A General Guide for Harmonizing Data: Drawing Lessons from Harmonizing COVID-19 PHSM Data

Cindy Cheng (cindy.cheng@hfp.tum.de)
Technical University of Munich (TUM) https://orcid.org/0000-0002-7345-827X
Luca Messerschmidt
Hochschule für Politik at the Technical University of Munich & Blavatnik School of Government, University of Oxford https://orcid.org/0000-0001-8069-381X
Issac Bravo
Hochschule für Politik at the Technical University of Munich (TUM)
Marco Waldbauer
Hochschule für Politik at the Technical University of Munich (TUM)
Rohan Bhavikatti
Independent Researcher
Caress Schenk
Nazarbayev University
Vanja Grujic
University of Brasilia
Tim Model
Delve
Robert Kubinec
New York University Abu Dhabi https://orcid.org/0000-0001-6655-4119
Joan Barceló
New York University Abu Dhabi https://orcid.org/0000-0001-6965-4342

Resource

Keywords: data harmonization, big data, COVID-19 government policies, public health and safety measures

Posted Date: February 17th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2595267/v1

Additional Declarations: There is **NO** Competing Interest.

A General Guide for Harmonizing Data: Drawing Lessons from Harmonizing COVID-19 PHSM Data

Cindy Cheng[1*], Luca Messerschmidt[1], Isaac Bravo[1], Marco Waldbauer[1], Rohan Bhavikatti[2], Caress Schenk[3], Vanja Grujic[4], Tim Model[5], Robert Kubinec[6] and Joan Barceló[6]

[1*]Hochschule für Politik, Technical University of Munich, Richard-Wagner Str. 1, Munich, 80333, Bavaria, Germany.
[2]Independent Researcher.
[3]School of Humanities and Social Sciences, Nazarbayev University, Kabanbay Batry Ave., 53, Astana, 010000, Kazakhstan.
[4]Faculty of Law, University of Brasilia, Campus Universitário Darcy Ribeiro Asa Norte, Brasilia , 10587, Brazil.
[5] Delve, 2225 3rd St, San Francisco, 94107, California, United States.
[6]Division of Social Science, New York University Abu Dhabi, Social Science Building (A5), Abu Dhabi, 129188, United Arab Emirates.

*Corresponding author(s). E-mail(s): cindy.cheng@hfp.tum.de; Contributing authors: luca.messerschmidt@hfp.tum.de; isaac.bravo@tum.de; marco.waldbauer@tum.de; rohanab10@gmail.com; cschenk@nu.edu.kz; vanjagrujiccoves@gmail.com; tim.model@delve.com; rmk7@nyu.edu; joan.barcelo@nyu.edu;

Abstract

Data harmonization is an important method for generating the requisite datasets to support big data analyses. To date however, articles about data harmonization are field-specific and highly technical, making it difficult for researchers to derive general principles for how to engage in and contextualize data harmonization efforts. This article provides general guidance and criteria for researchers who are considering undertaking such efforts or seek to evaluate the quality of existing ones. We derive these guidelines from the extant literature and our own experience in harmonizing data for the emergent and important new field of COVID-19 public health and safety measures (PHSM). We further introduce the methodology we employed for engaging in this data harmonization as a blueprint for researchers interested in engaging in manual data harmonization.

Keywords: data harmonization, big data, COVID-19 government policies, public health and safety measures

1 Introduction

Unprecedented technological advancements in information technology have ushered in a data science revolution, allowing scholars, companies and policy makers to conduct analyses at a scale, speed and granularity previously unimaginable [1, 2]. However as Elshawi et al. (2018) [3] note, "in practice, big data science lives and dies by the data. It mainly rests on the availability of massive datasets, of that there can be no doubt." From fields as varied as socio-economics [2, 4, 5] to ecology [6] and the 'Internet of Things' [7], data scientists report the lack of big data itself is a major bottleneck in using big data tools. Increasingly, data scientists must first sort through heterogeneous, incongruent, and fragmented datasets before any analyses can be conducted [8–13]. Such problems with data availability are often exacerbated in emergency situations where real time analyses are often stymied by unevenly documented or unclean data [14][pg. 358].

Because data harmonization entails untangling how highly complex data can be made to fit together, unsurprisingly, existing research on it has been both highly technical and field-specific[11, 15–19]. Correspondingly however, researchers interested in pursuing their own data harmonization or in evaluating the work of others currently lack general guidelines to help them. The need for broad guidelines will only grow given the increasing importance of data harmonization to big data science. To address this gap, we present a general set of (i) scope conditions outlining when data harmonization is possible and (ii) criterion for deciding whether it should be pursued given these scope conditions are fulfilled.

To develop these guidelines, we take a broad survey of existing data harmonization efforts from both the natural and social sciences and draw on our experience in harmonizing data on a pressing and emergent new research field, COVID-19 public health and safety measures (PHSM). Accordingly, our paper further introduces (iii) a novel, rigorous methodology for harmonizing PHSM data. Though dozens of research groups have sought to track PHSM, these individual data tracking efforts have succeeded in providing only an incomplete portrait of government COVID-19 responses, a situation exacerbated by the fact that many have stopped entirely, often due to funding constraints. Harmonizing PHSM data with due haste is desirable not only because of the emergency nature of the pandemic, but also for preserving the original sources underlying these data. We describe our efforts to seamlessly harmonize 8 different PHSM tracking efforts:

- ACAPS Government Measures (ACAPS)[20]
- COVID Analysis and Mapping of Policies (COVIDAMP)[21]
- Candian Dataset of COVID-19 Interventions (CIHI)[22]
- CoronaNet Research Project (CoronaNet)[23]
- John Hopkins Health Intervention Tracking for COVID-19 (HIT-COVID)[24]
- Oxford COVID-19 Government Response Tracker (OxCGRT)[25]
- World Health Organization EURO (WHO EURO) and US Center for Disease Control (WHO CDC) datasets on COVID-19 policies (retrieved from the WHO Public Health and Safety Measures (WHO PHSM)[26])

into the CoronaNet taxonomy with the help of 350+ research assistants around the world to provide a fuller picture of government responses to the pandemic.

We believe that our PHSM harmonization effort can serve as a guide for the harmonization of other datasets, especially in the social sciences. That is, in contrast to virtually all other harmonization efforts we could find, ours is largely implemented manually, providing us with an unusually intimate knowledge of data harmonization at the level of the individual observation (for more, see our Methodology section). In our main text, we illustrate the usefulness of our guidelines by applying them to the PHSM harmonization case and conclude with a discussion of how data harmonization illuminates complexities in the data generation process, which we hope will be of general interest to researchers writ large.

2 Results

Data harmonization is the practice of "reconciling various types, levels and sources of data in formats that are compatible and comparable, and thus useful for better decision-making" [27][pg. 360]. It is thus distinct from data integration, also known as data linkage [28], in that (successful) data harmonization results in a single cohesive dataset made from conceptually similar datasets (e.g. combining multiple datasets on COVID-19 PHSM) while data integration results in a multidimensional dataset made from conceptually different datasets (e.g. combining multiple datasets on COVID-19 PHSM, COVID-19 deaths, and GDP; e.g. the PERISCOPE Data Atlas [29, 30])[31]. To create a cohesive dataset, data harmonization can be understood as resolving differences along at least three dimensions [32][pg. 262]:

- Structure (i.e. conceptual schema)

- Syntax (i.e. data format)
- Semantics (i.e. intended meaning of words)

How harmonization across these dimensions is ultimately achieved can be broadly understood as stringent or flexible [33]. Stringent harmonization refers to the use of identical measures and procedures across studies. Meanwhile flexible harmonization ensures that different datasets are inferentially equivalent, but not necessarily identical.

Data scientists and researchers will increasingly grapple with data harmonization challenges as more and more data is generated without coordination. In our review of the available literature, existing general guidelines are nevertheless still targeted toward specific fields e.g. epidemiology [34] or medicine [35]. Indeed, most work on data harmonization make little attempt at generalizability, resulting in highly technical, field-specific guidelines [36–40]. Drawing both on our own harmonization efforts and a review of the work of others, we present a set of general criteria for data scientists to use when considering the trade-offs between harmonized and unharmonized data. These criteria help data scientists answer two separate but related questions (1) What are the *scope conditions* for pursuing data harmonization? (2) Given that data harmonization is possible, when *should* it be pursued? We hope these criteria can help researchers consider when to pursue data harmonization in their own research or better assess the efforts of others. Below, we elaborate on these criteria and provide concrete examples from our harmonization of PHSM data.

2.1 What are the scope conditions for pursuing data harmonization?

The strategy to pursue data harmonization must always be understood in relation to original data collection. At the most basic level, without original data, there is nothing to harmonize. Conversely however, the existence of original data is not itself a sufficient condition for pursuing harmonization. In what follows, we outline a set of scope conditions researchers should use to evaluate the feasibility of pursuing harmonization.

2.1.1 Are there (partial) non-overlapping original data?

Original data collection will virtually always allow data scientists to operationalize a given concept at least as or more precisely than harmonized data. However, original data collection may not always be possible. This can occur for any number of reasons, including: (i) difficulty in identifying data to collect in a timely manner[41]. E.g., for researchers studying rare diseases, it is virtually impossible to identify large samples of relevant patients quickly [42]. (ii) Data collection may be prohibitively expensive even if the sample size needed is relatively small. E.g., studies that rely on neuroimaging often use data from less than 50 individuals in part because of the high material costs of MRI imaging [11]. Relatedly (iii) the number of observations needed may be prohibitively large, which may make original data collection unfeasible due to insuffcent resources or lack of requisite authority. E.g., although (geo-)data of human populations is a lynch pin of social science research, many countries lack detailed census data due to insufficient resources. Meanwhile global census data is not possible due to both cost and jurisdictional constraints [19].

Moreover, original datasets not only must exist, but they also need to have least partially non-overlapping units of analysis. That is, data harmonization may only be possible if, e.g., there are datasets which track different patients with rare diseases but not if there are different datasets which track the same patients with rare diseases. This latter case would call for data integration rather than data harmonization because the non-overlap occurs in the conceptual variable collected rather than in the unit of analysis.

2.1.2 Are the different original datasets similar in terms of syntax, structure and semantics?

The availability of separate non-overlapping partial datasets is a necessary, but insufficient condition for proceeding with data harmonization. Also important to consider is similarity across the structure (i.e. conceptual schema), syntax (i.e. data format) and semantics (i.e. intended meaning of words) of different original datasets. [43]

In some cases, such differences in datasets can be controlled for when the data is analyzed. For instance, while MRI diffusion scans display variation even when using the same scanner (to say nothing of across-scanner variability), statistical techniques can be used to account for these differences [11, 44]. In other cases, trade-offs between the quantity of datasets that can be harmonized and the precision with which the harmonized data operationalizes the desired underlying concepts may need to be balanced [45–48].

Indeed, when there are differences in sampling across harmonized datasets as well as conceptualization of measures or data collection instruments, i.e., the data structure, researchers may need to discard many original data points to proceed with harmonization [43]. These issues are more pronounced when harmonizing data across different social contexts. Indeed, how a dataset is constructed, i.e., its syntax, is often “dependent upon the units and scale of measurement within each social system”[49][40]. As Boyden and Walnicki find [28], even when different datasets contain similar information about household wealth, standardising these measures across different survey rounds and national contexts was not possible, often due to semantic differences. They themselves split the difference by creating a multidimensional wealth index instead, which allows for the inclusion of more observations at the cost of less precision in the operationalization of the original measures. In short, while data harmonization can often be of the greatest value when it combines datasets across different countries or regions, researchers must account for the social contexts in which such datasets were conceptualized and gathered to ensure functional, linguistic and cultural equivalency of the variables. [50].

2.1.3 Is the desired data time sensitive?

Time plays a crucial role when adjudicating whether data harmonization is possible. The relative tradeoffs between original data collection and data harmonization largely come down to a tug of war between time and resources, versus the relative fit, quality and generalizability of the resulting data.

Here we further distinguish data harmonization into two archetypes: retrospective (i.e. ex-post harmonization or output harmonization [5]) and prospective harmonization (i.e. ex-ante harmonization or input harmonization [5]). Retrospective harmonization refers to harmonization of already collected datasets as opposed to the prospective harmonization, which entails harmonizing research designs or methodologies before the data is collected.

In some circumstances, original data collection is not possible and only retrospective harmonization is feasible. This is canonically true whenever researchers wish to analyze past events or behaviours [51]. For instance, those seeking to analyze survey data for a past time frame must rely on previous surveys, if any exist, and work with the set of questions asked at the time [52–58]. In other cases, while original data collection may theoretically be possible, the time impermanence of primary sources may render it unfeasible to fully implement. For instance, while harmonizing mobile phone usage presents a promising avenue for analyzing on mobility patterns, the rapid pace at which new providers arise and the extent to which users switch between services means that the accessibility or availability of this data are often difficult to maintain [59].

Prospective harmonization is a distinct form of original data collection where research methodologies are harmonized before data collection takes place. Though prima facie, prospective data harmonization would appear to be strictly dominant over both original data collection and retrospective data harmonization, its disadvantages are also substantial. First, it can be challenging to implement because it requires agreement on standardized measures among different researchers who likely have diverse research goals. These challenges are exacerbated when standardized measures must be created contemporaneously [60]. Moreover, if a given dataset already exists and possesses a substantial history and organizational support, it may require tremendous effort to overcome institutional resistance to coordinate different stakeholders around a new methodology [47]. Furthermore, there is no guarantee that the resulting harmonized methodology would be methodologically more robust compared to alternative strategies. Standardization can create winners and losers [61, 62], and the final harmonized methodology may better reflect the institutional power of those advocating for it [63, 64] rather than its scientific rigor. Meanwhile, even if these challenges are overcome but the desired data is part of a longer time series, then previously collected data cannot be included in prospective harmonization [65]. It is also not always possible to anticipate future data needs, and as the history of national accounts can attest to, in the worst case scenario, the same variable is used to measure different concepts over time [66–68]. Finally, there are often real world constraints which limit the utility of prospective data harmonization. For instance while the European Influenza Surveillance Scheme coordinate to monitor seasonal influenza strains, differences in health care systems and health insurance systems limit the extent to which there can be congruence in the output data [69].

2.2 When should data harmonization be pursued?

The possibility of pursuing data harmonization does not mean that it is ultimately the best methodological strategy. Even if the scope conditions outlined above are fulfilled, ultimately, researchers should not lose sight of the fact that the end goal is a complete and clean dataset of variables that best suit their analytical or research needs. To that end, the set of questions that we pose below encourages data scientists to consider how data harmonization, when it is possible, compares to original data collection, reliance on existing datasets, or other forms of data aggregation to furthering their analytical goals.

2.2.1 What can be gained from data harmonization?

Generally, analyses using harmonized data can increase the statisical power of subsequent analyses compared to those done on individual datasets [11, 70]. Deriving reasonable estimates of how data harmonization can forward a desired analyses either in terms of data completeness, data quality or data validity can help researchers assess the relative value of engaging in data harmonization. When it is not possible to derive such estimates beforehand, we suggest that researchers conduct pilot studies in order to gain a more concrete sense of what could potentially be gained.

2.2.2 What can be lost from data harmonization?

The existence of different datasets in a given field often underscores the possibility of having different, yet valid conceptualizations and operationalizations of a given topic. Harmonizing different datasets may increase the internal coherence at the expense of minimizing real and potentially important diversity in theoretical approaches toward a given topic. Researchers bent on creating standardized measures at all costs should be cognizant of the risk of pandering to the lowest common denominator to achieve comparability and thus losing "important meta data or disconnection from local meanings and circumstances." [28]. If after conducting an assessment of what may be lost from data harmonization, the researcher decides to proceeds, making these tradeoffs transparent for the research community overall can contribute to the rigor of analyses conducted in that field.

2.2.3 What are the limits of data harmonization?

Harmonizing data is rarely equivalent to building a complete dataset. That is, putting aside the issue of what may be lost in creating a harmonized dataset, the harmonized dataset itself often has limitations that should be acknowledged and evaluated.

One real limitation is that the observations included in the harmonized datasets are often only a subset of what could theoretically be collected. This may be true both when the underlying datasets to be harmonized do and do not contain information about the desired data. With regards to the former, former cultural or language barriers[43] may make it difficult impossible to harmonize data across variables like wealth or educational outcomes for instance even which such data are available [28][pg. 7]. Meanwhile, if the underlying datasets do not themselves contain the desired data, an issue that we flesh out more fully in our discussion of PHSM data below, then the harmonized data will also face the same data limitations.

Despite the importance of reporting the scope of one's data harmonization efforts, researchers do not appear to consistently report this information [38, 39, 41], with this being a particular problem some fields [71]. Meanwhile, other fields have made progress on providing ready-made tools and resources for researchers to perform such an evaluation [72]. Reporting on the limitations of data harmonization is important for helping a given field identify research gaps or giving researchers proper context for using a harmonized dataset.

2.2.4 What underlying biases may need to be accounted for when harmonizing data?

The data harmonization process may propagate existing errors from original datasets [73] or generate new ones during the data harmonization process. Scholars should both be aware of these potential issues and take measures to minimize them.

While data harmonization can mitigate the biases that existing errors in individual datsets may introduce to a given analyses (i.e., batch effects), its ability to do so is conditional on (i) the number of datasets harmonized (the more the better) and (ii) the extent to which a systematic error for a given dataset is random at the level of the dataset (i.e. the same type of error is not made in all individual datsets). In the worst case, the number of harmonized datasets are few and all exhibit the same types of error, which can compound the errors underlying the original dataset. However, even if the above conditions are fulfilled, it would still desirable to reduce batch effects in order to improve the overall precision of subsequent analyses conducted with the harmonized data.

Strategies to address batch effects head on may range from analog to technical. Analog methods include (i) recruiting larger sample sizes in the underlying data to more closely approximate the population-wide distribution or (ii), in the case of human subject datasets, recruiting a subset of subjects who can travel to multi-site locations to calibrate measurement errors or [74, 75]. Meanwhile, other researchers have proposed statistical technique to account batch effects, from simpler strategies like pre-processing [76] or outlier detection methods [77] to more complex model-based techniques, like e.g. linear or deep learning models, which tend to be quite specific to different fields and datasets [77–81].

2.2.5 What cooperative resources are available for harmonizing data?

Ultimately, data harmonization entails translating each dataset to speak the same taxonomical 'language'. Across the numerous harmonization efforts we have surveyed, virtually all emphasize the importance of cooperation and coordination among different partners. Such exchanges were useful for both resolving confusion or misunderstandings about different taxonomies or methodologies and increasing the capacity for piloting data harmonization efforts in a timely manner. That being said, researchers have also underscored that maintaining such cooperation can itself be a resource-intensive undertaking [35, 82].

While the form of communication and cooperation across different harmonization efforts will necessarily be idiosyncratic, given its importance to the success of data harmonization, accounting for these potential cooperative resources is important for evaluating the subsequent feasibility of data harmonization. Furthermore, providing some documentation in this regard can be helpful for evaluating the quality of the subsequently harmonized data and can increasing the transparency of the data generating process.

2.2.6 What are alternatives to data harmonization?

Thus far, we have focused on developing criteria for assessing how useful qualitative data harmonization of multiple datasets might be toward forwarding research in a given field. Note, the term qualitative harmonization refers to the stitching together of existing datasets without fundamentally changing the original units of analysis. Depending on the research question however, researchers may also consider alternatives strategies to synthesize data that results in datasets with different units of analyses or data as the original datasets: meta analyses, data imputation of harmonized data, and statistical harmonization. While it is beyond the scope of the paper to provide an in depth review of these different methods, we hope our brief discussion below can stimulate researchers to consider the suitability of these alternative strategies for their research needs.

Meta analyses is a methodological strategy of synthesizing different research studies which has become increasingly used over the past four decades across a variety of fields [83, 84]. In brief, it encompasses "different techniques for synthesizing summary statitistics" [85][pg. 21]. By contrast, mega-analysis, also known in some fields as individual participant data meta-analysis, entails synthesizing information by pooling the raw data, and can be understood as a form of data harmonization. Many of the same considerations outlined in previous sections are also applicable when considering whether meta or mega analyses may be more appropriate [35, 86]. Overall, while some studies suggest that mega-analyses can yield more precise results than meta-analysis [11, 87], others find that they can substantively be quite similar [88, 89] and that differences between the two can largely be explained by differences in modelling assumptions as opposed to intrinsic reasons [90].

Data imputation of harmonized data may be employed to address the missingness or limits that may be introduced as part of the data harmonization process. Multiple imputation, that is, deriving multiple imputed values of the missing data to account for statistical uncertainty around the true value of the imputed value, has increasingly been employed for these purposes [91–93]. However, some limitations to this approach can occur if e.g. there are not enough overlap measures across datasets or if the measurement of the same concept are on paradigmatically different scales [94]. As with all modeling approaches, researchers should further take note that the assumptions underlying multiple imputation can substantially affect the quality or appropriateness of the resulting imputed values [95].

Meanwhile, statistical harmonization seeks to combine data from multiple sources by building an index or measure of a given construct using the underlying data [86]. Latent variable models such as factor models or item response models can and have often been employed for this purpose [96–98]. The resulting index thus represents a model-based measure of the underlying construct that is derived from the raw data but is qualitatively different from it. For instance, whereas the qualitative data harmonization we engage in this paper can be understood as a patchwork quilt of different datasets, statistical harmonization is analogous to a blanket made from the deconstructed threads of various fabrics. Note, in some circumstances, such models can also be understood as a form of data imputation for harmonized data [96].

2.3 Harmonizing COVID-19 PHSM Data

To illustrate the ability of our criteria to provide guidance and context for the feasibility and relative value for pursuing data harmonization over other data generation efforts, we apply them to our efforts to harmonize COVID-19 PHSM data. Comprehensive, high quality and timely COVID-19 PHSM data is crucial for forwarding understanding of the pandemic but unfortunately no single dataset has been able to capture the full scope or scale of such data [99]. Not only can harmonizing this data get us closer to this goal, it can also ensure that the data collected by trackers that have stopped their work are not lost and that the original sources underlying this data are preserved.

