; Available from: https://www.tandfonline.com/doi/abs/10.1080/19325037.2023.2209620
19. Zuiderwijk A, Janssen M, Davis C. Innovation with open data: Essential elements of open data ecosystems. Information Polity. 2014 Jan 1;19(1-2):17–33.
20. California Open Data [Internet]. [cited 2023 Apr 21]. Available from: https://data.ca.gov/
21. Michigan | Open Data | Michigan | Open Data [Internet]. [cited 2023 Apr 21]. Available from: https://data.michigan.gov/
22. Data | Centers for Disease Control and Prevention [Internet]. Data.CDC.gov. [cited 2023 Apr 21]. Available from: https://data.cdc.gov/
23. OpenDataSUS [Internet]. [cited 2023 Apr 21]. Available from: https://opendatasus.saude.gov.br/
24. Aufderheide P, Boyles JL, Bieze K. Copyright, Free Speech, and The Public's Right to Know. Journalism Studies [Internet]. 2013 Oct 24 [cited 2023 Apr 21]; Available from: https://www.tandfonline.com/doi/abs/10.1080/1461670X.2012.739320
25. Samuelson P. Copyright's fair use doctrine and digital data. Publishing Research Quarterly. 1995 Mar;11(1):27–39.
26. Henderson P, Li X, Jurafsky D, Hashimoto T, Lemley MA, Liang P. Foundation Models and Fair Use. Social Science Research Network [Internet]. 2023 Mar 27 [cited 2024 Mar 28]; Available from: https://papers.ssrn.com/abstract=4404340
27. Gurstein MB. Open data: Empowering the empowered or effective data use for everyone? First Monday [Internet]. 2011 Feb 7 [cited 2024 Mar 28];16(2). Available from: https://firstmonday.org/ojs/index.php/fm/article/view/3316
28. Kitchin R. The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. SAGE; 2014. 241 p.
29. Kamocki P, Straňák P, Sedlák M. The Public License Selector: Making Open Licensing Easier. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016. p. 2533–8.

30. Willmers M, Van Schalkwyk F, Schonwetter T. Licensing Open Data in Developing Countries: The Case of the Kenyan and City of Cape Town Open Data Initiatives. Afr j inf commun (Online) [Internet]. 2015 Dec 15 [cited 2024 Mar 28];(16). Available from: https://ajic.wits.ac.za/article/view/13639
31. Rachum-Twaig O. Copyright Law and Derivative Works: Regulating Creativity. Routledge; 2018. 206 p.
32. Gaon, H. A. The Future of Copyright in the Age of Artificial Intelligence. Edward Elgar Publishing; 2021. 288 p.
33. Creative Commons — Attribution-ShareAlike 4.0 International — CC BY-SA 4.0 [Internet]. [cited 2023 Apr 21]. Available from: https://creativecommons.org/licenses/by-sa/4.0/
34. Alamo T, Reina DG, Mammarella M, Abella A. Covid-19: Open-Data Resources for Monitoring, Modeling, and Forecasting the Epidemic. Electronics. 2020 May 17;9(5):827.
35. Suzgun M, Melas-Kyriazi L, Sarkar SK, Kominers SD, Shieber SM. The Harvard USPTO Patent Dataset: A Large-Scale, Well-Structured, and Multi-Purpose Corpus of Patent Applications [Internet]. 2022 [cited 2023 Apr 21]. Available from: http://arxiv.org/abs/2207.04043
36. Shedding light on the legal approach to aggregate data under the GDPR & the FFDR | UNECE [Internet]. [cited 2023 Apr 21]. Available from: https://unece.org/statistics/documents/2021/12/working-documents/shedding-light-legal-approach-aggregate-data-under
37. GitHub - nytimes/covid-19-data: A repository of data on coronavirus cases and deaths in the U.S [Internet]. GitHub. [cited 2023 Apr 21]. Available from: https://github.com/nytimes/covid-19-data
38. CC0 [Internet]. Creative Commons. [cited 2023 Apr 21]. Available from: https://creativecommons.org/share-your-work/public-domain/cc0/
39. GNU Project-Free Software Foundation. The GNU General Public License v3.0 [Internet]. [cited 2023 Apr 21]. Available from: https://www.gnu.org/licenses/gpl-3.0.en.html
40. Resolution 40 | World Meteorological Organization [Internet]. [cited 2023 Apr 21]. Available from: https://community.wmo.int/en/resolution-40
41. Menne MJ, Durre I, Korzeniewski B, McNeill S, Thomas K, Yin X, et al. Global Historical Climatology Network - Daily (GHCN-Daily), Version 3 [Internet]. 2012 [cited 2023 Apr 21]. Available from: https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc:C00861/html
42. NOAA GSOD [Internet]. [cited 2023 Apr 21]. Available from: https://www.kaggle.com/datasets/noaa/gsod
43. Drummond B, Christie L. Sharing public sector data. 2023 Apr 21 [cited 2023 Apr 21]; Available from: https://post.parliament.uk/research-briefings/post-pn-0664/
44. 'Data API' [Internet]. The COVID Tracking Project. [cited 2025 Feb 10]. Available from: https://covidtracking.com/data/api.
45. Longpre S, Mahari R, Lee AN, Lund CS, Oderinwale H, Brannon W, et al. Consent in Crisis: The Rapid Decline of the AI Data Commons. The Thirty-Eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track. 2024. Available from: https://openreview.net/forum?id=66PcEzkf95
46. Chan A, Bradley H, Rajkumar N. Reclaiming the digital commons: A public data trust for training data. In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23). Association for Computing Machinery; 2023. p. 855–868. doi:10.1145/3600211.3604658. Available from: https://doi.org/10.1145/3600211.3604658
47. Gorman RA. Copyright Protection for the Collection and Representation of Facts. Harv Law Rev. 1963;76(8):1569–605.
48. Share-PSI Best Practice: Develop an Open Data Publication Plan [Internet]. [cited 2023 Apr 21]. Available from: https://www.w3.org/2013/share-psi/bp/odpp/