To summarize, our methodology combines automated and manual processes to harmonize data across 8 different datasets into the taxonomy for capturing COVID-19 PHSM developed by the CoronaNet Research Project. To that end, we followed the general methodology laid out in Figure 1. That is, after we evaluated the set of COVID-19 PHSM data to harmonize, we made taxonomy maps between the different external data and the CoronaNet taxonomy, removed policies from the external dataset irrelevant to the CoronaNet taxonomy, automatically deduplicated a portion of the external data. After

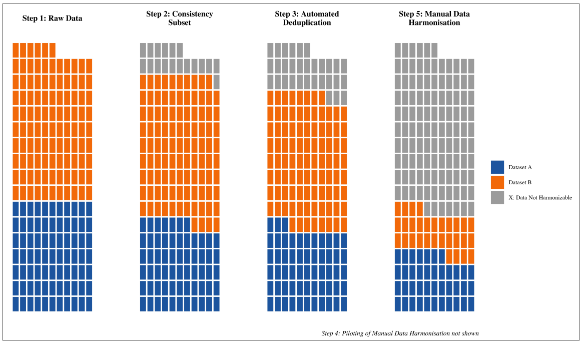

Fig. 1 Harmonization Methodology

having piloted manual harmonisation for a sample of the data, we are currently manually harmonizing the remaining external data into the CoronaNet dataset.

Below, we use our experience with harmonizing PHSM data to illustrate the utility of the guidelines that we elaborate above. For more detail, please see the Methodology section.

Are there (partial) non-overlapping original data? As Table A1 (see Appendix) shows, while there are more than 20 different datasets which capture data on government responses to the pandemic, no single dataset has been able to track all policies in part because the scope of the work has been too large for a single endeavor to handle with existing resources. Meanwhile, though there is clearly some duplication of effort among these datasts, the great variation in geographic coverage and temporal coverage in these datasets suggest that there is a high degree of non-overlapping observations in these datasets. PHSM data thus positively fulfill the first scope condition for pursuing data harmonization.

Are the different original datasets similar in terms of syntax, structure and semantics? COVID-19 PHSM data represent a relatively hard case in this regard insofar as each group tracking PHSM data developed their own taxonomy inferentially in response to real-time changes in the pandemic. To harmonize this data, we evaluated the similarity of the taxonomy for each dataset in terms of structure, syntax and semantics by creating taxonomy maps between it and the CoronaNet taxonomy. Because creating taxonomy maps is a high-cost endeavor (though note, other fields have made some progress in bringing down the costs [100]), we only made them for datasets with sufficiently large number of observations and geographic scope to warrant the effort, specifically: ACAPS, CIHI, COVIDAMP, CCCSL, HIT-COVID, WHO PHSM (WHO EURO and CDC) and ultimately decided to proceed with data harmonization with all except CCCSL (see subsection 4.2).

Meanwhile, in order to balance the tradeoff between data quantity and data precision, we chose to harmonize the external datasets into the CoronaNet taxonomy because it is the most detailed of the external datasets considered. As such, external policies recoded into the CoronaNet taxonomy are also augmented by the collection of additional data fields that were not systematically collected by other datasets.

Is the desired data time sensitive? The faster the access to high quality COVID-19 PHSM data, the more likely it can be used to understand the drivers and effects of the pandemic in real time. While we could have continued original data collection in sole accordance with the methodology outlined in Cheng et al. (2020) [23], our PHSM harmonization strategy allows us to straddle the best of both worlds insofar as relying on sources from external datasets likely helps reduce the search costs of finding original sources.

Since many PHSM trackers have stopped data collection due to funding constraints, data harmonization into the CoronaNet taxonomy also ensures that these data can live on and be used in an active research project[99]. Note that most data from external datasets do not save original PDF sources, leading to the gradual disappearance of the primary sources on which PHSM datasets are built. By recoding this data using the CoronaNet methodolgy, PDFs of such sources are saved before they disappear. If they have already disappeared, where possible, new sources of information are found and saved, mitigating the problem of digital expiration.

Note, that because of this methodology, our harmonization of PHSM data can be best understood as a case of stringent retrospective harmonization. It is stringent because all PHSM data from the 8 datasets will be harmonized to follow the same taxonomy. Meanwhile, it is retrospective because the data is coded from original primary sources, assuming the have not disappeared from the internet.

What can be gained from data harmonization? To our knowledge, no individual effort to document PHSM has been able to do so for all countries. Indeed, though at the time of writing, there are 128k+ observations unique to the CoronaNet dataset (150k+ total including already harmonized data), we identified 150,052 observations for the 7 datasets external to CoronaNet combined for data available until September 10, 2021. September 2021 was chosen as the cutoff date given our available resources and because most data tracking efforts had stopped or significantly slowed their data collection by this date except for OxCGRT, CIHI and WHO EURO (the latter two stopped in 2022). Should more resources become available we will expand our efforts to harmonize records for these datasets beyond this date. Based on our efforts so far, around 83% of external data do not overlap with the CoronaNet dataset, and of these around 44% can be recoded, suggesting there are potentially 55k additional observations to recode.

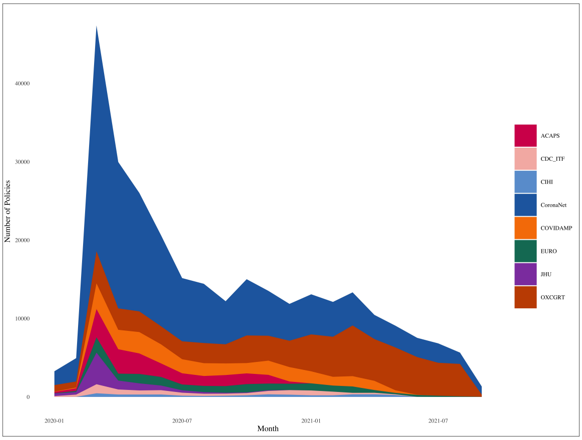

Fig. 2 Number of policies per date recorded by 8 different COVID-19 PHSM tracking efforts

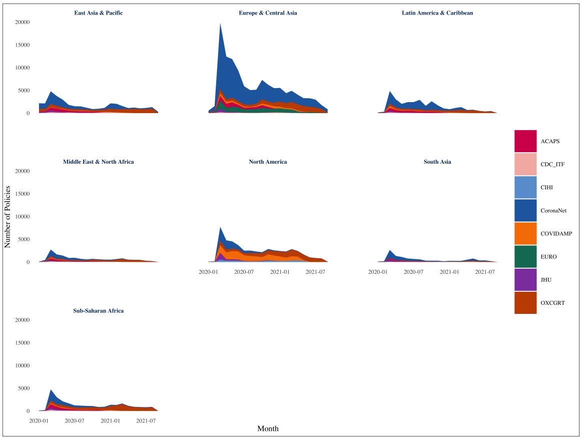

Fig. 3 Number of policies per date, grouped by region, recorded by 8 different COVID-19 PHSM tracking efforts

Data harmonization would thus lead to a dataset that is more complete and consistently coded across time and space then is currently available. Indeed, Figure 2 shows that while most datasets have fair coverage of PHSM until the summer of 2020, with data from CoronaNet being especially rich, data after this time is more limited especially for trackers that stopped data collection (e.g. HIT-COVID, ACAPS). OxCGRT meanwhile, has been able to document more policies for later months compared to other datasets.

Meanwhile Figure 3 illustrates differences in the number of policies captured across continents. Clearly, all trackers have asymmetrically focused on countries in Europe and North America. While data harmonization cannot compensate for this relative unevenness in data coverage, it can significantly improve coverage of non-European and non-North American countries in an absolute sense.

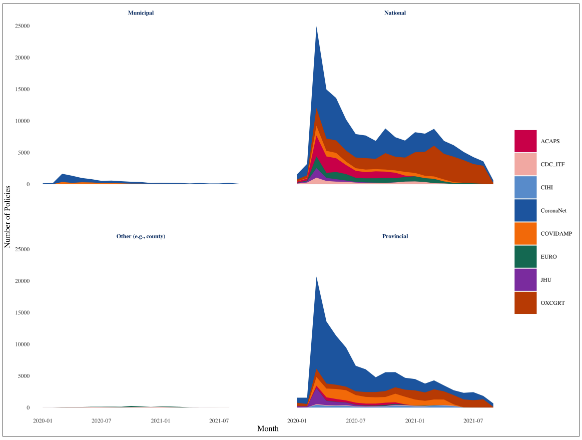

Fig. 4 Number of policies per date, grouped by the initiating level of government, recorded by 8 different COVID-19 PHSM tracking efforts

Moreover, as Figure 4 shows, most external datasets either focus on gathering national-level data for countries around the world or subnational data for a more limited number of countries, but rarely both. As such, data harmonization efforts will substantially improve the availability of PHSM data initiated at the national level and to some degree, the provincial level as well.

Overall, data harmonization greatly advances the completeness of PHSM data on a number of dimensions, including time, space, and administrative levels. Moreover, our data harmonization methodology also allows each policy in the external dataset to be evaluated independently, which can improve the quality of the PHSM data overall. This is all the more valuable given that while PHSM data has generally been made publicly accessible in close to real time because of the emergency nature of the pandemic, research groups have not been able to guarantee data cleanliness (see subsection 4.2). Progress on these dimensions greatly improve the research community's ability to conduct analyses on the COVID-19 pandemic which can yield results with both greater external validity and generalizability (in e.g. cross national analyses) as well as analyses that can yield outcomes with greater internal validity and with fewer potential confounders (in e.g. subnational analyses).

What can be lost from data harmonization? The main loss when harmonizing PHSM data into the CoronaNet taxonomy is with regards to measures that CoronaNet does not capture and for which, the benefit of its relative fine-grained taxonomy are moot. The most prominent of these measures are the economic ones, such as business subsidies or rental support. For measures for which there is conceptual overlap between the CoronaNet taxonomy and other taxonomies, the fact that the data were harmonized to the CoronaNet taxonomy, which by far has the most detailed taxonomy of the 8 datasets, minimizes the extent to which information was lost from the harmonization process.

Meanwhile, the benefits of data harmonization aside, there can be real scientific value when different researchers approach similar research topics with different research designs [101]. In support of this, we further make taxonomy maps between the CoronaNet taxonomy and the taxonomy of each respective dataset publicly available through our Supplementary Materials. These maps can not only help users better understand how to use different datasets, but can also provide robustness checks of COVID-19 related research and bolster the transparency and replicability of our data harmonization efforts.

What are the limits of data harmonization? While we believe that our efforts to harmonize data across 8 different datasets will provide the most complete picture possible of COVID-19 PHSMs, they will still fall short of a dataset that will reflect all COVID-19 PHSMs ever implemented. Though it is inherently impossible to assess how much data will still be missing after data harmonization is finalized — a complete dataset needs exist to make this assessment and it does not — we offer some insights as to where and why data may be incomplete. Specifically, our complete, harmonized dataset will still lack information on subnational policy making for a number of countries as well as from low state capacity governments.

Our review of projects gathering COVID-19 policies suggests that most projects focus on national level policies, limiting what data harmonization can achieve. Table B2 shows the coverage of data on subnational policy making for all datasets that we know to be in existence, using data available at the time of writing. Most datasets aside from CoronaNet do not collect subnational data and to the extent that they do, they overwhelmingly focus on the United States. Meanwhile, though the CoronaNet data does capture subnational data for some countries, given the volume of policies generated and limited resources, we are only able to capture this data for reduced time periods. However, available evidence suggests that subnational policy has taken place in many other countries beyond the ones listed in B2. Data from both Pandem[102] as well as CoronaNet's internal surveys suggest that there is subnational policy making in anywhere from 30 to 90 countries at any given point in time, as visualized in Extended Data Figures 5 and 6. Note that the CoronaNet internal surveys followed the same coding scheme as PanDem's [subvar] variable; at the time of writing, CoronaNet's internal assessment covers 98 countries for 6 quarters while Pandem's data covers 144 countries for 5 quarters, with 83 countries covered in common across both.

Meanwhile, we also identify how issues of low state capacity can make it difficult to document COVID-19 policies at all. Some problems that CoronaNet researchers have reported include:

- No announcement of policies in any official government sources: In the absence of any official government sources about a policy, research assistants must rely on media reports which can often have conflicting information about the nature or timing of a given policy. It is also not uncommon for

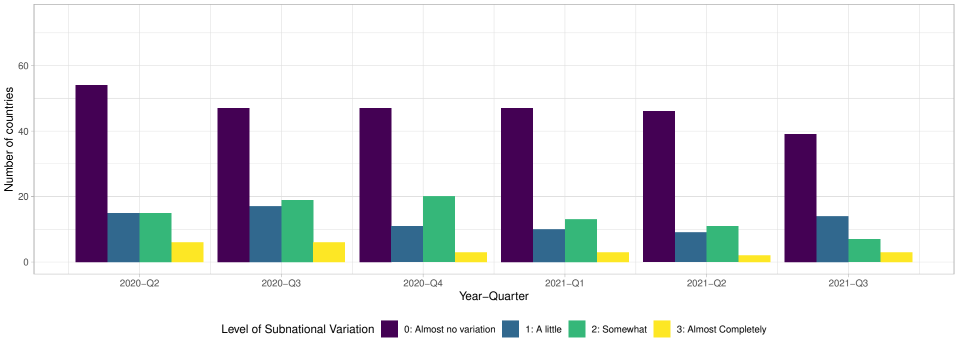

Fig. 5 Extent of policies made at the subnational level by quarter, from CoronaNet Research Project internal assesssment data.

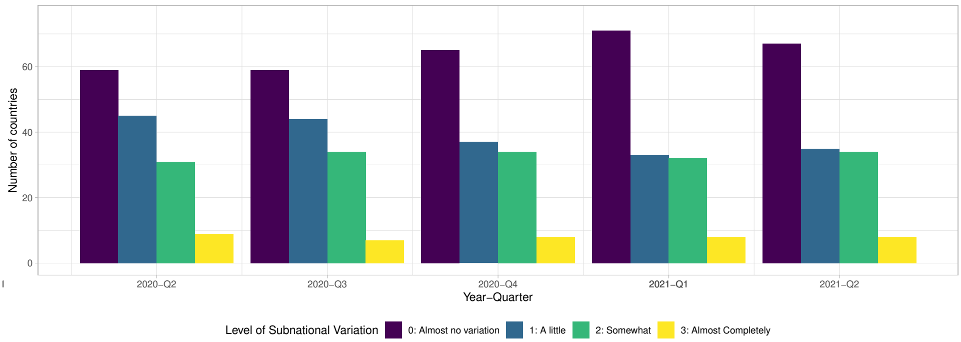

Fig. 6 Extent of policies made at the subnational level by quarter, from PanDem

governments to announce policies on social media without providing further information in the form of official government sources.

- Policies being communicated in mediums other than the Internet: In places with low internet connectivity, governments have been known to make policy announcements in non-digital forms used most prevalently by the local population e.g. radio, local news bulletins, town criers.
- NGOs and/or IOs implementing policies that are normally under the purview of governments: When governments lack the capacity to respond the COVID-19 pandemic, NGOs or IOs have been known to step in. While it is possible to capture these policies, policy trackers to date have largely focused on documenting government initiated policies.

In short, large scale data collection efforts of PHSM data have been predicated on: (i) the capacity to capture PHSM polices made at all different administrative levels (ii) the availability, access and durability of web-based documentation on PHSM policies and (iii) the assumption that governments are the primary policy responders to the COVID-19 pandemic. However, these conditions are not always present in low state capacity states. While the enormous undertaking described here will greatly advance our collective knowledge of COVID-10 PHSM policies, much more funding and support is needed to document all PHSM.

What underlying biases may need to be accounted for when harmonizing data? As we elaborate more fully in section 4.2, PHSM data is unusually challenging to harmonize because the emergency nature of the pandemic gave rise to multiple complex taxonomies and corresponding datasets that have had varying levels of quality, completeness, and underlying source material.

While we employ some automated processes to harmonize taxonomies and deduplicate data, our methodology is overwhelming reliant on the analog process of recoding external data based on the original sources found in the external data rather than relying directly on the observations available in the external data itself. In doing so, we can ensure that whatever errors might have been made in the automated taxonomy harmonization processes, which itself was adjusted to account for systemic errors in the external data (see Section 4.3.1), can be rectified manually later. Meanwhile we have also additionally vigorourly tested our automated deduplication strategies to ensure that we are biased towards keeping duplicates to be removed later manually rather than mistakenly removing observations that are not duplicates (see Section 4.3.3).

What cooperative resources are available for harmonizing data? External data partners were either co-hosts or participants in the two conferences hosted by CoronaNet: the PHSM Data Coverage Conference (February/March 2021) and the PHSM Research Outcomes Conference (October 2021). During both conferences though especially the first, trackers discussed common challenges and solutions to their data collection efforts, especially with regards to taxonomy and organization. Both the planning of the conferences and conferences themselves helped increase mutual understanding and collegiality among trackers[99]. For more information, please see https://covid19-conference.org or the shared statement written by conference participants outlining a framework for cooperation and collaboration (PHSM 2021).

Meanwhile, bilateral exchanges also played an important role in identifying and overcoming specific challenges with regards to mapping and harmonizing data for a given dataset. For instance, our ability to harmonize the CIHI dataset, was contingent on close cooperation with CIHI. Aside from explicit coordination on COVID-19 vaccines taxonomy, three volunteer researchers for CoronaNet were contracted to work on the CIHI database. This shared expertise greatly facilitated our ability to build a taxonomy map between CIHI and CoronaNet and to pilot our harmonization efforts.

Similarly, researchers from both CoronaNet and HIT-COVID were involved in building the HIT-COVID taxonomy map, which greatly facilitated the mapping process. They were also involved in piloting the data harmonization process, which also increased the speed at which it could be done. The fact that HIT-COVID and CoronaNet built their taxonomies for COVID-19 vaccine policies with mutual feedback from the other also facilitated the mapping of this particular policy type.

Meanwhile, ACAPS, COVIDAMP, and OxCGRT generously made themselves available for clarifying about confusions or misunderstandings about their respective taxonomies which helped make the mappings more accurate. However, despite repeat inquires to the WHO PHSM dataset to initiate such cooperation, we found them to be unresponsive which made the taxonomy mapping exercise with the WHO PHSM dataset comparatively difficult. Overall, we found that greater communication and cooperation between leaders of different datasets was an important intangible in facilitating the data harmonization process.

What are alternatives to data harmonization? While in this paper we concentrate on presenting our rationale and methodology for qualitatively harmonzing PHSM data, in a Kubinec (2021) we introduce a Bayesian item response model to create policy intensity scores of 6 different policy areas (general social distancing, business restrictions, school restrictions, mask usage, health monitoring and health resources) which combines data from both CoronaNet and OxCGRT [103]. As this latter paper shows, researchers should be cognizant that while statistical harmonization can be an effective form of data harmonization, the resulting indices or measures may sometimes need to be interpreted or used differently than the underlying raw data. For example, our policy intensity scores for mask wearing can be interpreted as the amount of time, resources and effort that a given policy-maker has devoted to the issue of mask restrictions in a given country compared to that in other countries. This is different from what the underlying raw data measures: whether a given mask restriction is in place or not. Researchers choosing to engage in statistical harmonization should thus provide a thorough accounting of the underlying concept that they seek to measure and a corresponding justification of why their statistical method provides a good operationlization of it.

3 Discussion

In this paper, we present a set of guidelines for helping data scientists and researchers undertsand (i) under what conditions data harmonization is possible and (ii) when it should actually be employed. Though each individual guideline is necessarily presented individually, they should be understood holistically and are not separable as such. Given that demand for data well outstrips supply and that there is a lack of general guidance on data harmonization as a potential solution to this problem, we hope that directing researchers to consider different dimensions of the harmonization process can be a valuable guide tor those considering pursuing such a strategy or those evaluating the work of others.

With regards to our own ongoing PHSM data harmonization efforts, we have shown that there are substantial gains to harmonizing PHSM data across 8 different datasets, particularly in terms of the time, spatial and administrative coverage of PHSM data. While some conceptual diversity is always lost when harmonizing data, we argue that by harmonizing PHSM data to the CoronaNet taxonomy, this issue is minimized due to the CoronaNet taxonomy's comparative richness. Data harmonization of these 8 datasets will still fall short of a complete PHSM dataset, especially for countries for which there is a great deal of subnational policy making or low state capacity but this effort nevertheless will provide the fullest picture yet of COVID-19 government policy making. Moreover, it substantially improves upon the existing WHO PHSM effort to harmonize data both in terms of scale and quality (see Appendix C). More resources would allow us to complete data harmonization more quickly, which given the ongoing nature of the COVID-19 pandemic, would be welcome. However, even if data harmonization is completed only after the pandemic is overcome, it will still present a tremendous historical resource for generations of researchers.

Our experience in data harmonization has underscored for us that the production of data may be understood not only as a mere reflection of reality, but a framing or even creation of reality. That is, by producing certain measures and not others, data can frame certain aspects of the world as more or less deserving of attention. Meanwhile creating a measure in the first place can bring forth concepts that previously did not exist in the public consciousness [66]. Harmonizing data cannot escape these dynamics and in fact invites greater scrutiny of them as it adds another layer of negotiation and complexity in terms of determining what is worthy of being measured and how to measure it. Undergirding all of this are social processes that produce data, harmonized or not, in the first place and which can have important influence on what data ultimately is or is not harmonized [63]. Though in a number of fields, researchers have developed novel platforms that aim to help facilitate data harmonization [17, 104], ultimately effective data harmonization requires researchers to identify clear goals for their harmonization process, a high level of attention to detail in designing a rigorous plan to carry out, and a strong working culture to ultimately successfully implement it. We hope that our guidelines and the lilustrative example of our experience with PHSM data harmonization can provide a roadmap for researchers embarking on similar journeys for their own research.

4 Methodology

In this section, we provide greater detail as to the methodology we employed to semi-manually harmonize data from 7 PHSM datasets into the CoronaNet taxonomy for policies implemented by governments before September 10, 2021. Our methodology can be summarized as follows:

1. Step 1: Create taxonomy maps for each external dataset and CoronaNet, which we make publicly available in the Supplementary Information. Based on these maps, we then mapped data available for each external dataset, into the CoronaNet taxonomy
2. Step 2: Perform basic cleaning and subsetting of external data to only observations clearly relevant existing CoronaNet data collection efforts.
3. Step 3: Remove a portion of duplicated policies using customized automated algorithms with respect to:
   - Duplication within each respective external dataset
   - Duplication across the different external datasets
4. Step 4: Pilot our data harmonization efforts for a select few countries (over the summer of 2021)
5. Step 5: Release the resulting curated external data to our community of volunteer research assistants to
   - Manually assess the overlap between PHSM data found in the CoronaNet dataset with that found in the ACAPS, COVIDAMP, CIHI, John Hopkins HIT-COVID, OxCGRT, the WHO EURO and CDC respectively and;
   - Manually recode data found in the external datasets that were not already in the CoronaNet dataset into the CoronaNet taxonomy.

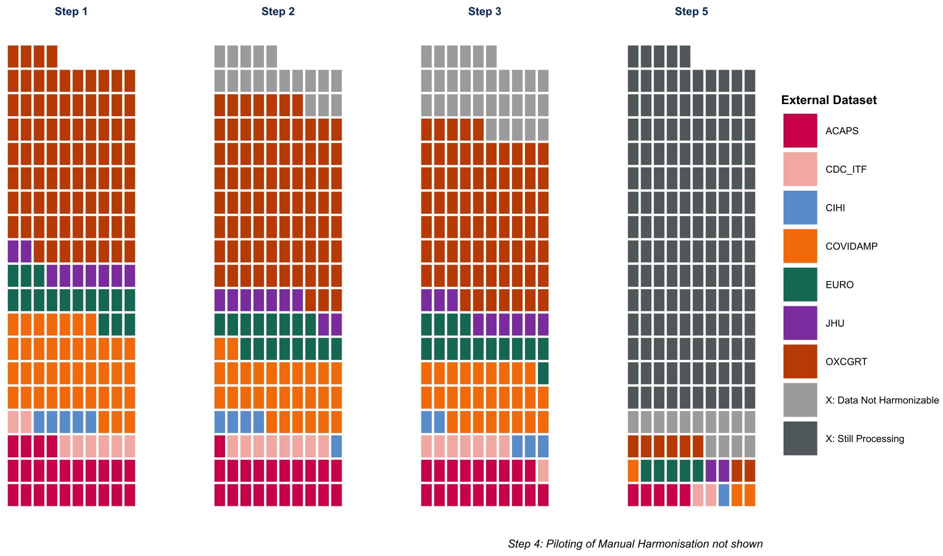

Fig. 7 PHSM Data Harmonisation Process

Our data harmonization methodology thus combines both automated and manual processes in create a more complete dataset of PHSM policies in the CoronaNet taxonomy relative to what had been originally researched by the CoronaNet Research Project alone. With this in mind, in Table 1, for each of the external datasets, we show the total amount of raw external data (Step 1), the data after observations were removed to maintain consistency with CoronaNet data collection efforts (Step 2) and the data after duplicated observations identified through automated algorithms were removed (Step 3). Manual harmonization of data (Step 5) is still ongoing but in Table 1, we provide further information on i) how much of the external data has been assessed for overlap, ii) how much of the external dataset has been assessed for harmonization iii) how much external data has been recoded into the CoronaNet taxonomy. Note subsection 4.3.7 details our pilot harmonization process (Step 4) and section 4.4 provides more detail on progress made in data harmonization. A note to the reader: unless explicitly noted, any subsequent analysis or description of the external data refers to data recorded by September 10, 2021. For a visualization of this overall process, see Figure 7.

To give context to these methodological decisions however, we first outline why we chose to harmonize these particular datasets, the challenges we faced in harmonizing data before going into greater detail as to how we implemented each of the steps laid out above. To learn about how our effort compares to a similar effort by the WHO, please see Appendix C.

Table 1 State of External Data at different steps of the data harmonization process

Dataset	Step 1 Raw Data Data (#)	Step 2 Consistency Subset (#)	Step 3 Automated Deduplication (#)	Step 5 % Overlap Assessment Completed	% Integration Assessment Completed	%Recoded into CoronaNet
All Data	180842	162991	150052	44.73	16.59	9.66
ACAPS	23926	20842	18699	63.62	22.73	13.41
CDC_ITF	7985	7405	7096	58.08	18.42	13.27
CIHI	4417	4235	4210	13.49	1.76	1.70
COVIDAMP	39332	27703	26473	23.68	8.25	4.95
EURO	15258	15071	14220	73.45	40.94	23.22
JHU	8917	8606	8142	40.47	8.83	5.47
OxCGRT	81007	79129	71212	40.33	15.20	5.63

4.1 Which datasets to harmonize?

In choosing which datasets to harmonize, we had to weigh the potential benefits of data harmonization among a number of different dimensions, including the:

- Geographical coverage of the dataset
- Temporal coverage of the dataset
- Volume of data collected by the external dataset
- Relative similarity of policy taxonomies to the CoronaNet taxonomy
- Relative capacity of external dataset partners for collaboratiion

As can be seen in Table A1 in the Appendix, we identified more than 20 datasets for consideration for harmonization. We ultimately chose datasets to harmonize that (i) aspired to world-wide geographic coverage with (ii) at least ten thousand observations in each dataset and were (iii) based on original coding of sources (as opposed to recoding of existing sources). Datasets that fit this criterion were ACAPS, COVIDAMP, HIT-COVID, OxCGRT, and CCCSL (though as explained further below, though we did initiate an effort to harmonize CCCSL data, we ultimately did not do so). One clear exception to this criteria was the inclusion of the CIHI dataset, which focuses on Canadian policies and had fewer than ten thousand policies. We decided to include the CIHI dataset for consideration because i) it already formed a substantial part of subnational data collection for other data collection efforts, including the OxCGRT dataset and ii) because of substantial cooperation and access to researchers in expertise in both the CoronaNet and CIHI taxonomies. Similarly, though the WHO Euro dataset also aims for a regional, rather than a world-wide focus, given the large number of policies in this dataset as well partial funding support that the CoronaNet Research Project receives from the EU Commission to support EU data collection, we decided to include it for harmonization. Because the WHO CDC dataset follows the same taxonomy as the WHO EURO dataset and also contains a substantial number of policies (close to 8,000), it was also included for harmonization.

4.2 Challenges of Data harmonization

Defining the scope of data to harmonize is only the first step in the harmonization process. There are numerous common challenges in harmonizing data from different datasets; we elaborate and explain how we address them in the main text.

Because of the emergency situation created by the pandemic however, on top of these normal challenges, we additionally had to deal with the fact that standards which researchers usually abide by before releasing their data were not observed. Normally, researchers generate datasets based on events that have already happened, not while they are happening. Indeed a given event needs to have run its course in order for researchers to both i) conceptualize the event being captured into a structured and logically organized taxonomy ii) estimate the amount of work needed in order to build a dataset based on this taxonomy. Moreover, because dirty data can significantly affect subsequent research findings, researchers often err on the side of caution by spending substantial additional time rigorously cleaning and validating their data before release. Researchers also have personal incentives to delay the release of a dataset given that i) they generally wish to be the first to conduct analyses on data that they themselves have collected and ii) unclean datasets can significantly negatively affect professional reputations. Meanwhile, to promote replicability and transparency about the data generating process, copies of original sources and coding decisions are often extensively documented so that other researchers may better understand how the data was generated. Due to the pandemic, however, PHSM data exceptionally were:

- Collected based on taxonomies that were developed inferentially from research group to research group while the pandemic was still ongoing.
- Released without extensive cleaning.
- Inconsistently preserved with regards to data for original raw sources.
- Absent regular updates of taxonomies.

There were a number of research-based reasons to prioritize speed over rigor. Not only did launching data collection during rather than after the pandemic help jump start early research on the pandemic, in many cases it was critical to document these policies in as close to real time as possible because primary sources of information about the pandemic can and have disappeared from the Internet over time. Though many COVID-19 trackers surely would have continued to improve their data quality, unfortunately many have had to stop their efforts because of lack of funding support. Our efforts to harmonize this external data into the CoronaNet dataset thus not only ensures that their substantial contributions can live on, but are also improved insofar as any errors in the data or discrepancies between datasets are resolved before being harmonized. This job is made more difficult however, because many trackers did not have rigorous guidelines for preserving raw sources. In what follows, we expand upon how each of these these additional challenges have affected our data harmonization efforts and methodology.

4.2.1 The challenge of harmonizing different taxonomies

Different conceptualizations of what ultimately ‘counts’ as PHSM data lies at the root of different taxonomic approaches to collecting such data. While one benefit of independently developing taxonomies is that it encourages greater flexibility and adaptability in conceptualizing the drivers and effects pandemic while simultaneously validating common themes that independently appear across taxonomies, it also makes reconciling the differences among taxonomies more challenging. A particular challenge with our data harmonization efforts is that the CoronaNet taxonomy on the whole captures more dimensions of policies than other datasets do. While this means that our data harmonization efforts will yield much more fine-grained information, mapping from a simpler taxonomy into a more complex taxonomy is also a much more challenging task than vice-versa. In what follows, we discuss what challenges we faced when mapping taxonomies for COVID-19 policy types in particular as well as for other important dimensions of COVID-19 policies.

There are at least four broad issues to consider when mapping the substance of different COVID-19 policies: (i) when taxonomies use the same or similar language to describe a policy but rely on different conceptualizations to code these policies (ii) when taxonomies have the same or similar conceptual understandings of a given event but use different taxonomical structures to capture it (iii) when taxonomies have similar but ultimately different conceptual understandings of a given event (iv) when taxonomies capture and conceptualize different events. We elaborate with examples for each of these issues in what follows:

An example of why it is important to be sensitive to semantics can be seen with regards to the term 'restrictions on internal movement.' While all datasets that use this terminology understand this to entail policies that restrict movement, some have different understandings of the phrase 'internal.' For instance, because the OxCGRT dataset generally codes policies from the perspective of the country[1], their 'C7_Restrictions on internal movement' indicator captures any restriction of movement within a country. Meanwhile, because CoronaNet codes policies from the perspective of the initiating government, its 'Internal Border Restrictions' policy type captures policies that restrict movement within the jurisdiction of a given initiating government while policies that restrict movement outside a given jurisdiction are coded as 'External Border Restrictions'. As such, if the state of California restricts its citizens from leaving the country, this would be captured in OxCGRT's 'C7_Restrictions on internal movement' indicator but would be coded as an 'External Border Restriction', not an 'Internal Border Restriction' using the CoronaNet taxonomy. Parsing out these differences can only be automated to a limited extent, especially if the given taxonomy being mapped simply does not make the same distinctions.

Meanwhile an example of how different datasets implemented different taxonomical structures to capture a similar conceptual understanding of a policy is how they captured policies related to the elderly. Different trackers took a variety of approaches to capturing such policies. OxCGRT organized its taxonomy by creating an ordinal variable, "H8_Protection of elderly people" index[2], which focuses on capturing policies specifically targeted toward the

[1]This is true for countries outside of those that the OxCGRT dataset also documents subnational data for: the United States, Canada and China. Note that it also collects subnational data for Brazil but in this case, it appears that their subnational Brazilian data is also coded at the level of the country.

[2]Note this index records"policies for protecting elderly people (as defined locally) in Long Term Care Facilities and/or the community and home setting"

elderly on an ordinal scale.[3] In contrast, the CoronaNet and COVIDAMP taxonomies documents policies toward the elderly not in its policy type variable but in a separate variable[4] which records the demographic targets of a given policy. Both datasets record whether a policy is targeted toward 'People in nursing homes/long term care facilities'. CoronaNet additionally makes it possible to document whether a policy is targeted toward 'People of a certain age' (where the ages are captured separately in a text entry) or 'People with certain health conditions' (where the health conditions are captured separately in a text entry) while COVIDAMP additionally makes it possible to document whether a policy is targeted toward 'Older adults/individuls with underlying medical conditions'. When mapping different taxonomies, these differences in taxonomical structure must additionally be taken into account.

Furthermore, taxonomies may capture similar, yet conceptually still quite distinct events which makes one to one matching between datasets difficult, if not impossible. For instance, the CIHI taxonomy's policy type of 'Travel-restrictions' does not make any distinctions between restrictions made within or outside of a given government's borders. Meanwhile, to revisit the example of policies related to the elderly, John Hopkins and the WHO PHSM taxonomy capture conceptually similar but still quite distinct categories that cannot be directly mapped onto policies related to the elderly. By developing a 'nursing homes' category, John Hopkin's taxonomy targets not the elderly per se, but the institutional settings in which they are likely to be the most vulnerable. The WHO PHSM dataset generalizes this idea in its policy category of 'Measures taken to reduce spread of COVID-19 in settings where populations reside in groups or are restrained or limited in movement or autonomy (e.g., some longer-term health care settings, seniors' residences, shelters, prisons). May include limiting visitors or outside excursions, cohorting of infected persons or green zones." This taxonomy implicitly suggests that it may be prudent to investigate not only the effects of policies on the elderly but for all those with limited mobility at the expense of easily extractable information on the elderly in particular. These cases are perhaps the most difficult to resolve as it is impossible to directly map distinctions that one taxonomy makes into other taxonomies where no such distinctions are made.

Finally, while all datasets generically sought to capture policies governments made in response to COVID-19, different datasets focused on different policy areas. For instance, virtually all external datasets have separate policy categories to capture economic or financial policies (e.g. government support of small businesses) while such policies are not systematically captured in

[3] It takes on a value of 0 if no measures are in place, 1 if 'Recommended isolation, hygiene, and visitor restriction measures in LTCFs and/or elderly people to stay at home', 2 for 'Narrow restrictions for isolation, hygiene in LTCFs, some limitations on external visitors and/or restrictions protecting elderly people at home' and 3 for 'Extensive restrictions for isolation and hygiene in LTCFs, all non-essential external visitors prohibited, and/or all elderly people required to stay at home and not leave the home with minimal exceptions, and receive no external visitors'.

[4] In CoronaNet, this is the 'target_who_gen' variable while in COVIDAMP it is the 'policy_subtarget' variable

the CoronaNet taxonomy. In these cases, such policies are thus simply not mappable.

The fact that different projects undertook such a variety of approaches in capturing such policies also underscores the idea that there is no one correct taxonomy for capturing such policies; each has its own pros and cons. For instance, aggregating all policies towards the elderly in one indicator as OxCGRT does facilitates research on how the pandemic has affected the elderly but makes it difficult to easily compare the effect of the pandemic on other vulnerable populations for example. Meanwhile though the CoronaNet and COVIDAMP approach allows more flexibility in what kind of policies toward the elderly can be captured, it also lacks the cohesiveness that having all such policies clearly labeled as being related to the elderly that OxCGRT has. With regards to data harmonization meanwhile, the sheer variety of approaches, does substantially increase the challenge of transforming this data to adhere to one taxonomy.

Indeed, despite a strong partnership with CCCSL, we decided not to harmonize data from the CCCSL dataset because of these taxonomical challenges. We found CCCSL's structure and semantics were too different from CoronaNet's, such that we estimated we would ultimately only be able to use less than half of CCCSL's observations. To illustrate by example, an observation with the CCCSL id of 4547 notes in its description that "Ski holiday returns should take special care." Such an observation would not be considered a policy in the CoronaNet taxonomy because it is does not provide specific enough information about what is meant by special care and the link for the original source of this observation is dead. While many observations do contain high quality information and descriptions, a substantial number do not contain any or only very minimal descriptive information. Combined with the difficulty in accessing original sources, we decided the relative effort required to consistently map the remaining observations into the CoronaNet taxonomy would be too high, especially considering that we are also harmonizing similar data from 7 other datasets.

So far we have only discussed the challenge of mapping taxonomies specific to policy types. However all datasets also capture additional important contextual information for understanding, analyzing and comparing government COVID-19 policies. In Table 2 below, we show the variety of approaches different datasets undertook to capture some of the most important of these dimensions including: the structure of the data (Structure), whether a given dataset captures end dates for a policy (End Dates?), has a protocol for capturing and linking updates of a policy to its original policy (Updates?), has a standardized method for documenting policies occurring at the ISO-2 (provincial) level (Location standardized at ISO-2 level), captures information about the geographic target of a policy (Geog. Target?) or captures information about the demographic target of a given policy (Demog. Target).

The dimensions highlighted in the table were chosen to underscore difficulties in harmonizing even the most basic information about a given policy. As

Table 2 Comparison of dimensions captured across different datasets

Dataset	Structure	End Dates?	Updates?	Location standardized at ISO-2 level?	Geog. Target?	Demog. Target?
CoronaNet	Event data	Yes	Yes	Yes	Yes	Yes
ACAPS	Event data	No	No	No	No	No
CIHI	Event data	Extractable through description text field	No	Yes	Yes	No
COVIDAMP	Event data	Yes	Yes	Yes	Yes	Yes
JHU	Event data	No	Yes	Yes	No	No
OxCGRT	Panel data	Yes	NA	Yes	No	No
WHO (CDC and WHO EURO)	Event data	Yes	Yes	No	No	No

the table shows, while most external datasets are formatted in event dataset format which facilitates comparability across these datasets, OxCGRT data is available only in panel format, which presents unique challenges. With regards to the data structure, in order to facilitate data harmonization, the Oxford data must be reformatted to an event format (see the Supplementary Information to access the taxonomy map). However, the panel structure also has knock-on effects on how other policy dimensions are captured, which we discuss more later in this section.

Datasets also differ with regards to how they capture the timing of a policy. Although knowing the duration of a policy is crucial for understanding its subsequent impact, if any, both ACAPS and the HIT-COVID dataset did not systematically capture information about policy end dates. Though CIHI did make this data available through its textual description, it was not available as a separate field and had to be separately extracted. When harmonizing data from these datasets then, additional work must be done to provide information on end dates.

Relatedly, datasets have also taken inconsistent approaches to capturing policy updates, if they do at all. Taxonomies that can capture such updates are arguably better equipped to capture the messiness and uncertainty of the COVID-19 policy making process (e.g. policy makers for example often lengthen or shorten the timing of a given policy in response to changing COVID-19 conditions). ACAPS and CIHI however do not separately capture and link policy updates to the original policy. Meanwhile, OxCGRT's inability to capture information on how policies may be linked is largely due to its panel dataset structure. In contrast, though both the CoronaNet and COVIDAMP taxonomy have rules for linking policies together, these differ across datasets.[5]

While all datasets use standardized taxonomy for documenting country level information about where a policy originated from, some datasets did not use a standardized taxonomy for capturing this information at the subnational ISO-2 level, in particular ACAPS and the WHO. Even when the taxonomy was standardized within a given dataset, different datasets used slightly different taxonomies at both the country and subnational levels which also necessitates further reconciliation and standardization before the data can be harmonized.

Of all the datasets processed for data harmonization, only the CoronaNet and COVIDAMP datasets capture information on both the particular geographic (e.g. country, province, city) and demographic targets (e.g. general population, asylum seekers) of a given policy. To the extent that other datasets also capture this information, it is either too broad or not standardized enough. For instance, though the various indicators in the OxCGRT data capture whether a policy overall applies to the general population or a targeted population, no further information about the specific targets is provided. Meanwhile, the WHO PHSM dataset does have a separate field which documents demographic targets but these entries are not standardized and with more than 5900 unique entries, many of which have typos (see Appendix C for more). It is thus impossible to use them for analysis without substantial additional cleaning and harmonization.

All in all, we illustrate how harmonizing different datasets can be quite the challenging exercise when considering only two taxonomies, much less 8. This is true not only with regards to taxonomy specific to the substance of COVID-19 poicies themselves but also to for additional policy dimensions like policy timing and targets.

---

[5]Coronanet links policies together if there are any changes to the original policies time duration, quantitative amount (In particular, if the i) length of quarantine has been changed ii) the amount of health resources has been changed (e.g. 100 vs 200 hospital beds iii) the number of people restricted from gathering has been changed iv) the monetary amount or number of vaccines for purchase, distribution or production has been changed or v) the time of curfew has been changed), direction of the policy (whether a policy is inbound, outbound or both), the travel mechanism (e.g. whether a policy bans flights, ships or buses), the compliance (That is, whether a policy is recommended or mandatory) or the enforcer (The institutional enforcer of a given policy (e.g. military, Department of Justice). Information on which observations are linked can be found in the 'policy_id' variable while information as to how they are linked can be found in the 'entry_type', 'update_type', 'update_level' and 'update_level_var' columns. COVIDAMP meanwhile, has separate fields to document i) whether an original policy was extended over time. This information is captured in the 'Prior row ID linked to this entry' column or ii) whether a given policy implemented at the local level has a relationship with a higher level of government. This information is captured in the fields: 'Parent policy number', 'Parent policy relationship' and 'Additional notes for parent/child relationship'

4.2.2 The challenge of harmonizing dirty data

Dirty data refers to data that is miscoded according to a given taxonomy. In our investigation of the cleanliness of different datasets, we distinguish between policies that are (i) inaccurately coded relative to a given taxonomy or (ii) incomplete or missing. Harmonizing dirty data would be challenging even if taxonomies across datasets were the same; these problems are only compounded when the taxonomies are different. Unfortunately, because of the emergency situation of the pandemic, all datasets (both external and CoronaNet) suffer from problems with dirty data.

For instance, although within the ACAPS taxonomy, all policies related to a curfew should theoretically be coded as 'Movement Restrictions' and 'Curfew' in their 'category' and 'measure' fields respectively, text analysis of the descriptions accompanying these observationssuggests that curfew policies were mistakenly coded in at least 8 other policy categories[6]. Data can also be dirty for other important policy dimensions, e.g. the start dates of a given policies. Many governments simply maintain websites where they note the most current policies without detailed information as to when the policy started or will end[7] and in some cases, coders will simply note the date that they accessed the policy as the start date as opposed to the true start date. That these types of issues were found across all datasets is no surprise given the unusual circumstances that such data are collected and released. Nevertheless, they can pose immense challenges for data harmonizing as blindly automating the harmonization of such data risks compounding the original errors in the data.

While it is difficult to quantify the relative cleanliness of different datasets (and thus, how much of an issue it poses to data harmonization), we provide some sense of the relative data quality of different datasets with regards to the quality of their textual descriptions in Table 3 below. Good textual descriptions of a given policy are crucial for helping users understand what policies a given dataset is actually documenting and organizing. We first try to get a sense of how informative these descriptions are by counting the average number of characters each description has per dataset (Description Length (Average)), how many descriptions have less than 50 characters (Desciptions with less than 50 characters (Total))[8] and how many observations have no descriptions at all (Missing Descriptions (Total)). The table shows that textual descriptions from

[6]Policies relevant to curfews which should have been coded as 'Movement Restrictions - Curfew' were also found as being coded under 'Lockdown – Partial Lockdown', 'State of Emergency', 'Movement Restrictions - Border closures', 'Movement restrictions = Surveillance and monitoring', 'Movement Restrictions - Domestic travel restrictions', 'Public health measures - Isolation and quarantine policies', 'Governance and socio-economic measures - Emergency administrative structures activited or established'; 'Governance and socio-economic measures - Military deployment'

[7]See this archived https://web.archive.org/web/20210621102402/https://covid19.gov.lv/en/support-society/how-behave-safely/covid-19-control-measures for an example from the Latvian government.

[8]Generally, descriptions with less than 50 characters contain only limited information about a given policy. Examples of descriptions of with less than 50 characters include: "Albania banned all flights to and from the UK." (CoronaNet); "Blida extended until at least 19. april 2020" (ACAPS), "Lockdown extended. Lockdown extended" (CDC_ITF), "The state of emergency in WA has been extended." (COVIDAMP), "Delay of international flights have been extended" (EURO), "Extends school closures until March 16" (JHU), "orders extended until April 30" (OxCGRT).

the ACAPS dataset have on average the least number of characters compared to other datasets, with more than two thousand having descriptions of less than 50 characters and more than 100 having no description at all. While OxCGRT has the third highest average description length, it also has the most number of descriptions with less than 50 characters. Meanwhile John Hopkins has the most number of policies without any description, at more than 1600.

With regards to the content of the descriptions, only the CoronaNet and CIHI databases appear standardize what should be included in this textual description (see 'Description Standardized?' column in Table 3).[9] For each dataset, we randomly selected one description that accorded to the average description length for that dataset to illustrate what kind of information could be gleaned from them in the 'Example of Average Description' column in Table 3. These descriptions suggest that while the CoronaNet and CIHI descriptions include information about the date the policy is enacted and the policy initiator, this information is not always reliably made available for descriptions from other datasets. While this information is generally also subsequently captured in separate variable fields, having detailed textual descriptions are important for helping to adjudicate whether the subsequent coding of these separate policy dimensions is accurate or not.

While it would be useful to have a similar quality assessment for other important variables of each dataset, as far as we know, only the CoronaNet dataset provides an empirical assessment of the quality of its data. CoronaNet implements a multiple validation scheme in which it samples 10% of its raw sources for three independent coders to separately code. If 2 out of 3 of the coders document a policy in the same way, then it is still considered valid. Though data validation is still ongoing, preliminary data suggests that there is high inter-coder reliability, around 80%, for how its policy type variable is coded, which is generally accepted to be indicative of high inter coder reliability [105–107] An exception to the generally high validity of the policy type variable is the relatively poor coder interreliability for the 'Health Testing' and 'Health Monitoring' categories. This is likely related to changes in the CoronaNet taxonomy, which while important to make to better adapt to the changing policy-making environment, also increases the dirtiness of the data. A full accounting of taxonomy changes can be found here. Other external datasets likely also have faced similar issues which subsequently affect their data quality although we were unable to locate public documentation of these changes.[10]

[9] Coders for CoronaNet are instructed to include the following information in their textual descriptions: (i) the name of the country from which a policy originates (ii) the date the policy is supposed to take effect (iii) information about the 'type' of policy (iii) if applicable, the country or region that a policy is targeted toward (iv) if applicable, the type of people or resources a policy is targeted towards (vi) if applicable, when a policy is slated to end. The CIHI descriptions take a regularized format in which the government initiating the policy is clearly specified, the policy type is described and the end date of a given policy is recorded if applicable. With regards to the other datasets, we were unable to find any documentation that suggested that text descriptions should follow a standardized format nor were we able to find evidence of one by reading through a sample of the text descriptions themselves.

[10] Note, if there were any taxonomy changes for OxCGRT or JHU HIT-COVID, they are likely recoverable from their git repository histories (available respectively https://github.com/OxCGRT/covid-policy-tracker/tree/master/documentation and

Table 3 Assessment of Textual Descriptions

	Description Length (Average)	Descriptions with less than 50 characters (Total)	Missing Descriptions (Total)	Description Standardized?	Example of Average Description
CoronaNet	348	203	0	Yes	On 2 April 2020, the Australian Capital Territory government announced the construction of a temporary COVID-19 Emergency Department through a partnership with local healthcare provider Aspen Medical. The package also provides funding to our hospitals to purchase more equipment and more personal protective equipment for our nurses and doctors
ACAPS	172	2147	118	No	IKR extended until at least 23.4.2020 All movements in the Kurdistan Region are banned between the hours of midnight and 0600, except for security officials and ambulances.
CDC_ITF	537	126	47	No	Curfew extented. Guatemala s President on 7 June announced he extended curfew and measures imposed to contain the coronavirus pandemic for one more week. The partial curfew, in force from 22 March, will continue between 18H00 and 05H00 starting 8 June, and transit between provinces remains prohibited.Since March, measures such as the suspension of public transport and classes have been in force in Guatemala. Social, religious, sporting and cultural activities are also prohibited. Air, sea and land borders are closed to foreigners.
CIHI	254	0	0	Yes	Who: Government of Yukon What: Updated school health and safety guidelines for K-12 to reduce the requirement for 2-metre distancing between students in the classroom and to make masks mandatory in common areas outside of the classroom. Effective until
COVIDAMP	227	950	0	No	Extension: Indoor events with over 500 spectators/ attendees cannot exceed 50% of the venue's room's capacity. Indoor events attended by non-students must adhere to social distancing requirements and face covering requirements
EURO	297	509	662	No	Extension through 1 March 2021 Taking into account the analysis of the current situation and the epidemiological situation in the UK, Italy, Germany, Denmark, Austria, Australia, the Netherlands and South Africa, extension of suspension of international flights entry, exit and transit flights .
JHU	230	635	1622	No	In case the person has any history in the last 14 days and the person is symptomatic as per the case definition of COVID 19, the person must be isolated in a hospital as per protocol and will be tested for COVID19 as per protocol.
OxCGRT	329	4265	0	No	The two week self-quarantine has been lifted for those traveling to New Hampshire from surrounding New England States (Maine, Vermont, Massachusetts, Connecticut, Rhode Island). Those traveling to New Hampshire from non-New England states for an extended period of time are still asked to self-quarantine for a two-week period.

The closest similar information that other datasets provide on data quality are with regards to their cleaning procedures. More information on the steps other datasets took to ensure data quality can be found in their respective

https://github.com/HopkinsIDD/hit-covid/tree/master/documentation) but we could find no explicit documentation of any such changes.

documentation [11]. Given that a number of external trackers had to stop their data collection efforts as well as the relatively high level of data quality of the CoronaNet data for the dimensions that we have information on, we can cautiously infer that harmonizing external data to the CoronaNet dataset will help improve the quality of the subsequently harmonized data.

Table 1: Inter-Coder Reliability Measures for On-Going Validation (round 1)

Policy	(n)	Percentage Agreement	Cohen's Kappa (k)
Curfew	19	100	1
Hygiene	2	100	1
Declaration of National Emergency	54	96.3	0.96
Restrictions of Mass Gatherings	92	94.6	0.94
External Border Restrictions	112	94.6	0.94
Closure and Regulation of Schools	48	93.8	0.93
Restriction and Regulation of Businesses	63	90.5	0.9
Lockdown	28	85.7	0.85
Restriction and Regulation of Government Services	63	84.1	0.83
Health Resources	91	83.5	0.82
Quarantine	47	83	0.82
Internal Border Restrictions	22	81.8	0.81
Social Distancing	37	73	0.71
New Task Force, Bureau or Administrative Configuration	22	72.7	0.71
Public Awareness Measures	51	68.6	0.67
Health Testing	17	52.9	0.5
Health Monitoring	13	23.1	0.18

Summary Inter-coder Reliability Scores	
Percentage Agreement	81.07
Cohen's Kappa	0.8
Krippendorff's alpha	0.82

Fig. 8 Caption

Data completeness is also an important factor in a dataset's overall quality. The more complete a datatset is, the more accurate subsequent analyses based on this data can be. All datasets harmonized here are by definition incomplete given that they made their datasets publicly available while their data collection efforts are ongoing. This issue is compounded by the fact that many datasets have had to stop or substantially slow their data collection efforts, particularly ACAPS, JHU, CDC_ITF and COVIDAMP. Because policies often continue past the lifetime of the external group collecting the data itself, issues of data incompleteness only grow over time for datasets that stop collecting data. While a full assessment of the completeness of each dataset is not possible (one would need a perfectly complete dataset in order to judget the completeness of other datasets) in Table 4 below, we provide some sense of each datasets relative completeness by assessing how many policies lack end dates, the average start and end dates of policies and the last submission date of a given policy.

Based on this table, following ACAPS and JHU which do not collect information on any end dates at all, CIHI has the highest percentage of policies missing while CoronaNet has the most number of missing information on end dates. Meanwhile though the average start date and end dates for all datasets center around the last half of 2020, the earliest average start dates are the ACAPS, JHU and CoronaNet datasets, with Oxford, CIHI and the EURO

[11] See the following for their respective documentation: CoronaNet [23]; ACAPS [20] (OxCGRT [25], JHU [24] and the WHO PHSM[108]). Note, no documentation on data quality procedures were found for CIHI

datasets being relatively farther along. Meanwhile the CDC_ITF, CoronaNet and EURO datasets have the earliest average end dates while OxCGRT, CIHI and COVIDAMP have the latest average end dates. The last submission date (relative to September 2021) underscores that ACAPS, JHU and the CDC_ITF, have stopped data collection while COVIDAMP has significantly slowed its efforts. At the time of writing, only the OxCGRT and CoronaNet datasets appear to be actively collecting PHSM data. Overall then, this table suggests then that data harmonization may substantially raise the data completeness of the CoronaNet dataset.

Table 4 Assessment of Data Completeness

dataset	Missing End Dates (Total)	Missing End Dates (%)	Start Date (Average)	End Date (Average)	Last Submission Date
CoronaNet	31672	35.3	2020-08-16	2020-11-03	2023-02-08
ACAPS	18699	100.00	2020-06-13	NA	2020-12-08
CDC_ITF	1384	19.50	2020-08-16	2020-09-06	2021-12-05†
CIHI	3483	82.73	2020-10-06	2020-11-15	2021-08-12‡
COVIDAMP	4970	18.77	2020-08-28	2020-10-09	2021-09-21‡
EURO	2137	15.03	2020-09-10	2020-09-24	2021-06-21*
JHU	8142	100.00	2020-04-15	NA	2020-12-15
OXCGRT	2558	3.59	2020-12-27	2021-01-07	2021-09-21‡

† There is likely some issues with this variable for the CDC_ITF as we retrieved the WHO PHSM data on 2021-09-10 so the last submission date could not have been after this date.

*There was only one submission date recorded for the EURO dataset (2020-04-01), however according to the WHO PHSM website, the last update was 2021-06-21

‡These datasets do not separately record this information. Instead the date the dataset was retrieved or the self-reported last submission date for the entire dataset was used.

As outlined above, all datasets considered in this paper suffer in various degrees from problems of miscoded or missing or incomplete data. However, though dirty data substantially raises the complexity and challenge of accurate data harmonization, the data harmonization process can also improve the quality of such data, which we will discuss in more detail later on.

4.2.3 The challenge harmonizing data with missing information on original sources

Given both the challenges in harmonizing (i) data coded from multiple different taxonomies as well as (ii) dirty data, it is essential to have access to the original raw source of data for a given policy to harmonize the data accurately. Reference to the original source used to code the policy is necessary for instance, to resolve any confusion or disagreement about a given coding decision.

In Table 5, we illustrate differences among each dataset in terms of how they make source data available (Source Data) and how many observations do not have any source data attached to it (Missing Links (Total)). The table also shows, relative to external data that has already been assessed for harmonization, the percentage of observations that have been found to be based on sources with dead links for which corroborating information was unable to be found after a good faith effort (Unrecoverable links (Percent of total integrated)) as well as the percentage of observations which have been found to be based on dead links but for which corroborating information was subsequently recovered (Recovered Links (Percent of total integrated)).

We find that while all datasets provide reference to the URL links used to code a given policy, only CoronaNet, COVIDAMP and HIT-COVID also provide links to static PDFs of raw sources which ensure that this information will continue to be available in the future.[12] With regards to the extent to which a given observation is missing a URL or PDF link to its raw source, the WHO EURO and OxCGRT datasets have the most number of missing links while this is not an issue for the CoronaNet and CIHI datasets. Meanwhile, based on the amount of external data that has been harmonized thus far, around 10.2% of the external data is based on links that were dead which were not possible to recover corroborating information for. This was a particular problem for the WHO EURO and WHO CDC_ITF datasets though not an issue for the CIHI or COVIDAMP datasets. Meanwhile around 4.7% of the external dataset assessed for harmonization to date, were based on dead links but for which it was possible to recover corroborating information. Because these data points are recoded using the CoronaNet taxonomy, PDFs of these recovered links were also uploaded, ensuring that they will continue to be preserved for future records. Observations coded by the WHO EURO database were found to be particularly recoverable. Note that we do not make an assessment for unrecoverable or recovered links for CoronaNet because the CoronaNet methodology ensures that PDFs are always saved (the data is collected via a

[12] Note however, COVIDAMP has around 150+ observations which only have a URL link and no PDF link attached to it while early observations entered into the JHU HIT-COVID dataset also only have URL links with no accompanying PDF links

survey and uploading a PDF is mandatory for a policy response to be considered valid). All told, at least 17% of the external data (3% of the external data have no links, 10.2% of the data are based on links with unrecoverable information and 4.7% of the data are based on links with recoverable information) are based on data with some issues with regards to their original sources, which only increases the challenge of smoothly harmonizing information from different datasets.

Table 5 Assessment of Raw Sources

Dataset	Source Data	Missing Links (Total)	Unrecoverable links (Percent of total integrated)	Recovered Links (Percent of total integrated)
CoronaNet	URL and PDF links	0	NA	NA
ACAPS	URL links	26	0.012	0.05
CDC_ITF	URL links	24	0.08	0.06
CIHI	URL links	0	0.00	0.01
COVIDAMP	URL and PDF links	8	0.02	0.05
EURO	URL links	3011	0.20	0.07
JHU	URL links for early data; URL links and PDF links for later data	14	0.05	0.05
OxCGRT	URL links	1385	0.07	0.03

4.3 COVID-19 PHSM Harmonization Methodology

The challenges posed by harmonizing multiple complex taxonomies of dirty data based on inconsistently preserved original sources led us to the conclusion that ultimately, only manual harmonization would allow us to harmonize data from different PHSM trackers in a way that would ensure high data quality and validity. Given the sheer number of policies in the external dataset however, to the extent possible, we sought to support these manual harmonization efforts with automated tools, specifically with automated taxonomy mappings and initial data dedpulication efforts. In what follows, we outline in greater detail each of these different steps.

4.3.1 Step 1. Making Automated Taxonomy Maps

Given the variety and complexity of approaches that different groups have taken to document PHSM policies, asking research assistants to not only become experts in one taxonomy but multiple taxonomies would have been unfeasible. Instead, we created maps between the CoronaNet taxonomy and other datasets so that all datasets could be understood in the CoronaNet taxonomy for a number of principal fields including:

- Policy timing
  - The start date of the policy
  - When available, the end date of policy
- Policy initiator
  - The country from which a policy is initiated from
  - When available, the ISO-2 level region from which a policy is initiated from
- Policy Type
  - Broad policy type
  - When possible, the policy sub type
- Sources/URLs
  - URL links
  - When available, links of original pdfs
- Textual description

When possible, other fields, such as the geographic and demographic targets, are also matched. As outlined in section 4.2.1, because of conceptual and organizational differences across different taxonomies, one to one mappings were not always possible especially with regards to the substance of COVID-19 policies. In such cases, one to two or one to three mappings were suggested. For the COVIDAMP and WHO PHSM mappings (relevant for the WHO EURO and WHO CDC datasets), we also employed machine learning models to predict the most likely policy type an observations was likely to be in the CoronaNet taxonomy based on the textual description of the policy. Both because one to one mappings based on the taxonomies themselves were often not possible and because of issues with dirty data, in some cases, the mappings were often adjusted to so that they were based not only on the formal taxonomy but also on when certain key words were used in the dataset. For example, though policies originally coded in the WHO taxonomy of 'Social and physical distancing measures (Category) - Domestic Travel (Sub-Category) - Closing internal land borders (Measure)' might reasonably map onto CoronaNet's 'Internal Border Restriction' policy type, when the word 'quarantine' appears in the text description of such policies, we reclassify them in the taxonomy map as a 'Quarantine' policy instead. As such, these taxonomy mappings are not always based strictly on how different policy types theoretically should map onto each other, but attempt to account for mistakes and miscodings in the external data to create the best mapping possible between the existing data and the CoronaNet datasets. In this first automated step, our aim was to ensure that *most* mappings were correctly mapped but did not take pains to make sure that *every* mapping was correctly mapped, because, as we explain later on, each observation was ultimately assessed and evaluated for harmonization by human coders who are better equipped to make these more fine-grained and nuanced judgements.

As part of this mapping exercise, in order to keep track of the original dataset that each observation came from, we also ensured that each record was associated with its own unique identifier (unique_id). In some cases, the data had to be somewhat reformatted in such a way that also impacted how the unique_id assigned by the original dataset was formatted though this was always done so in a way that makes it possible to trace back to the original dataset.[13] In the case of OxCGRT, no unique identifiers are provided in the original dataset and in this case we generate them using a combination of the policy indicator, date, country and where applicable, province.

Please see the Supplementary Information for more information about how to access the specific taxonomy mappings we created between CoronaNet and other datasets.

4.3.2 Step 2. Basic cleaning and subsetting of external data

With the help of the taxonomy maps, we were able to roughly transform the external datasets into the CoronaNet taxonomy. Before moving forward with manual data harmonization, we first implemented some basic cleaning and subsetting of the data. Because, as discussed in subsection 4.2.1, most datasets do not use a consistent reference for identifying policies originating from the ISO-2 provincial level, we created code to clean these text strings up as much as possible. Given the sheer number of observations that needed such cleaning, we could not ensure full standardization for these text strings. However, we took pains to ensure that the 430+ provinces for which CoronaNet is systematically seeking to collect subnational data for were consistently documented in the external data[14].

Next we subset the external data to exclude regions that CoronaNet is currently not collecting data for. In particular, we excluded from our harmonization efforts observations from the COVIDAMP dataset documented at the county or tribal level in the United States as well as observations for Greenland, the United States Virgin Islands and Guam. In addition, we also subset the

[13] For example, in the JHU dataset border restrictions for people leaving or entering a country are coded in separate observations. However, in the CoronaNet dataset, if a policy for restricting both entry and exit to or from the same countr(ies) on the same date, they are coded as one observation. In this case, the JHU data is collapsed to fit into one observation and the unique identifier is also collapsed such that two or more of the original unique identifiers are collapsed into one when they are mapped to the CoronaNet taxonomy.

[14] Specifically, these are subnational provinces for the following countries: Brazil, China, Canada, France, Germany, India, Italy, Japan, Nigeria, Russia, Spain, Switzerland, and the United States

external dataset to exclude policy types that CoronaNet is currently not collecting data for, in particular economic or financial measures taken in response to the pandemic.

4.3.3 Step 3. Automated Deduplication

After making taxonomy maps for each external dataset to the CoronaNet taxonomy and conducting some basic cleaning of the data, we also took steps to deduplicate the data using automatic methods as much as possible. Deduplication was assessed along three criteria: i) duplicates within each external dataset ii) duplicates across the external datasets and iii) duplicates between the CoronaNet and external datasets. We outline the steps we took to assess the level of duplication along each of these criteria and when possible, to remove duplicates accordingly. All in all, we took a conservative approach in our automated deduplication efforts insofar as we rather left many potential duplicates in the dataset rather than removed too many policies which may have not been duplicates.

4.3.4 Step 3a. Deduplication within External Datasets

Given the sheer amount of data collected and coordination needed to collect such data, it is not surprising that there is some duplication within datasets. Duplicates can occur for a number of reasons including (i) structural differences between taxonomies (ii) the lack of one to one matching between taxonomies (e.g. a policy that may be coded as several policies in one taxonomy may only be coded as one policy in the CoronaNet taxonomy) (iii) coder error.

We first needed to deal specifically with duplication that occurs as a result OxCGRT's method of collecting data to fit a panel data. In particular, OxCGRT coders are generally instructed to provide an assessment of whether a policy was in place or not for each given day that they are either recording the policy or for which they have evidence for a policy being in place or not. For instance, if a coder finds that the same policy has been in place over several weeks, the same textual description may be copied and pasted into the notes section for each day that the coder happened to review the status of policy-making for that indicator, even if the numerical indicator itself does not change. When initially extracting and reshaping the OxCGRT data into an event dataset format, each textual description is initially retained, even though it may not contain new information. To deal with this, we built a custom function to identify policies that repeated the exact same description, keeping the 'latest' instance of the policy description and removing earlier ones (see the OxCGRT-CoronaNet taxonomy map available through the Supplemntary Information for more detail).

We also needed to implement a custom procedure to deal with a related practice of documenting 'no change' in a policy indicator which was unique to OxCGRT's methodology for documenting policies. Specifically, when an OxCGRT coder does not identify any change in a policy indicator, it is customary for the coder to note something to the effect of 'No change' in the textual description for that particular day. This information can be extremely valuable if one desires to know the status of a given indicator in the 'present' as it allows researchers to distinguish whether there was truly no change in government policy makers or whether there was simply no one actively documenting government policy making for a given region and indicator. As the present becomes the past however, this information becomes less useful. For instance, while the value of knowing that there was 'no change' in a given indicator 'today' is quite high, knowing that there was 'no change' for a given indicator in e.g. March 2020 is not very informative especially if there was subsequently a lot of policy making activity for that indicator. Given that we initially retained each textual description from the OxCGRT data when transforming it from a panel to event dataset format, our initial efforts created an OxCGRT event dataset format that was filled with observations that documented variations of the sentiment 'No change.' Because the CoronaNet taxonomy does not document when there are no policy changes, to the extent possible then, we sought to remove such observations from the OxCGRT dataset. The difficulty in doing so was compounded by the fact that (i) there appears to be no standard language that OxCGRT coders follow in communicating that a policy had no change (ii) not infrequently, a textual description will start by noting that there has been no change to a policy, but will then subsequently provide a long and detailed description of the policy. In these cases, it is unclear whether there actually was no change to a policy and the coder is simply noting what the policy was or if there was no change to the policy that could be captured by the OxCGRT taxonomy, but there were actually some changes made by the government and the coder is documenting them qualitatively in the text. To deal with the former issue, we looked through hundreds of OxCGRT policies to try to identify as many phrases that conveyed the sentiment 'no change' as possible. To deal with the latter issue, we did not remove observations over a certain character limit even when they noted that there was 'no change' in case there actually was a substantive change that could be captured in the CoronaNet taxonomy. These choices were consistent with our general conservative approach towards automated deduplication.

Following this specialized deduplication for the OxCGRT dataset, we then sought to identify duplicates within each dataset more generally. We experimented with identifying policies that had identical values for a variety of different policies and ultimately found the following set of variables as being able to accurately identify a large number of duplicates:

- 'description': records the textual description used to describe each observation[15]
- 'country': records the country that a policy originates from, where the list of countries are standardized,

[15] Note, for the purposes of deduplication, the descriptions were stripped of punctuation and special characters and transformed to all lower cased letters in order to decrease the likelihood that stray superfluous symbols would prevent the identification of duplicates.

- 'province': records the province that a policy originates from, where the list of policies are semi-standardized (see Section 4.3.2 for more info),
- 'link': records the URL link used as the raw source of information for a given policy

Theoretically, we believed that the likelihood of identifying true duplicates with the above variable fields are quite high given that all descriptions are all written in free form and that URL links can act as fairly robust unique identifiers. With this set of variables, we identified 6955 policies that were duplicated.[16] To check this assumption, we sampled 100 groups of policies that were found to be duplicates, (which was equivalent to 393 total observations), and through manual investigation, found that 99 of these groupings were indeed duplicates, for an accuracty of 99%. We further manually checked groups of policies that were identified as having particularly high number of duplicates (7 or more, the maximum being 19) and found that our criterion accurately identified these groups of policies as having duplicates. Because this automated deduplication method proved to be quite accurate, we subsequently used this criterion to remove likely duplicates within each dataset. We show the distribution of policies we found to be duplicates according to this criterion in Table 6.

Table 6 Assessment of duplicates within datasets

ACAPS	CDC_ITF	CIHI	COVIDAMP[17]	EURO[18]	JHU	OxCGRT[19]
246	45	1	437	304	373	5549

As can be seen, we identified a particularly high number of duplicates within the OxCGRT dataset. This is consistent with our knowledge that

[16]Note that we excluded from this procedure, policies that had the textual description 'Extension' or 'extend' in their descriptions. As part of our investigation, we found that it was common for coders to copy and paste the same description with this word every time a policy was extended in time and as such we would have inaccurately removed many policies had we not excluded such observations from our deduplication efforts.

[17]We found that many duplicates for COVIDAMP were due to the fact in some cases, when the same policy was used to target different cities, the same description, country, province and link were used. Because the CoronaNet taxonomy documents different target cities within one observation instead of multiple observations, when deduplicating this data, the code was adjusted such that the information on the target cities in the COVIDAMP data was aggregated and preserved in the observation that was kept while the other observations were discarded.

[18]We found that many duplicates for the WHO EURO data was due to the fact that policy sub types were recorded as separate observations while using the same description, country, province and link. Since initial taxonomy mapping exercise was often only precise to the policy type, not the policy sub type, it was fine to discard these observations as duplicates given that the core information would still be retained and manually recoded more precisely in terms of policy sub types by human coders later on.

[19]Duplicates in the OxCGRT data were found for a number of reasons including (i) often, coders would use the same description and links across separate OxCGRT indicators for a given country and start date. Because these policy types were already mapped as a 'one to many' policy type mapping, it was fine to discard these policies as duplicates. (ii) A fair number of policies are 'no change' policies that were not previously caught in the custom procedure described to capture such policies (iii) A fair number of descriptions only had a link in the description.

duplication is a particular problem with OxCGRT data because of their methodology for data collection as well as what we knew to be a conservative approach in our custom method of deduplicating OxCGRT data.

4.3.5 Step 3b. Deduplication across External Datasets

The data was also evaluated for duplicates across datasets. Data duplication across datasets happens because different policy trackers have only coordinated their work in collecting PHSM data to a limited extent. As such, the same policy may be independently documented by coders in different datasets. While this is desirable from the point of view of data validation, it is a hindrance from the point of view of data harmonization.

As a first step in deduplicating data across datasets, we were able to remove a number of observations that were by definition duplicates. Specifically, since the OxCGRT subnational data for Canada is based in large part on the data collected by CIHI, we removed OxCGRT data for Canada from our dataset and instead chose to prioritize the more fine-grained version of the data documented by the CIHI dataset. Note that the full WHO PHSM dataset actually includes data from ACAPS, John Hopkins and OxCGRT. These observations were removed from the dataset as well following similar a logic. That is, it seemed likely that a direct translation from e.g. the ACAPS/JHU to CoronaNet taxonomy would lead to fewer errors than using the version of the data that first translates ACAPS/JHU to the WHO PHSM taxonomy and then to the CoronaNet taxonomy. Second, it further allows us to maintain evaluate the full ACAPS and JHU datasets; whereas in the WHO PHSM dataset the ACAPS data has already been deduplicated according to the WHO PHSM taxonomy.

Following this, we then experimented with identifying duplicates across datasets more generally. In addition to exploring which set of variables most reliably identified groups of true duplicates (as we did for identifying duplicates within datasets), when duplicating across datasets, we further had to decide from which dataset observations should be retained when duplicates were found. With regards the former, we found that identifying duplicates based on the following variables[20] to yield the most accurate results :

- type : records the broad policy area of a given COVID-19 policy. E.g. a policy related to schools will be coded as 'Closure and Regulation of Schools' type.
- type_sub_cat: The specific policy area of a given COVID-19 policy. This is hierarchically determined such that only certain type and type_sub_cat combinations can go together. E.g. a polcy related to primary schools will have a sub type of 'Primary Schools' and will by definition have a policy type of 'Closure and Regulation of Schools' .

[20]We considered other variables but found that they were not adequate because they were not broadly collected across different external datasets. E.g. enforcer is only collected by CIHI; target_country, target_province is only collected by COVIDAMP; target_direction, institution_status is only collected by JHU; type_mass_gathering is only collected by WHO EURO and WHO CDC,date_announced is only collected by COVIDAMP and CIHI

- country: records the country that a policy originates from, where the list of countries are standardized,
- province : records the province that a policy originates from, where the list of policies are semi-standardized (see Section 4.3.2 for more info),
- target_who_what : if applicable, records the citizenship (citizen or non-citizen) or travel status (traveller or resident) which a given policy is targeted toward
- date_start : records the start date of a given policy

Meanwhile, with regards to the issue of what observations we should ultimately retain when duplicates were identified, we developed a protocol for prioritizing given datasets based on both our qualitative experience working and transforming each dataset during the taxonomy mapping exercise in Step 1 as well as the quantitative assessment of the data quality of each dataset which we outlined in Section 4.2.2. When there was only one duplicate identified for a given observation, we chose to retain information from the dataset that had the most number of characters in its textual description of that observation. When more than one duplicate was identified per grouping however, we developed the following protocol for prioritizing which observation to retain:

- Priority 1: For Canadian data, CIHI is prioritized first because this dataset specializes in collecting Canadian data.
- Priority 2: COVIDAMP data is prioritized second for all data except for Canadian data based on both our qualitative and quantitative assessment of COVDIAMP data quality. Based on our experience creating the taxonomy map between COVIDAMP and CoronaNet, we found that COVIDAMP's taxonomy was very similar to the CoronaNet taxonomy, mitigating the challenge of taxonomy mapping and potential attendant errors. In terms of our quantitative assessment of COVIDAMP data quality, we found it to be relatively high quality insofar as there are very few missing links and relatively high quality of textual descriptions. Note however, that COVIDAMP only collects data for 64 sovereign countries (while 95 are available in its dataset, these include policies for United States Native American tribes).
- Priority 3: WHO CDC_ITF and WHO EURO is prioritized third for all data except for Canadian data. These data were prioritized together because they have already been mutually assessed for deduplication within the WHO PHSM dataset. In terms of data quality, the CDC_ITF data appears to have higher quality descriptions compared to OxCGRT, ACAPS and JHU based on the average length of the description, the number of descriptions with less than 50 characters, while the WHO EURO data appears to have higher quality descriptions than ACAPS and JHU based on the average description length and higher quality descriptions than ACAPS, JHU and OxCGRT based on the number of descriptions with less than 50 characters. Meanwhile, both datasets also have fewer missing end dates then ACAPS, JHU and OxCGRT.
- Priority 4: OxCGRT data is prioritized fourth for all data except for Canadian data because the OxCGRT data has some information on end dates

and based on our qualitative assessment, has more informative descriptions of policies than the JHU and ACAPS. This is supported quantitatively as well given that OxCGRT descriptions are on average longer and have less missingness than JHU and ACAPS descriptions.

- Priority 5: John Hopkins is prioritized fifth for all data except for Canadian data because compared to the ACPAS taxonomy, the John Hopkins taxonomy is relatively similar to the CoronaNet taxonomy and it is relatively rich in subnational data. It was prioritized after the other datasets in part because it has no information on end dates
- Priority 6: ACAPS data is prioritized sixth for all data except for Canadian data this is because the compared to the other datasets, its textual descriptions are of poorer quality and because it has no information on end dates.

Using the above methodology, we identified 5989 duplicate observations. The distribution of policies identified as duplicates is shown in Table 7. Here we can see that observations from OxCGRT and ACAPs were discarded most often given these criteria. We then sampled 100 groups of observations identified to be duplicates, for a total of 425 observations, using this algorithm and found that 74.5% to be true duplicates, meaning that likekly, around 1500 observations were discarded as being duplicates in this process that likely were unique observations. Given that we identified around 180k observations to harmonize to begin with and that most policies discarded were from datasets that we had previously found to have a higher likelihood of duplication (OxCGRT) or to be comparatively of lower quality (ACAPS), we made the judgement call that it was acceptable to discard this small percentage of observations without threatening the rigor of the data harmonization exercise writ large. Moreover, discarding these policies for consideration for manual harmonization at this point does not preclude doing so at a later state should resources allow for reassessing the value of harmonizing these policies.

Table 7 Assessment of duplicates across datasets

ACAPS	CDC_ITF	CIHI	COVIDAMP	EURO	JHU	OxCGRT
1909	273	22	753	519	92	2421

4.3.6 Step 3c. Deduplication between CoronaNet and External Datasets

Lastly, we also evaluated the extent to which there were duplicates between the CoronaNet dataset and the external datasets. Such duplication can occur for the same reason that there is duplication across datasets: there has not been coordination between CoronaNet and these other datasets in terms of collecting policies and as such it is quite possible that there are duplicates across these datasets.

Like our attempts to identify duplicates both within and across the external datasets, we also experimented with different sets of variables that could accurately identify true duplicates across the CoronaNet and external datasets. However, ultimately we were not able to find a combination that yielded sufficiently high accuracy. Our best attempt used the following variables to identify true duplicates:

- country
- province
- date_start
- init_country_level
- link

Based on this criteria, we sampled 100 groups of policies found to be duplicates (equivalent to 764 observations) but found that only 14 were true duplicates, for an accuracy of 14%. Subsequent efforts with other sets of variables did not improve on this percentage. As such, we were unable to automate deduplication of the external dataset across the external and CoronaNet datasets and limited our automated deduplication efforts to deduplication within and across external datasets. .

As a last step, we adjusted the dataset at this stage for the sample of policies that we manually inspected for duplication in Steps 3a and 3b. In other words, we recovered the policies that the algorithm falsely identified as being duplicates and added them back to the dataset to be evaluated for manual harmonisation. In so doing, we additionally identified observations that would not be considered policies in the CoronaNet dataset from this sample (around 50) and removed them for consideration from manual harmonisation.

4.3.7 Step 4. Piloting of Manual harmonization Efforts

Steps 1 through 3 yielded an external dataset for which automated taxonomy mappings provided a rough first translation of the external data to the CoronaNet taxonomy and automated deduplication was able to remove the most obvious instances of duplicates within the external dataset.

As the challenges of harmonizing data from different, unclean data with inconsistently preserved raw sources revealed themselves, it became clear that the bulk of the work in data harmonization would need to be manual. While automated methods were able to reduce the size of the external dataset from around 180k to around 150k records, this still represents a tremendous number of policies to harmonize. As such, the CoronaNet Research Project has recruited hundreds of volunteers from around the world to help us complete this task.

Before rolling out these efforts to the entire project however, we first piloted data harmomization for a subset of each external dataset in order to i) validate the accuracy of the automated taxonomy mappings in Step 1 and ii) learn about potential difficulties and pitfalls as well as useful strategies to data harmonization so as to provide better guidance to future volunteers.

Table 8 describes the scope of our pilot harmonization efforts. The 'assessment time frame' refers to the actual time frame spent on piloting the data harmonization efforts (as opposed to the when the policies themselves were implemented). Part of the reason for these staggered time frames is that each taxonomy map itself took around 3-4 weeks to create; once a taxonomy map was created, it was immediately piloted for a given geographical scope. The choice to pilot certain countries and regions depended both on the availability of data for a given region for a given dataset and CoronaNet's own prioritization of harmonizing European countries first given its partial funding from an EU Horizon 2020 grant. While relatively more assessments were done for taxonomies that were piloted earlier, fewer policies were assessed later on in part because i) taxonomy maps became better given the experience building the earlier ones and ii) assessment capabilities became higher given the experience of asessing earlier taxonomies. The rollout of starting with mapping taxonomies from certain datasets as opposed to others was largely a function of how much capacity for cooperation the partner dataset was able to provide in building a given taxonomy map.

As can be seen in Table 8, initially we sought to also include CCCSL in our pilot harmonization efforts. Unlike for the other taxonomy maps, the taxonomy map in this case was spearheaded by CCCSL partners. However, as part of the pilot assessment exercise, we found that both the CoronaNet and CCCSL were too complex to create high-accuracy maps. As previously discussed, given that CCCSL also had only around 11k observations, relatively few observations compared to other trackers with aspirations to track policies world-wide, inconsistently preserved sources, and unstandardized descriptions, we decided to depriortize harmonizing CCCSL data.

In piloting this data harmonization process more generally, research assistants reported that vague or incomplete descriptions and missing or dead links increased the difficulty of the work. It was not uncommon to encounter duplicate policies or external policies that needed to be broken down into smaller pieces in order to translate properly into the CoronaNet taxonomy. The pilot harmonization process also produced a pool of strategies and tips that future research assistants could draw on in their own efforts.[21] Ultimately, these experiences helped us finalize the procedure we developed to manually harmonize the data, which we describe more in the following section.

4.4 Step 5: Manual harmonization of Data

After having piloted our manual data harmonization efforts for each external dataset separately, we then finalized our plans for manual harmonization of

[21] Some strategies include (i) reading through the descriptions of all observations for a given country or region first in order to catch potential errors in the dataset (ii) using the Way Back Machine to recover dead links (iii) being aware that national level data from the OxCGRT dataset may include information about subnational policies because of the particulars of their methodology.

Table 8 Assessment Pilot harmonization Efforts

Dataset	Assessment Time Frame	Geographical Scope	No. Policies Assessed
CCCSL	March to April 2021	Liechtenstein, United Kingdom	approx 600
JHU	May to August 2021	India subnational (Andhra Pradesh, Bihar, Chandigarh, Chhattisgarh, Dadra and Nagar Haveli, Goa, Gujarat, Hayana, Jammu and Kashmir, Madha Pradesh, Puducherry, Punjab, Tamil Nadu, Tripura, Uttar Pradesh), Slovenia, Luxembourg	687
CIHI	July to September 2021	Saskatchewan, New Brunswick, Alberta, Manitoba	469
WHO CDC and WHO EURO	June to September 2021	Slovenia, North Macedonia, Estonia	330
COVIDAMP	August to September 2021	United States subnational (North Carolina, Maryland, Wyoming, Alaska, Georgia)	262
OxCGRT	September 2021	Luxembourg	68
ACAPS	September 2021	Bulgaria	40

the full combined external dataset into two main steps. First, each observation is assessed for whether it is already documented within the CoronaNet dataset or not. This information is saved internally under the column name 'overlap_assessment'. Second, observations that are currently not in CoronaNet are recoded using the CoronaNet taxonomy and harmonized into the CoronaNet dataset. This information is saved internally under the column name 'integrate_assessment'. We elaborate on each of these steps in the below.

In order to allow coders to manually assess the external data according to this criterion, we wrote the external data into Google Sheets, which we refer internally as the 'Data Integration Sheets', and grouped each sheet by country or subnational region and added conditional formatting to help facilitate their assessments. A note here on language: at the beginning of our harmonization process, we inaccurately referred to our efforts as 'data integration' instead of 'data harmonization'. For the sake of replicability, we keep this language now in our discussion that follows, with apologies to the reader.

By using Google Sheets, we were able to provide an editable, centralized place for numerous different people to assess the external data. In addition to the 'overlap_assessment' and 'integrate_assessment' columns as well as columns to record which human coder made a given assessment, these sheets also provide information about the:

- Unique identifier for a given external observation (unique_id)
- Dataset that it belongs to (dataset)
- Textual description of the observation (description)
- Timing of the policy (date_start; date_end)
- Likely policy type. The type and type_sub_cat: provides the direct mapping while type_alt and type_alt_2 provides the machine learning prediction of the policy type, where available
- Demographic targets of a policy when available (target_who_what, target_who_gen)
- Geographic information about the policy initiator (country, province, city, init_other)
- Geographic target of the policy (target_country, target_province, target_city, target_other)
- Compliance of the policy (compliance)
- Types of travel the policy affected if applicable (travel_mechanism), and
- Raw source of the policy either in terms of the original URL (link) or a PDF of the source (pdf_link).

We summarize each of the steps below before then providing by an example of how the Data Integration Sheets are used following this methodology. Though manual harmonization of the data is still ongoing, we close the section by providing an assessment of our progress to date and a discussion of tools and resources we have developed to support this process.

4.4.1 Step 5a. Manual assessment of overlap between external and CoronaNet data

For each observation in the external dataset, a human coder evaluates whether this observation has previously captured in the CoronaNet dataset or not. This evaluation is stored in the column 'overlap_assess' in the Data Integration Sheets and can take on the values of 'Yes', 'No', or 'NA'. The meaning of each of these values is as follows:

- *'Yes'*: this means that the external observation had already been independently captured in the CoronaNet dataset. In this case, the research assistant should copy and paste the corresponding CoronaNet unique identifier, which is stored in its record_id variable, into the matched_record_id column in the Data harmonization Sheet.

- *'No'*: this means that the external observation has not been previously captured in the CoronaNet dataset. In this case, the human coder should move onto the second step of manually harmonizing the data.
- *'NA'* this means that no one has yet been able to make an assessment of whether a given observation is or is not already in the CoronaNet dataset.

4.4.2 Step 5b. Manual harmonization of data

If a given observation is found to be in the external dataset but not in the CoronaNet dataset, then the human coder should move onto to second step of harmonizing this external data into the CoronaNet taxonomy. To do so, they are instructed to treat the external observation just as they would any other potential source of information about a COVID-19 policy. In particular, they are asked to first go to the raw source of information using either the URL or PDF links (if available) provided for a given policy. That is, they are asked to recode the data based on the raw source of information provided in the Data harmonization Links, rather than from the textual description of the observation provided by the external data.

Once they have read through the raw information source, they can then either recode the information into the CoronaNet taxonomy using the normal procedure for documenting policies at CoronaNet (that is, they can document this information into a Qualtrics survey customized for this purpose. See the Methodology section in [23] for more information) or they can provide another assessment of the external data. In the 'integrate_assess' column, they can make one of the following 6 assessments:

- 'Integrated'; this means that the coder has recoded it into the CoronaNet taxonomy.
- 'Integrated with additional original research': this means that the coder had to do some additional research before coding the observation into the CoronaNet taxonomy. This could be for any number of reasons. E.g. the information that from the URL or PDF links in the external dataset may be unclear or require additional context/knowledge to code well.
- 'Integrated with additional work to find a new link' means that the original link for the policy is dead but that the RA was able to find a new link that corroborates the information described in the 'description' column.
- 'Integrated with additional original research AND with additional work to find a new link': means the RA fulfilled both the criterion under: 'Integrated with additional original research' and 'Integrated with additional work to find a new link'. See above for more information.
- 'Duplicated policy': this means that there were multiple external policies that were duplicates of each other. In this case, the coder is asked to only harmonize one of them and to mark the other ones as being duplicates.

overlap_assessment	matched_record_id	Notes	integrate_assessment	unique_id	dataset	description
No	NA	NA	Not a relevant policy	OXCGRT_Hungary_20210728_mask	OXCGRT	No policy changes. "It is no longer mandatory to wear a face mask, except in hospitals and social institutions."
Yes	R_3NXmQbf9TrzN3XU	NA	NA	EURO_730824_1	EURO	School trips abroad are forbidden. Already booked school trips abroad must be cancelled. The foreign language study program is also suspended
NA		NA	NA	OXCGRT_Hungary_20200311_school	OXCGRT	On 11 March 2020, Hungary's government declared a state of emergency and closed university campuses. Archive link doesn't work consistently, so including original for reference too: UNESCO data confirms a partial closure between March 11 and March 15, before shifting to a complete closure:

Fig. 9 Example of Data harmonization Sheets for France

- 'Not a relevant policy': this means that after having taken a closer look at the link for the observation is not one that would be coded in the CoronaNet taxonomy.
- 'Link dead, no other link found' means that the original link for the policy as noted in the CoronaNet Data harmonization sheet is dead and the coder was unable to i) use the Way Back Machine to find the original data ii) find another link to corroborate this information. In this case, the coder is instructed to not recode this policy

Figure 9 provides a visual example of this data harmonization exercise for three policies in Hungary. The first policy was found to not have been in the CoronaNet dataset. As such, the coder marked the overlap_assessment as 'No'. After looking through the URL or PDF link, the coder then subsequently assessed the policy as being an irrelevant policy to the CoronaNet dataset and thus 'Not a relevant policy' was chosen in the integrate_assessment column.

Meanwhile, the second observation was found to have already been coded in the CoronaNet dataset; as such the coder marked the overlap_assessment as being 'Yes' and copied and pasted the corresponding record in the CoronaNet dataset, R_3NXmQbf9TrzN3XU into the matched_record_id column.

Finally, at the time of writing, the third policy has not been assessed for harmonization yet. As such, both the overlap_assessment and integrate_assessment columns take the value of NA.

Step 5 of manually harmonizing the data is still ongoing. However, based on the X observations that we have assessed so far, we have found that on average 83% of policies in the external dataset were not previously in the CoronaNet dataset. Table 9 provides a breakdown of the overlap assessment by dataset. Overall, the JHU and CoronaNet datasets have the most amount of overlap at 36% while the OxCGRT and CoronaNet datasets have the least amount of overlap at 12%.

Meanwhile Table 10 shows the breakdown to date of the harmonization assessment. Recall, that these assessments are only done for policies that are found to not currently be in the CoronaNet dataset, or in other words, for 83% of the external data assessed to date. The integration assessments show that around 44% of the data found to not currently be in the CoronaNet dataset is subsequently recoded into the CoronaNet taxonomy and dataset, with around 11% of requiring either additional research or work to find a new link before this is possible. Meanwhile, 24% of the observations are assessed to be duplicates,

Table 9 Overlap Assessment by Dataset

overlap assessment	Total	ACAPS	CDC_ITF	CIHI	COVID-AMP	EURO	JHU	OxCGRT
No	0.83	0.79	0.75	0.84	0.82	0.79	0.64	0.88
Yes	0.17	0.21	0.25	0.16	0.18	0.21	0.36	0.12

21% are found to be not relevant policies and 10% do not have a recoverable link and thus cannot be substantiated and subsequently recoded.

There is however substantial variation for each assessment across the different datasets. Overall, it appears that data from CIHI is often harmonized without the need for substantial extra work and without problems with relatively low issues with duplicated policies, dead links or irrelevant policies. This may in part be due to the fact that CIHI data focuses on subnational Canadian data and is relatively high quality given that it is collected not by volunteers, but by paid contractors. Meanwhile it appears possible to recover and recode signficantly more information from the COVIDAMP datasets compared to other datasets, likely because they also provide PDF links to their original sources. With regards to duplication, ACAPS, the CDC_ITF, COVIDAMP and OxCGRT appear to have about the similar amount of duplication while there is comparatively little duplication for WHO EURO and as previously mentioned CIHI. The fact that the rate of duplicate data for OxCGRT data is relatively in line with those found for other datasets also suggests that we did not go overborad with our custom OxCGRT deduplication efforts in Step 3a. Dead links appear to be a particular problem for WHO EURO sources while irrelevant policies appear to be particularly high with regards to OxCGRT data. This is largely due to previously mentioned differences in OxCGRT and CoronaNet methodology; while OxCGRT documents policies that have 'No change', CoronaNet does not (see section 4.3.4 for more details).

We conclude by noting that since the last step in the harmonization of the different taxonomies into CoronaNet taxonomy is manual and requires the enlistment of a substantial labor force, we have made signficant investments in training research assistants and providing supportive resources for them to minimize the possibility of systematic coding errors. These include:

- Regular workshops for managers and research assistants about data harmonization. These are mandatory for new research assistants and they receive this training along with the original training that we developed to onbaord them into the project [23].
- The design and diffusion of reference material to the research assistants, such as: manuals, spreadsheets, presentations, info-graphics and videos.
- Monitoring and rectification of inconsistencies identified in both the overlap assessment and data harmonization stages of the harmonization process by

Table 10 Harmonization Assessment by Dataset

integrate assessment	Total	ACAPS	CDC ITF	CIHI	COVID-AMP	EURO	JHU	OxCGRT
Integrated	0.36	0.42	0.46	0.91	0.47	0.36	0.41	0.28
Integrated with additional work to find a new link	0.03	0.03	0.03	0.01	0.03	0.04	0.04	0.02
Integrated with additional original research	0.05	0.04	0.09	0.03	0.06	0.05	0.07	0.04
Integrated with additional original research AND additional work to find a new link	0.02	0.01	0.03	0.00	0.01	0.03	0.01	0.02
Duplicated policy	0.24	0.21	0.20	0.05	0.23	0.18	0.32	0.29
Not a relevant policy	0.21	0.16	0.11	0.00	0.17	0.14	0.10	0.30
Link dead, no other link found	0.10	0.12	0.08	0.00	0.02	0.20	0.05	0.07

both managers and automated code. If there is an error in the data harmonization process, it is noted and communicated as feedback to research assistants to rectify before it is accepted as a valid harmonized entry.
- Open communication channels for research assistants to recieve asynchronous feedback on questions they may have on the data harmonization process through Slack.

While our harmonization efforts are still ongoing, we hope that the methodology we have outlined here can prove useful to others seeking to harmonize similar data or to evaluate the work of others.

Supplementary information. Interested readers are encouraged to see the supplementary information file which contains further information and links to the taxonomy maps we created to map each external dataset to the CoronaNet taxonomy.

Acknowledgments. We deeply thank the very large number of RAs who have been working on harmonizing PHSM data. Of the more than 350 individuals who have helped harmonize PHSM data to date, the following deserve special acknowledgement for their work in onboarding and training research assistants in the process of data harmonization: Vanessa Zwisele, Augusto Teixeira, Constanza Schönfeld, Muneeba Rizvi, and Clara Fochler. Augusto Teixeira also deserves special recognition for his role in monitoring and providing feedback on data harmonization errors to research assistants and Muneeba Rizvi for tirelessly providing weekly progress updates on our manual harmonization efforts.

We also recognize Audrey Firrone and Natalia Filkina Spreizer for their help in creating the taxonomy map between CIHI and CoronaNet. We are moreover grateful to the following individuals for their work in helping us pilot our data harmonization efforts in the summer of 2021:

Joseph Shim, Natalia Filkina Spreizer, Audrey Firrone, Silvia Biagioli, Mayuiri A.,Maanya Cheekati, Laura Eckoff, Fiona Valad, Katelyn Thomas, Humza Q, Amy Nguyen, Rawaf al Rawaf, Sella Devita, Paula Ganga, Tim Bishop, Jaimi Plater, Rose Pasty, Natalie Ellis, Maryam Al Hammadi, Shreeya Mhade, Shrajit Jain, Kyle Oliver, Shaila Sarathy, Alisher Shariyazdanov, Emma Baker, Jurgen Kadriaj, Celine Heng and Augusto Teixeira.

Finally, the following research assistants have done at least 100 overlap assessments or 50 integration assessments on the dataset and we are deeply appreciative for their contribution to this work:

Abylaikhan Mamyrov, Abylay Abdrakhmanov, Aden Littlewood, Aidana Askar, Aigerim Kenes, Aina Kabuldinova, Aisulu Kossanova, Aiym Smadil, Aizada Omen, Aizhan Token, Aizhan Zhaksygaliyeva, Aleksiina Kallunki, Alexandr Lopukhov, Ali Kahraman, Aliya Ibrayeva, Altynay Askerova, Amina Abylkassymova, Amina Yensegenova, Aneliya Kassymova, Anke Horn, Arailym Kalkeshova, Arailym Kanatkyzy, Arailym Myrzagaliyeva, Aruzhan Ilanbay, Aruzhan Mardissadyk, Assel Abdukadyrova, Assel Kambar, Assemay Toganbay, Audrey Firrone, Audrey Tey, Augusto Teixeira, Avirat Desai, Ayan Tlepova, Ayaulym Saduakas, Ayaulym Tokenova, Bayan Tleumagambet, Bekzat Kabenov, Bekzat Sabyrov, Bibinur Salykova, Botagoz Amangeldiyeva, Celine Heng, Clara Fochler, Damina Askarova, Damira Nygmetova, Damira Shertay, Dana Sultanbekova, Danagul Bakhytova, Daniyarbek Bakhytzhan, Darya Gorbacheva, Dilnaz Medenova, Dilnaz Zhangazinova, Dilyara Begalykyzy, Dulat Minas, Ella Goeckner-Wald, Ethan

Fulton, Fariza Shermakhanova, Fiona Valade, Fiona Zhong, Giovanna Monteiro de Sá, Guang Shi, Gulzhan Tulbayeva, Humaira Hossain, Ida Steineck Nilsson, Jellen Olivares-Blanco, Jia Wei Liu, Jianing Li, Jose Muriel, Jurgen Kadriaj, Kader Saygili, Kamila Bissenkulova, Kamila Isataeva, Kanat Kenzhetayev, Kelly Dang, Komila Nassyrova, Constanza Schönfeld, Kudaibergen Aldakayev, Lailim Azetova, Larissa Yugay, Laula Bibolat, Laura Duisebayeva, Liliya Mukhamejanova, Madina Altenova, Madina Kupzhassarova, Madina Mamybekova, Madina Tanbayeva, Maksat Akimzhanov, Mari Pulido, Mario Hernandez, Matthew Cottrell, Melody Bechler, Meruyert Ospan, Milana Kapezova, Mimi Suhyoon Lee, Minar Mekekyzy, Moza Alhuraiz, Mukhtabar Sabyrova, Muneeba Rizvi, Nargiz Turtayeva, Nargiza Kozhanova, Natalia Filkina-Spreizer, Nazerke Baimukan, Nazerke Kanatbekova, Nazym Malikova, Nicole Lawrence, Nicole Mattson, Nidia Michinge, Nur Hazwani Binti Shariff, Nurbibi Altayeva, Nursaya Alpysbes, Paloma Laye, Rauan Alguatov, Rawaf al Rawaf, Sabit Kyrykbay, Sai Teha Muramalla, Samat Sayakov, Samet Berk Oksuz, Shynggyskhan Yeshev, Silvia Biagioli, Silvia Paolucci, Simon Jenke, Sonia Mago, Steve Giang, Stian Jenssen, Sukhrab Turdiyev, Suyuan Wang, Symbat Mombayeva, Tafadzwa Wandayi, Tasia Wagner, Tom Wiederkehr, Vicky Liu, Waldemar Hartmann, Wing Shea, Wissam Gaith, Xiatian Ye, Yaqi Ren, Yeligay Segizbay, Yelnaz Ramazanova, Zarina Irgaliyeva, Zhadyra Yeshentay, Zhanel Jamenkova, Zhaniya Abdrakhman, Zhanna Bakytbek, Zhansaya Nurzhaubayeva, Zhuldyz Ramazanova, Zi Yu Zhang, Akbota Sarzhanova, Altynay Bakbayeva, Amina Darayeva, Aruzhan Balgynbayeva, Assem Bazarbek, Axcel Jasso, Ayaulym Matisheva, Azhar Alisheva, Dana Aitpenbet, Dilnaz Tussupayeva, Jeff Huang, Muslimbek Temirlan, Naela Elmore, Nurdana Kalaganova, Symbat Kusherova, Talitha Gower, Yelnur Nurlan, Yulduzkhan Turgunova, Zhandossay Bakyt, Zilin Tao.

We also thank Svanhildur Thorvaldsdottir for her helpful guidance and comments during the initial development of this paper and Hannah Löffler for her comments and edits of the paper. We are also grateful for Hans-Peter Nowak's help in making our harmonization efforts accessible on our website.

Finally we are also very grateful for the support that we have received from the Chair of International Relations (Tim Büthe), Hochschule für Politik at the Technical University of Munich (TUM) for this work. Moreover, our work would not be possible without the support and generosity that other COVID-19 tracking projects have provided. We especially would like to recognize Amélie Desvars-Larrive and Michael Gruber (CCCSL), Sophia Zweig, Alex John Zapf, Kyle Oliver, and Hanmeng Xu (JHU HIT-COVID), Erin Pichora and Christina Cately (CIHI), Alex Howe, Steve Penson, Angeliki Niki and Claudia Manili (ACAPS), Alaina Case and Ellie Graeden (COVIDAMP) and Yuxi Zhang, Anna Petherick, Thomas Hale and Toby Phillips (OxCGRT).

Declarations

Funding. We are grateful for the generous funding support that we have received from the Chair of International Relations (Tim Büthe), Hochschule für Politik at the Technical University of Munich (TUM), New York University Abu Dhabi (Covid-19 Facilitator Research Grant), Leibniz Research Alliance Group "Crises in a Globalised World", "CoronaNet in Eurasia: Leveraging the Comparative Moment in COVID-19 Research". Grant number 832-06g. This project has also received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016233 for EU data collection.

Conflict of interest/Competing interests. We declare no competing interests.

Ethics approval . Not applicable.

Consent to participate. Not applicable.

Consent for publication. Not applicable.

Availability of data and materials. Users interested in which observations in the CoronaNet dataset were harmonized from external datasets can reference either the 'collab' or 'collab_id' columns in the raw event dataset made publicly available on the CoronaNet github repository here: https://github.com/CoronaNetDataScience/corona$_t$scs/tree/master/data/CoronaNet. The 'collab' variable notes which external dataset, if any, an observations was harmonized from and the 'collab_id' variable documents the original unique ID which matches the corresponding observation in the original data.

Code availability. Please refer to the Supplementary Materials in order to access the taxonomy maps made for Step 1 of the data harmonization process. Access to these taxonomy maps is also provided through our website here: https://www.coronanet-project.org/external$_d$ata$_h$armonization.html. Meanwhile, we make the code and data for replicating Steps 2 to 3 available in the following folders in the CoronaNet public git repository: https://github.com/CoronaNetDataScience/corona$_t$scs/tree/master/RCode/collaboration https://github.com/CoronaNetDataScience/corona$_t$scs/tree/master/data/collaboration.

Author's contributions. Author's contributions to the paper are as follows: Conceptualization (C.C., L.M.), Methodology (C.C., I.B., M.W., R.B.), Software (C.C.), Validation (C.C.), Investigation (I.B., M.W., R.B.), Resource (C.C., L.M. C.S., V.G, R.K, J.B.), Data Curation (C.C.), Writing - Original Draft (C.C.), Writing - Review & Ediging (C.C., L.M., I.B., M.W., R.B., T.M.), Visualisation (C.C., L.M., I.B., R.B.), Supervision (C.C., L.M., C.S., V.G., T.M., R.K., J.B.), Project adminstration (C.C., L.M., I.B., M.W., R.B. V.G., R.K., J.B.), Funding (C.C., L.M., C.S., R.K., J.B.).

Appendix A COVID-19 Trackers

The following table provides an overview of the 24 largest COVID-19 PHSM data collection efforts. For each, we provide a short description [Description], an estimation of the number of policies it has documented at the time of writing [Records], its geographic scope [Geographic scope], whether it is actively collecting data [Still collecting data?], the last date the dataset was retrieved or updated [Last retrieved; updated], the sources that the dataset relies on [Sources], a link to its URL [Website], and where the data tracking effort is locating geographically [Based in].

We hope that the inclusion of this table will help readers better contextualize the decision that we made to harmonize certain datasets as opposed to others. We further believe this table can in general provide readers with a comprehensive overview of available PHSM data.

Table A1: Summary of existing COVID-19 PHSM Trackers

Dataset	Description	Records	Geographic scope	Still collecting data?	Last retrieved; updated	Sources	Website	Based in
ACAPS	The COVID-19 Government Measures Dataset puts together all the measures implemented by governments worldwide in response to the Coronavirus pandemic. Data collection includes secondary data review. The researched information available falls into five categories. Each category is broken down into several types of measures.	23,923	Worldwide	No	10-01-2020	Governments, media, United Nations, and other organisations	https://www.acaps.org/covid-19-government-measures-dataset	Switzerland
COVID-AMP	The dataset documents policies and plans to address the COVID-19 pandemic.	40,429	Worldwide	Yes	03-16-2022	Governments and media.	https://www.covidamp.org/about/amp	United States

Table A1: Summary of existing COVID-19 PHSM Trackers

Dataset	**Description**	**Records**	**Geographic scope**	**Still collecting data?**	**Last retrieved; updated**	**Sources**	**Website**	**Based in**
CIHI	CIHI is maintaining a comprehensive scan of federal, provincial and territorial government interventions, announcements and other measures to reduce the spread of and improve the health outcomes related to COVID-19.	5,413	Canada	Yes	09-30-2021	Governments and media.	https://www.cihi.ca/en/covid-19-intervention-scan	Canada
HIT-COVID	The dataset documents tracks the implementation and relaxation of public health and social measures (PHSMs) taken by governments to slow transmission of SARS-COV-2 globally.	13,658	Worldwide	No	06-01-2021	Governments, and other organisations.	https://github.com/HopkinsIDD/hit-covid	United States
OxCGRT	OxCOVID19 Database is a large, single-centre, multimodal relational database consisting of information related to COVID-19 pandemic.	128,891 / 85,295	Worldwide	Yes	11-14-2021	Governments, and other organisations.	https://covid19.oii.ox.ac.uk ; https://github.com/covid19db/data/blob/master/data-government-response/covid19db-government-response-GOVTRACK.csv.bz2	United Kingdom

Table A1: Summary of existing COVID-19 PHSM Trackers

Dataset	Description	Records	Geographic scope	Still collecting data?	Last retrieved; updated	Sources	Website	Based in
WHO Euro	The dataset document by european countries, territories and areas that enforce rules or guidelines to limit the spread of COVID-19	27,100	Worldwide	Yes	08-16-2022	Governments, media and other organisations.	https://who.maps.arcgis.com/apps/dashboards/ead3c6475654481ca51c248d52ab9c61	Switzerland
WHO CDC	The dataset document by countries, territories and areas that enforce rules or guidelines to limit the spread of COVID-19	7,940	Worldwide	Yes	06-28-2021	Governments, and other organisations.	https://covid19.who.int/info	Switzerland
Bogazici University	Dataset that track the number of economic policy responses taken by governments worldwide to a face the effects of the pandemic.	169	Worldwide	No	05-07-2021	Media	http://web.boun.edu.tr/elgin/COVID.htm	Turkey

Table A1: Summary of existing COVID-19 PHSM Trackers

Dataset	Description	Records	Geographic scope	Still collecting data?	Last retrieved; updated	Sources	Website	Based in
Complexity Science Hub COVID19 Control Strategies List (CCCSL)	The dataset describes the implemented NPIs for 57 countries, including the Diamond Princess cruise ship. Measures implemented at the subnational level (state, region, city) are also included.	11,512	57 countries	No	n.a	Public sources	https://covid19-interventions.com	Austria
COVID Border Accountability Project	The dataset documents country-level travel and immigration bans introduced in response to COVID-19.	2,928	Worldwide	No	12-21-21	ACAPS CoronaNet Project	https://covidborderaccountability.org/	France
International Institute for Applied Systems Analysis (IIASA)	The dataset stock of COVID-19 datasets for 26 European countries at the regional NUTS3 or NUTS2 level.	1,210	26 European countries	Yes	March 2022	Governments, and other organisations.	https://asjadnaqvi.github.io/COVID19-European-Regional-Tracker/	Austria

Table A1: Summary of existing COVID-19 PHSM Trackers

Dataset	**Description**	**Records**	**Geographic scope**	**Still collecting data?**	**Last retrieved; updated**	**Sources**	**Website**	**Based in**
Covid-19 Policy Tracker	This project aims to fill in this gap by providing a full record of state/provincial-level (or equivalent first-level subdivisions) PHSMs in six countries: Japan, Korea, UK, Germany, Brazil and China (to be added), which cover different continents and approaches towards Covid-19.	870	6 countries	No	08-01-2020	Governments, and other organisations.	https://citiesandregions.cn/research/Covid-19-Policy-Tracker	China
COVID-19 State Policy Project	The dataset documents US state-level distancing policies to the 2019 novel coronavirus (SARS-CoV-2), the cause of COVID-19.	16,512	1 country	No	09-08-2021	Governments, and other organisations.	https://github.com/COVID19StatePolicy/SocialDistancing	United States

Table A1: Summary of existing COVID-19 PHSM Trackers

Dataset	Description	Records	Geographic scope	Still collecting data?	Last retrieved; updated	Sources	Website	Based in
Eurofound	PolicyWatch aims to map measures introduced to cushion the social and economic effects on businesses, workers and citizens. It also includes information on the role played by social partners in the design and implementation of the measures.	1,465	EU Member States + Norway	Yes	March 2022	Governments, and other organisations.	https://www.eurofound.europa.eu/data/covid-19-eu-policywatch	European Union
Grattan Insitute: coronavirus announcements tracker	The dataset documents Australian Government policy announcements in response to the COVID-19 crisis	326	1 country	No	12-04-2021	Governments, and other organisations.	https://docs.google.com/spreadsheets/d/1ZqnCmSueVD26Xrw1hMszi5Q_aQwsZJe78Y5HCLwXQ64/edit#gid=447719678	Australia

Table A1: Summary of existing COVID-19 PHSM Trackers

Dataset	**Description**	**Records**	**Geographic scope**	**Still collecting data?**	**Last retrieved; updated**	**Sources**	**Website**	**Based in**
IAE Paris - Université Paris I Panthéon-Sorbonne	The Response2covid19 dataset aims at rigorously tracking and comparing governments' responses to face the COVID-19 pandemic. This dataset includes economic measures taken by governments	4,985	Worldwide	Yes	05-10-2020	ACAPS, IMF and UNESCO	https://response2covid19.org/	France
International Network for Government Science Advice	The dataset keeps track of how policy interventions are being made by various national and sub-national (state, province, etc.) governments across the world.	5,848	Worldwide	No	11-01-2021	Governments, and other organisations.	https://ingsa.org/covid/policy making-tracker/	New Zealand

Table A1: Summary of existing COVID-19 PHSM Trackers

Dataset	Description	Records	Geographic scope	Still collecting data?	Last retrieved; updated	Sources	Website	Based in
European Centre for Disease Prevention and Control	The dataset documents non-pharmaceutical interventions (or response measures) based on information available from official public sources, and may not capture measures being taken by countries that are not reported on publicly available websites.	2,036	30 European countries	Yes	05-03-2022	Governments, and other organisations.	https://www.ecdc.europa.eu/en/publications-data/download-data-response-measures-covid-19	Sweden
European Centre for Disease Prevention and Control and Joint Research Centre (JRC) of the European Commission	The dataset documents non-pharmaceutical interventions (or response measures) based on information available from official public sources, and may not capture measures being taken by countries that are not reported on publicly available websites.	5,254	30 European countries	Yes	03-28-2022	Governments, and other organisations.	https://covid-statistics.jrc.ec.europa.eu/RMeasures	Sweden

Table A1: Summary of existing COVID-19 PHSM Trackers

Dataset	Description	Records	Geographic scope	Still collecting data?	Last retrieved; updated	Sources	Website	Based in
COVID-19 Global Gender Response Tracker - UNDP	The dataset monitors responses taken by governments worldwide to tackle the pandemic, and highlights those that have integrated a gender lens. It captures two types of government responses: women's participation in COVID-19 task forces and national policy measures taken by governments. It analyzes which of the policy measures address women's economic and social security, including unpaid care work, the labour market and violence against women.	4,968	Worldwide	No	12-31-2022	Governments, and other organisations.	https://data.undp.org/gendertracker/	United States
The Health Foundation - Covid-19 policy tracker	The dataset documents national government and health and social care system responses to COVID-19 in England in 2020.	1,149	England	No	12-31-2021	Government institutions	https://www.health.org.uk/news-and-comment/charts-and-infographics/covid-19-policy-tracker	England

Table A1: Summary of existing COVID-19 PHSM Trackers

Dataset	Description	Records	Geographic scope	Still collecting data?	Last retrieved; updated	Sources	Website	Based in
Yale SOM-Tobin Center State and Local COVID Restriction Database	The dataset documents business and related restrictions issued by all U.S. states and counties in response to the COVID-19 pandemic.	n.a.	United States	n.a.	n.a.	Government institutions	https://som.yale.edu/covid-restrictions	United States
COVID-19 U.S. State Policy (CUSP)	The COVID-19 U.S. State Policy (CUSP) database documents the dates all 50 states and the District of Columbia implemented health and social policies to respond to the COVID-19 pandemic and its economic ramifications.	n.a.	United States	n.a.	n.a.	Government institutions	https://statepolicies.com/data/graphs/line/	United States

Table A1: Summary of existing COVID-19 PHSM Trackers

Dataset	**Description**	**Records**	**Geographic scope**	**Still collecting data?**	**Last retrieved; updated**	**Sources**	**Website**	**Based in**
Project Lockdown	The dataset documents the different policies that governments are undertaking during the COVID-19 crisis by mapping them and measuring a number of relevant metrics.	n.a.	Worldwide	n.a.	n.a.	Government institutions	https://github.com/TheIOFoundation/ProjectLockdown/wiki	Malaysia

Appendix B Coverage of subnational policy-making by country and time coverage

The following table provides an overview of subnational coverage of COVID-19 policies based on a review of the datasets covered in A1. Note the time coverage within a given dataset provides an average date across different subnational regions. For example, while the table notes that CoronaNet provides subnational data for Australia until December 2020, in effect this means that for some subnational regions the time coverage goes beyond December 2020 and for other subnational regions it stops before December 2020, with December 2020 being an approximate average date across Australia.

country	dataset (time coverage within the dataset)
Australia	CoronaNet (December 2020), OxCGRT (December 2022)
Brazil	CoronaNet (December 2020), OxCGRT (December 2022)
Canada	CIHI, CoronaNet (December 2020), OxCGRT (December 2022)
China	CoronaNet (January 2021), OxCGRT (January 2023)
France	CoronaNet (May 2021)
Germany	CoronaNet (April 2021)
India	CoronaNet (January 2021), HIT-COVID (December 2020), OxCGRT (December 2022)
Italy	CoronaNet (March 2021)
Japan	CoronaNet (January 2021)
Kazakhstan	CoronaNet (February 2021)
Nigeria	CoronaNet (January 2021)
Switzerland	CoronaNet (January 2021)
Spain	CoronaNet (May 2021)
Russia	CoronaNet (April 2021)
United Kingdom	COVIDAMP (December 2020), OxCGRT (December 2022), HealthUK[109] (December 2020)
United States	CoronaNet (December 2020), COVID-19 US State Policies (CUSP) (mid 2021) [110], COVID-19 State Policy Tracker[111] (August 2021), COVIDAMP (July 2022), HIT-COVID (November 2020), OxCGRT (December 2022), State Policy Responses to COVID-19 (SPRC19)[112] (April 2020), Yale SOM-Tobin Center State and Local COVID Restriction Database[113] (2021; data not publicly available so this is an approximation)

Table B2 Subnational data coverage by dataset and time.

Appendix C Comparison between CoronaNet and WHO PHSM data harmonization efforts

We are aware of at least one other effort to harmonize PHSM data from different datasets: the World Health Organization's (WHO) PHSM dataset. The World Health Organization's (WHO) PHSM dataset was first published in the summer of 2020 and harmonizes data from five projects which we also include in our data harmonization efforts: OxCGRT, ACAPS, HIT-COVID, WHO EURO and CDC). Aside from the fact the WHO does not include data from CoronaNet, COVIDAMP or CIHI, a crucial difference between our data harmonization efforts and the WHO effort is that the WHO PHSM dataset does not collect original data policies but rather focuses on merging different data sources. Having mapped and evaluated the quality of the WHO PHSM dataset as part of our own data harmonization exercise, we argue that our data harmonization effort improves on their efforts in several respects with regards to the scale and quality of the resulting harmonized data.

The obvious benefit of the WHO PHSM harmonization effort over ours is (i) that they have harmonized data past September 2021 and (ii) that they had been making release weekly updates which harmonizes the latest observations from each underlying dataset. Since August 2022 these weekly updates have stopped however and on their website they report that they have concluded their harmonization exercise. Despite these advantages in time coverage, we argue their approach had come at a substantial cost to data quality. We contend that combining CoronaNet's general methodology of (i) concentrating on a more limited time period and smaller set of countries through to September 2021 (ii) recruiting volunteers all around the world dedicated towards documenting policies for a given country and (iii) using a survey instrument to collect policies [23] with (iv) following a manual data harmonization effort has allowed us to create a more standardized, coherent and valid, dataset compared to the WHO effort. We elaborate on both how our data harmonization efforts compare in terms of scale and quality in the following sections. We note, that in contrast to our analysis of the subset of the WHO data that we harmonized and discuss in the Methodology section of the paper which was limited to data harmonized before September 10, 2020, in our comparison below we assess the differences between our harmonization efforts and their latest harmonized data, which contains data until August 2022.

C.1 Comparing the scale of harmonization efforts

Overall, we argue that the CoronaNet ongoing harmonization efforts have lead to a dataset that is more compact, insofar as it limits itself to policies made before September 2021, but as such more complete and high quality, than the WHO PHSM harmonization effort, which has harmonized data until August 2022.

We start with a broad comparison of the two harmonization efforts by volume of policies documented. We note that the latest, and final version of the PHSM dataset (dating to August 2022) contains around 121,000 policies, which is close to 30k policies less than the size of the existing CoronaNet dataset, which at the time of writing documents more than 150,000 policies. On the basis of the number of policies alone, our ongoing harmonization efforts, almost certainly yields a dataset that is more complete than the WHO effort for the time period up until September 2021. By comparison, for this same time period, the WHO PHSM dataset documents close to 96k policies.

Indeed, when we breakdown our harmonization efforts by dataset, we can infer that the WHO PHSM data has less complete data coverage than CoronaNet in part because it does not harmonize data from CoronaNet, COVIDAMP or CIHI. Meanwhile, figure C1 further allows us to break down the amount of data in the harmonized WHO PHSM data by dataset and finer slices of time. As it shows, over time, it has come to increasingly rely on data from OXCGT and WHO EURO datasets, as ACAPS, HIT-COVID and CDC stopped data collection.

To take a closer look at how the two data harmonization efforts compare with regards to coverage over time, although the WHO PHSM dataset has indeed been able to harmonize data past September 2021, we believe that this has come at the cost of overall data completeness and quality. That is, given that the pandemic was very much still in full swing from September 2021 to August 2022, with most countries focusing on COVID-19 vaccination in particular, we believe that the 22k+ observations that the WHO PHSM dataset has been able to harmonize from September 2021 to August 2022 can present only a very incomplete picture of the pandemic. Indeed, on further observation, we find that these 22k+ documents policies for 186 countries, with a mean of 120 policies per country. By comparison, the WHO PHSM dataset documented around 44k policies for the same time period one year before, that is, from September 2020 to August 2021 for 228 countries, with a mean of 186 policies per country. For further comparison, we can look at numbers from the CoronaNet dataset from September 2021 to August 2021. Here we find that CoronaNet documented data for 182 countries, with a mean of 228 policies per country. These numbers suggest that by focusing on a more limited period of time, the CoronaNet data harmonization effort is arguably able to build a more coherent dataset for a given time period.

With regards to geographical coverage, though the PHSM dataset provides coverage of 233 regions while our data harmonization efforts only cover 201, these additional covered regions exclusively consist of small island nations or overseas territories which are on average, undercoded within the WHO PHSM dataset[22]. Meanwhile, the WHO PHSM data harmonization effort puts

[22] In the WHO PHSM dataset, there are on average 59 policies which on average covers policies made until mid August 2020 for the following 38 islands and overseas territories and which are not covered in our data harmonization efforts: American Samoa, Anguilla, Aruba, Bermuda, Bonaire, British Virgin Islands, Cayman Islands, Cook Islands, Curacao, Falkland Islands (Malvinas), Faroe Islands, French Guiana, French Polynesia, Gibraltar, Greenland, Guadeloupe, Guam, Guernsey, Isle Of Man, Jersey, Martinique, Mayotte, Montserrat, New Caledonia, Niue, Northern Mariana

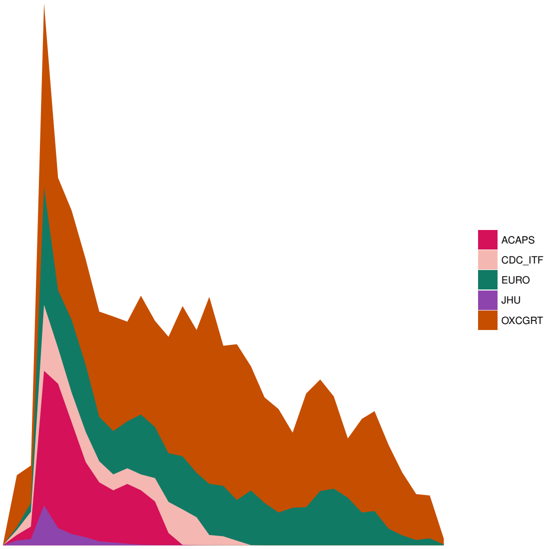

Fig. C1 Number of policies by tracker overtime in the WHO PHSM dataset]

relatively little emphasis on harmonizing subnational data; around 34% of the data it harmonizes is at the sub-national level, compared to 51% for our data harmonization efforts. This is all the more important given that there is substantial subnational variation in the policy making process for many countries, which we discuss in greater detail in the next section.

C.2 Comparing the quality of harmonization efforts

Overall, we have found that the WHO PHSM harmonization efforts suffers from significant problems with regards to data standardization, data coherence as well as source data compared to the Coronanet efforts.

With regards to data standardization, we have identified a number of inconsistencies in the WHO PHSM dataset which makes it difficult to use their data without additional processing. For example, while the WHO dataset captures rich information on the targets of its compiled data in its 'targeted' variable, the usefulness of this variable for analysis is diminished by the fact that it contains more than 13,390 unique entries.[23] While the CoronaNet dataset captures much of this same information, it organizes the information into different

Islands, Commonwealth Of The, Pitcairn Islands, Puerto Rico, Reunion, Saba, Saint Barthelemy, Saint Helena, Saint Martin, Saint Pierre and Miquelon, Sint Eustatius, Sint Maarten, Turks And Caicos Islands, United States Virgin Islands, Wallis And Futuna

[23] Upon closer examination, by simply performing some simple automated cleaning procedures on these categories like removing special characters and making all characters lower case, the actual number of unique entries is closer to 5,900. However even after performing this procedure, the point about lack of standardization still stands. E.g. various observations read: 'al schools', 'all school'; 'all schools' when it would be more useful to use one standard phrasing to refer to all schools.

fields in more manageable numbers of categories within each, which facilitates a researcher's ability to quantitatively or qualitatively compare different observations. For instance, while the WHO's 'targeted' variable includes entries as varied as 'secondary schools', 'citizens', and 'All flights', CoronaNet documents this information in separate fields ('secondary schools' can be found in the 'type_sub_cat' variable, which generally captures information on policy sub types. In this case secondary schools is as it is a sub type of the broader 'Closure and Regulation of Schools' [type] variable. Meanwhile 'citizens' can be found in the [target_who_what] variable which captures information on demographic targets, and 'All flights' can be found in the [travel_mechanism] variable which generally captures information as to the mode of travel that is restricted). Moreover, though the WHO notes that they standardise names for 'country, territory or area' in their dataset downloaded, we have consistently found that data on subnational geographic areas are inconsistently documented (documented in their 'area_covered' variable). For instance, the province of Jammu and Kashmir of India is alternatively coded as 'jammu and kashmir', 'Jammu and kashmir' or 'Jammu and Kahsmir' in the WHO dataset. Because CoronaNet uses a survey instrument to document this data, problems with typos which can make standardization difficult to achieve are avoided.

Additionally, we have found a substantial degree of policy incoherence in the WHO PHSM dataset, both in terms of the quality of the observations harmonized in the dataset as well as in terms of observations not included in the data. With regards to the former, as of August 2022, the WHO PHSM dataset lacks a textual description of a given policy for more than 890 measures and reports 2,911 policies without a start date. These issues are not present in the CoronaNet data collection methodology because of these dimensions are collected as forced responses in the survey. Meanwhile, with regards to the latter, we have found that there is still a great deal of incoherence in the external data when one simply compiles data from different datasets without doing additional research to fill in the blanks. For instance, while our data harmonization efforts of 7 different datasets have identified 844 external policies for Romania, we found that there were were a substantial number of policies that were not captured by any external data. For instance, even though we identified more 40 policies in the external dataset which could be considered as having the policy type 'Lockdown' in the CoronaNet taxonomy, further investigation revealed more than 400 such lockdown policies in Romania because of the government's strategy of implementing lockdowns in different geographical regions over time. Because CoronaNet also engages in original data collection, such policy gaps can be filled in in conjunction with our data harmonization efforts, although not in the WHO data harmonization efforts. These observations match with our experience that PHSM policies can be very complex and require a) experts who can do the research to substantiate not only policies that are in external dtasets but which the governments have actually implemented b) evaluate and clean existing policies in external datasets in c) a standardized manner.

Finally problems in the WHO PHSM dataset with regards to missing raw sources and lack of transparency around the data generation process hinders the ability to evaluate the validity of the WHO PHSM dataset. In the current WHO PHSM dataset, there are 5700 missing links, 20k+ additional links that the WHO PHSM have found to be dead, and 25k+ links which the WHO declared as being 'unknown' in terms of whether they are live or not, but for which no follow up sources are provided. In contrast, CoronaNet only includes data points that have a working link or a screenshot of the original PDF source attached. When CoronaNet research assistants encounter missing or dead links as part of the data harmonization process, they are instructed to either attempt to recover active links with the same information (to date, around 5.8% of the harmonized data) or the observation is not included in the dataset (around 7.3% of the external data). Access to the raw sources is paramount for researchers to independently ascertain the validity and reliability of the subsequent data coded. With regards to the WHO PHSM data generation process, though in XXX (2020) they provide a basic description of how they process the data, given the issues with data quality outlined above, greater transparency as to what criterion they use to determine that "the clean, verified data is ready to be shared with WHO and other researchers." would be welcome.

C.3 Discussion

By laying out the contrast between our data harmonization effort and the WHO data harmonization effort, we hope that readers gain not only a deeper appreciation of the complexity of harmonizing PHSM data, but also for the relative merits of our efforts. Given the volume and complexity of PHSM data as well as the reality of limited resources, we believe that our decision to harmonize data for a more limited period of time results in a higher quality, more complete dataset that can provide a more rigorous foundation for researcher on the COVID-19 pandemic.

References

[1] Demchenko, Y., Zhao, Z., Grosso, P., Wibisono, A., De Laat, C.: Addressing big data challenges for scientific data infrastructure. In: 4th IEEE International Conference on Cloud Computing Technology and Science Proceedings, pp. 614–617 (2012). IEEE

[2] Ruggles, S.: The minnesota population center data integration projects: Challenges of harmonizing census microdata across time and place. In: In Proceedings of the American Statistical Association, Government Statistics Section, pp. 1405–1415 (2006). Citeseer

[3] Elshawi, R., Sakr, S., Talia, D., Trunfio, P.: Big data systems meet machine learning challenges: towards big data science as a service. Big data research **14**, 1–11 (2018)

[4] Solt, F.: The standardized world income inequality database. Social science quarterly **97**(5), 1267–1281 (2016)

[5] Hoffmeyer-Zlotnik, J.H.: Standardisation and harmonisation of socio-demographic variables (version 2.0) (2016)

[6] Salguero-Gómez, R., Jackson, J., Gascoigne, S.J.: Four key challenges in the open-data revolution. Journal of Animal Ecology **90**(9), 2000–2004 (2021)

[7] Ahmed, E., Yaqoob, I., Hashem, I.A.T., Khan, I., Ahmed, A.I.A., Imran, M., Vasilakos, A.V.: The role of big data analytics in internet of things. Computer Networks **129**, 459–471 (2017)

[8] Cuzzocrea, A., Song, I.-Y., Davis, K.C.: Analytics over large-scale multidimensional data: the big data revolution! In: Proceedings of the ACM 14th International Workshop on Data Warehousing and OLAP, pp. 101–104 (2011)

[9] Cai, L., Zhu, Y.: The challenges of data quality and data quality assessment in the big data era. Data science journal **14** (2015)

[10] Turner, M.C., Mehlum, I.S.: Greater coordination and harmonisation of European occupational cohorts is needed. BMJ Publishing Group Ltd (2018)

[11] Zhu, A.H., Moyer, D.C., Nir, T.M., Thompson, P.M., Jahanshad, N.: Challenges and opportunities in dmri data harmonization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 157–172 (2019). Springer

[12] Bergeron, J., Massicotte, R., Atkinson, S., Bocking, A., Fraser, W., Fortier, I., member cohorts' principal investigators, R.: Cohort profile: Research advancement through cohort cataloguing and harmonization (reach). International journal of epidemiology **50**(2), 396–397 (2021)

[13] Stowell, D., Kelly, J., Tanner, D., Taylor, J., Jones, E., Geddes, J., Chalstrey, E.: A harmonised, high-coverage, open dataset of solar photovoltaic installations in the uk. Scientific Data **7**(1), 1–15 (2020)

[14] Zagorecki, A.T., Johnson, D.E., Ristvej, J.: Data mining and machine learning in the context of disaster and crisis management. International Journal of Emergency Management **9**(4), 351–365 (2013)

[15] Chen, Y., Sabri, S., Rajabifard, A., Agunbiade, M.E.: An ontology-based spatial data harmonisation for urban analytics. Computers, Environment and Urban Systems **72**, 177–190 (2018)

[16] Tax, C.M., Grussu, F., Kaden, E., Ning, L., Rudrapatna, U., Evans, C.J., St-Jean, S., Leemans, A., Koppers, S., Merhof, D., *et al.*: Cross-scanner and cross-protocol diffusion mri data harmonisation: A benchmark database and evaluation of algorithms. NeuroImage **195**, 285–299 (2019)

[17] Chen, T., Abadi, A.J., Lê Cao, K.-A., Tyagi, S.: multiomics: A user-friendly multi-omics data harmonisation r pipeline. F1000Research **10**(538), 538 (2021)

[18] Wey, T.W., Doiron, D., Wissa, R., Fabre, G., Motoc, I., Noordzij, J.M., Ruiz, M., Timmermans, E., Van Lenthe, F.J., Bobak, M., *et al.*: Overview of retrospective data harmonisation in the mindmap project: process and results. J Epidemiol Community Health **75**(5), 433–441 (2021)

[19] Lloyd, C.T., Chamberlain, H., Kerr, D., Yetman, G., Pistolesi, L., Stevens, F.R., Gaughan, A.E., Nieves, J.J., Hornby, G., MacManus, K., *et al.*: Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets. Big earth data **3**(2), 108–139 (2019)

[20] ACAPS: Acaps government measures dataset readme version 1.1 (2020)

[21] Katz, R., Graedn, E.: Covidamp.org: A global resource for governance in the time of covid. Research brief 2020-07, Washington: Georgetown University Center for Global Health Science & Security (2020)

[22] for Health Information, C.I.: Canadian data set of covid-19 interventions. Technical report, Canadian Institute for Health Information (2021)

[23] Cheng, C., Barceló, J., Hartnett, A.S., Kubinec, R., Messerschmidt, L.: Covid-19 government response event dataset (coronanet v. 1.0). Nature human behaviour **4**(7), 756–768 (2020)

[24] Zheng, Q., Jones, F.K., Leavitt, S.V., Ung, L., Labrique, A.B., Peters, D.H., Lee, E.C., Azman, A.S.: Hit-covid, a global database tracking public health interventions to covid-19. Scientific data **7**(1), 1–8 (2020)

[25] Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S., *et al.*: A global panel database of pandemic policies (oxford covid-19 government response tracker). Nature Human Behaviour **5**(4), 529–538 (2021)

[26] Organization, W.H., of Hygiene, L.S., Medicine, T.: Public health and

safety measures. Technical report, World Health Organization (2022)

[27] Zeb, A., Soininen, J.-P., Sozer, N.: Data harmonisation as a key to enable digitalisation of the food sector: A review. Food and Bioproducts Processing **127**, 360–370 (2021)

[28] Boyden, J., Walnicki, D.: Leveraging the Power of Longitudinal Data: Insights on Data Harmonisation and Linkage from Young Lives. Oxford: Young Lives (2021)

[29] Parimbelli, E., Larizza, C., Urosevic, V., Pogliaghi, A., Ottaviano, M., Cheng, C., Benoit, V., Pala, D., Casella, V., Bellazzi, R., *et al.*: The periscope data atlas: A demonstration of release v1. 2. In: International Conference on Artificial Intelligence in Medicine, pp. 412–415 (2022). Springer

[30] Pala, D., Parimbelli, E., Larizza, C., Cheng, C., Ottaviano, M., Pogliaghi, A., Đukić, G., Jovanović, A., Milićević, O., Urošević, V., *et al.*: A new interactive tool to visualize and analyze covid-19 data: The periscope atlas. International Journal of Environmental Research and Public Health **19**(15), 9136 (2022)

[31] Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A., Tarczy-Hornoch, P.: Data integration and genomic medicine. Journal of biomedical informatics **40**(1), 5–16 (2007)

[32] Fichtinger, A., Rix, J., Schäffler, U., Michi, I., Gone, M., Reitz, T.: Data harmonisation put into practice by the humboldt project. International Journal of Spatial Data Infrastructures Research **6**, 234–260 (2011)

[33] Fortier, I., Doiron, D., Burton, P., Raina, P.: Invited commentary: consolidating data harmonization—how to obtain quality and applicability? American journal of epidemiology **174**(3), 261–264 (2011)

[34] Fortier, I., Raina, P., Van den Heuvel, E.R., Griffith, L.E., Craig, C., Saliba, M., Doiron, D., Stolk, R.P., Knoppers, B.M., Ferretti, V., *et al.*: Maelstrom research guidelines for rigorous retrospective data harmonization. International journal of epidemiology **46**(1), 103–105 (2017)

[35] Stewart, L.A., Tierney, J.F.: To ipd or not to ipd? advantages and disadvantages of systematic reviews using individual patient data. Evaluation & the health professions **25**(1), 76–97 (2002)

[36] Nan, Y., Del Ser, J., Walsh, S., Schönlieb, C., Roberts, M., Selby, I., Howard, K., Owen, J., Neville, J., Guiot, J., et al.: Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic

review, meta-analysis and future research directions. Information Fusion (2022)

[37] Dewitte, O., Jones, A., Spaargaren, O., Breuning-Madsen, H., Brossard, M., Dampha, A., Deckers, J., Gallali, T., Hallett, S., Jones, R., *et al.*: Harmonisation of the soil map of africa at the continental scale. Geoderma **211**, 138–153 (2013)

[38] Bagepally, B.S., Chaiyakunapruk, N., Attia, J., Thakkinstian, A., *et al.*: Meta-analysis of economic evaluation studies: data harmonisation and methodological issues. BMC health services research **22**(1), 1–10 (2022)

[39] Lopes, C., Quental, L., De Oliveira, D.P.S., Filipe, A., Pereira, A.: Inspire data harmonisation of mineral resources: contribution of minerals4eu project. Mapping: Revista Internacional de Geomática y Ciências de la Tierra **27**(187), 56–63 (2018)

[40] Kveder, A., Galico, A.: Guidelines for cleaning and harmonization of Generation and Gender Survey data. Generations and Gender Programme (2008)

[41] Hutchinson, D.M., Silins, E., Mattick, R.P., Patton, G.C., Fergusson, D.M., Hayatbakhsh, R., Toumbourou, J.W., Olsson, C.A., Najman, J.M., Spry, E., Tait, R., Degenhardt, L., Swift, W., Butterworth, P., Horwood, L., Consortium., C.C.R.: How can data harmonisation benefit mental health research? An example of the Cannabis Cohorts Research Consortium. Australian & New Zealand Journal of Psychiatry **49**(4), 317–323 (2015)

[42] Almeida, J.R., Silva, L.B., Bos, I., Visser, P.J., Oliveira, J.L.: A methodology for cohort harmonisation in multicentre clinical research. Informatics in Medicine Unlocked **27**, 100760 (2021)

[43] Bath, P.A., Deeg, D., Poppelaars, J.: The harmonisation of longitudinal data: a case study using data from cohort studies in The Netherlands and the United Kingdom. Ageing & Society **30**(8), 1419–1437 (2010)

[44] Dinsdale, N.K., Jenkinson, M., Namburete, A.I.: Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. NeuroImage **228**, 117689 (2021)

[45] Cooper, R., Hardy, R., Aihie Sayer, A., Ben-Shlomo, Y., Birnie, K., Cooper, C., Craig, L., Deary, I.J., Demakakos, P., Gallacher, J., McNeill, G., Martin, R., Starr, J., Steptoe, A., Kuh, D., HALCyon, S.T.: Age and gender differences in physical capability levels from mid-life onwards: the harmonisation and meta-analysis of data from eight UK cohort studies. PloS one **6**(11), 27899 (2011)

[46] Tomson, T., Battino, D., Craig, J., Hernandez-Diaz, S., Holmes, L.B., Lindhout, D., Morrow, J., French, J., on Therapeutic Strategies, I.C.: Pregnancy registries: differences, similarities, and possible harmonization. Wiley Online Library (2010)

[47] Christensen, L., Hubert, J.-P., Järvi, T., Kagerbauer, M., Vázquez, N.S., Weiß, C.: Improving comparability of survey results through ex-post harmonisation a case study with twelve european national travel surveys. In: 10th International Conference on Transport Survey Methods (2014)

[48] Sanchez-Niubo, A., Egea-Cortés, L., Olaya, B., Caballero, F.F., Ayuso-Mateos, J.L., Prina, M., Bobak, M., Arndt, H., Tobiasz-Adamczyk, B., Pająk, A., Leonardi, M., Koupil, I., Panagiotakos, D., Tamosiunas, A., Scherbov, S., Sanderson, W., Koskinen, S., Chatterji, S., Haro, J., Consortium., A.: Cohort profile: the ageing trajectories of health–longitudinal opportunities and synergies (ATHLOS) project. International journal of epidemiology **48**(4), 1052–1053 (2019)

[49] Przeworski, A., Teune, H.: The Logic of Comparative Social Inquiry, Florida. Robert E. Krieger (1970)

[50] Dawes, A.: Measuring the development of cognitive skills across time and context: Reflections from young lives. Young Lives, 1–12 (2020)

[51] Hurtt, G.C., Chini, L.P., Frolking, S., Betts, R., Feddema, J., Fischer, G., Fisk, J., Hibbard, K., Houghton, R., Janetos, A., Jones, C.D., Kindermann, G., Kinoshita, T., Goldewijk, K.K., Riahi, K., Shevliakova, E., Smith, S., Stehfest, E., Thomson, A., Thornton, P., van Vuuren, P., Wang, Y.P.: Harmonization of land-use scenarios for the period 1500–2100: 600 years of global gridded annual land-use transitions, wood harvest, and resulting secondary lands. Climatic change **109**(1), 117–161 (2011)

[52] Roberts, G., Binder, D.: Analyses based on combining similar information from multiple surveys. In: Survey Research Methods Section of the Joint Statistical Meetings (JSM), pp. 2138–2147 (2009)

[53] Tomescu-Dubrow, I., Slomczynski, K.M.: Democratic values and protest behavior: Data harmonization, measurement comparability, and multilevel modeling in cross-national perspective **23**, 103–114 (2014)

[54] Slomczynski, K.M., Tomescu-Dubrow, I., Wysmulek, I.: Survey data quality in analyzing harmonized indicators of protest behavior: A survey data recycling approach. American Behavioral Scientist **66**, 412–433 (2021)

[55] Kołczyńska, M.: Combining multiple survey sources: A reproducible

workflow and toolbox for survey data harmonization. Methodological Innovations **15**(1), 62–72 (2022)

[56] Wysmułek, I., Tomescu-Dubrow, I., Kwak, J.: Ex-post harmonization of cross-national survey data: advances in methodological and substantive inquiries. Quality & Quantity **56**(3), 1701–1708 (2022)

[57] Wittwer, R., Hubrich, S., Wittig, S., Gerike, R.: Development of a new method for household travel survey data harmonisation. Transportation research procedia **32**, 597–606 (2018)

[58] Armoogum, J., Bonsall, P., Browne, M., Christensen, L., Cools, M., Cornelis, E., Diana, M., Harder, H., Reinau, K.H., Hubert, J.-P., Kagerbauer, M., Kuhnimhof, T., Madre, J.-L., Moiseeva, A., Polak, J., Reinau, C.H., Schulz, A., Tébar, M., Vidalakis, L.: Survey harmonisation with new technologies improvement (SHANTI) (2014)

[59] Hughes, C., Zagheni, E., Abel, G.J., Sorichetta, A., Wi'sniowski, A., Weber, I., Tatem, A.J.: Inferring migrations: traditional methods and new approaches based on mobile phone, social media, and other big data: feasibility study on inferring (labour) mobility and migration in the European Union from big data and social media data (2016)

[60] Hamilton, C.M., Strader, L.C., Pratt, J.G., Maiese, D., Hendershot, T., Kwok, R.K., Hammond, J.A., Huggins, W., Jackman, D., Pan, H., *et al.*: The phenx toolkit: get the most from your measures. American Journal of Epidemiology **174**(3), 253–260 (2011)

[61] De Frahan, B.H., Vancauteren, M.: Harmonisation of food regulations and trade in the single market: evidence from disaggregated data. European Review of Agricultural Economics **33**(3), 337–360 (2006)

[62] Veermani, C.: Analysing trade flows and industrial structure of india: the question of data harmonisation. Unpublished working paper, Centre for Development Studies, Trivandrum, India (2001)

[63] Owino, B.: Harmonising data systems for cash transfer programming in emergencies in Somalia. Journal of International Humanitarian Action **5**(1), 1–16 (2020)

[64] Mattli, W., Büthe, T.: Setting international standards: technological rationality or primacy of power? World Politics **56**(1), 1–42 (2003)

[65] Haver, M.A.: The statistics corner: The NAICS is coming. Will we be ready? Business Economics **32**(4), 63–65 (1997)

[66] Desrosières, A.: Measurement and its uses: Harmonization and quality in

social statistics. International Statistical Review **68**(2), 173–187 (2000)

[67] Mügge, D.: Studying macroeconomic indicators as powerful ideas. In: Ideas, Political Power, and Public Policy, pp. 108–125 (2018). Routledge Taylor & Francis Group.

[68] Jerven, M.: An unlevel playing field: national income estimates and reciprocal comparison in global economic history. Journal of Global History **7**(1), 107–128 (2012)

[69] Uphoff, H., Cohen, J., Fleming, D., Noone, A.: Harmonisation of national influenza surveillance morbidity data from EISS: a simple index. Euro Surveillance **8**(7), 156–164 (2003)

[70] Button, K.S., Ioannidis, J., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafò, M.R.: Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience **14**(5), 365–376 (2013)

[71] Griffith, L.E., Van Den Heuvel, E., Fortier, I., Sohel, N., Hofer, S.M., Payette, H., Wolfson, C., Belleville, S., Kenny, M., Doiron, D., *et al.*: Statistical approaches to harmonize data on cognitive measures in systematic reviews are rarely reported. Journal of clinical epidemiology **68**(2), 154–162 (2015)

[72] Fortier, I., Doiron, D., Little, J., Ferretti, V., L'Heureux, F., Stolk, R.P., Knoppers, B.M., Hudson, T.J., Burton, P.R.: Is rigorous retrospective harmonization possible? application of the datashaper approach across 53 large studies. International journal of epidemiology **40**(5), 1314–1328 (2011)

[73] Sekot, W., Fillbrandt, T., Zesiger, A.: Improving the international compatibility of accountancy data: the 'dach-initiative'. Small-scale Forestry **10**(2), 255–269 (2011)

[74] Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., *et al.*: Harmonization of resting-state functional mri data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. PLoS biology **17**(4), 3000042 (2019)

[75] Maikusa, N., Zhu, Y., Uematsu, A., Yamashita, A., Saotome, K., Okada, N., Kasai, K., Okanoya, K., Yamashita, O., Tanaka, S.C., *et al.*: Comparison of traveling-subject and combat harmonization methods for assessing structural brain characteristics. Human brain mapping **42**(16), 5278–5287 (2021)

[76] Baume, O., Skøien, J.O., Heuvelink, G.B., Pebesma, E.J., Melles, S.: A geostatistical approach to data harmonization–application to radioactivity exposure data. International Journal of Applied Earth Observation and Geoinformation **13**(3), 409–419 (2011)

[77] Bento, M., Fantini, I., Park, J., Rittner, L., Frayne, R.: Deep learning in large and multi-site structural brain mr imaging datasets. Frontiers in Neuroinformatics **15**, 82 (2022)

[78] Moyer, D., Ver Steeg, G., Tax, C.M., Thompson, P.M.: Scanner invariant representations for diffusion mri harmonization. Magnetic resonance in medicine **84**(4), 2174–2189 (2020)

[79] Wachinger, C., Rieckmann, A., Pölsterl, S., Initiative, A.D.N., *et al.*: Detect and correct bias in multi-site neuroimaging datasets. Medical Image Analysis **67**, 101879 (2021)

[80] Baume, O.P., SKØIEN, J., Heuvelink, G.B., Pebesma, E.J.: Data harmonization with geostatistical tools: a bayesian extension. Geostats2008, Santiago, Chile (2008)

[81] Borisov, N., Buzdin, A.: Transcriptomic harmonization as the way for suppressing cross-platform bias and batch effect. Biomedicines **10**(9), 2318 (2022)

[82] Adhikari, K., Patten, S.B., Patel, A.B., Premji, S., Tough, S., Letourneau, N., Giesbrecht, G., Metcalfe, A.: Data harmonization and data pooling from cohort studies: a practical approach for data management. International journal of population data science **6**(1) (2021)

[83] Borenstein, M., Hedges, L.V., Higgins, J.P., Rothstein, H.R.: Introduction to Meta-analysis. John Wiley & Sons, ??? (2021)

[84] Rao, S.R., Graubard, B.I., Schmid, C.H., Morton, S.C., Louis, T.A., Zaslavsky, A.M., Finkelstein, D.M.: Meta-analysis of survey data: application to health services research. Health Services and Outcomes Research Methodology **8**(2), 98–114 (2008)

[85] Eisenhauer, J.G.: Meta-analysis and mega-analysis: A simple introduction. Teaching Statistics **43**(1), 21–27 (2021)

[86] Griffith, L., van den Heuvel, E., Fortier, I., Hofer, S., Raina, P., Sohel, N., Payette, H., Wolfson, C., Belleville, S.: Harmonization of cognitive measures in individual participant data and aggregate data meta-analysis (2013)

[87] Luningham, J.M., McArtor, D.B., Hendriks, A.M., van Beijsterveldt, C.E., Lichtenstein, P., Lundström, S., Larsson, H., Bartels, M., Boomsma, D.I., Lubke, G.H.: Data integration methods for phenotype harmonization in multi-cohort genome-wide association studies with behavioral outcomes. Frontiers in genetics **10**, 1227 (2019)

[88] Boedhoe, P.S., Schmaal, L., Abe, Y., Ameis, S.H., Arnold, P.D., Batistuzzo, M.C., Benedetti, F., Beucke, J.C., Bollettini, I., Bose, A., *et al.*: Distinct subcortical volume alterations in pediatric and adult ocd: a worldwide meta-and mega-analysis. American Journal of Psychiatry **174**(1), 60–69 (2017)

[89] Jahanshad, N., Ganjgahi, H., Bralten, J., den Braber, A., Faskowitz, J., Knodt, A.R., Lemaitre, H., Nir, T.M., Patel, B., Richie, S., et al.: Do candidate genes affect the brain's white matter microstructure? large-scale evaluation of 6,165 diffusion mri scans. BioRxiv, 107987 (2017)

[90] Burke, D.L., Ensor, J., Riley, R.D.: Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. Statistics in medicine **36**(5), 855–875 (2017)

[91] Siddique, J., Reiter, J.P., Brincks, A., Gibbons, R.D., Crespi, C.M., Brown, C.H.: Multiple imputation for harmonizing longitudinal non-commensurate measures in individual participant data meta-analysis. Statistics in medicine **34**(26), 3399–3414 (2015)

[92] Kunkel, D., Kaizar, E.E.: A comparison of existing methods for multiple imputation in individual participant data meta-analysis. Statistics in medicine **36**(22), 3507–3532 (2017)

[93] Shishegar, R., Cox, T., Rolls, D., Bourgeat, P., Doré, V., Lamb, F., Robertson, J., Laws, S.M., Porter, T., Fripp, J., *et al.*: Using imputation to provide harmonized longitudinal measures of cognition across aibl and adni. Scientific reports **11**(1), 1–11 (2021)

[94] Siddique, J., de Chavez, P.J., Howe, G., Cruden, G., Brown, C.H.: Limitations in using multiple imputation to harmonize individual participant data for meta-analysis. Prevention Science **19**(1), 95–108 (2018)

[95] Pasteels, J.-M.: Review of best practice methodologies for imputing and harmonising data in cross-country datasets. ILO internal report (2013)

[96] Gibbons, R.D., Perraillon, M.C., Kim, J.B.: Item response theory approaches to harmonization and research synthesis. Health Services and Outcomes Research Methodology **14**(4), 213–231 (2014)

[97] Curran, P.J., Hussong, A.M., Cai, L., Huang, W., Chassin, L., Sher,

K.J., Zucker, R.A.: Pooling data from multiple longitudinal studies: the role of item response theory in integrative data analysis. Developmental psychology **44**(2), 365 (2008)

[98] Gross, A.L., Tommet, D., D'Aquila, M., Schmitt, E., Marcantonio, E.R., Helfand, B., Inouye, S.K., Jones, R.N.: Harmonization of delirium severity instruments: a comparison of the drs-r-98, mdas, and cam-s using item response theory. BMC medical research methodology **18**(1), 1–18 (2018)

[99] Cheng, C., Desvars-Larrive, A., Ebbinghaus, B., Hale, T., Howes, A., Lehner, L., Messerschmidt, L., Nika, A., Penson, S., Petherick, A., *et al.*: Capturing the covid-19 crisis through public health and social measures data science. Scientific Data **9**(1), 1–9 (2022)

[100] Veldsman, W.P., Campli, G., Dind, S., de Laval, V.R., Drage, H., Waterhouse, R.M., Robinson-Rechavi, M.: Taxonbridge: an r package to create custom taxonomies based on the ncbi and gbif taxonomies. bioRxiv (2022)

[101] Cohen, J.A., Trojano, M., Mowry, E.M., Uitdehaag, B.M., Reingold, S.C., Marrie, R.A.: Leveraging real-world data to investigate multiple sclerosis disease behavior, prognosis, and treatment. Multiple Sclerosis Journal **26**(1), 23–37 (2020)

[102] Edgell, A., Lührmann, A., Maerz, S.F., Lachapelle, J., Grahn, S., God, A.F.G., Lundstedt, M., Natsika, N., Kaiser, S., Kolvani, P., et al.: Pandemic backsliding: Democracy during covid-19 (pandem), version 6. Technical report, Varieties of Democracy (V-Dem) Institute (2020)

[103] Kubinec, R., Barceló, J., Goldszmidt, R., Grujic, V., Model, T., Schenk, C., Cheng, C., Hale, T., Hartnett, A.S., Messerschmidt, L., et al.: Statistically validated indices for covid-19 public health policies (2021)

[104] Parmesan, S., Scaiella, U., Barbera, M., Tarasova, T.: Dandelion: from raw data to datagems for developers. In: ISWC (Developers Workshop), pp. 1–6 (2014)

[105] O'Connor, C., Joffe, H.: Intercoder reliability in qualitative research: debates and practical guidelines. International journal of qualitative methods **19**, 1609406919899220 (2020)

[106] Miles, M.B., Huberman, A.M.: Qualitative Data Analysis: An Expanded Sourcebook. sage, ??? (1994)

[107] Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics, 159–174 (1977)

[108] WHO: Global dataset of public health and social measures data harmonization, processing flow, and data dictionaries for stage 1 and stage 2 databases)

[109] Dunn, P., Allen, L., Cameron, G., Alderwick, H.: The Health Foundation, COVID-19 policy tracker: A timeline of national policy and health system responses to COVID-19 in England. 2020

[110] Skinner, A., Flannery, K., Nocka, K., Bor, J., Dean, L.T., Jay, J., Lipson, S.K., Cole, M.B., Benfer, E.A., Scheckman, R., *et al.*: A database of us state policies to mitigate covid-19 and its economic consequences. BMC Public Health **22**(1), 1–8 (2022)

[111] Fullman, N., Bang-Jensen, B., Reinke, G., Magistro, B., Castellano, R., Erickson, M., Walcott, R., Dapper, C., Amano, K., Wilkerson, J., Adolph, C.: State-level social distancing policies in response to covid-19 in the us, version 1.145. Technical report (2020)

[112] Boehmke, F.J., Desmarais, B.A., Eastman, A., Grassel, I., Harden, J.J., Harper, S., Kaboli, L., Ko, H., Oster, E., Saunders, T.M.: Sprc19: A database tracking us state policy responses to covid-19 (2022)

[113] Spiegel, M.: Yale som-tobin center state and local covid restriction database. Technical report, Yale SOM-TObin Center (2020)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryinformationdataharmonizationphsm.pdf