# Global dataset of public health and social measures

**Data harmonization, processing flow, and data dictionaries for Stage 1 and Stage 2 databases**

The global dataset of public health and social measures (PHSM) is a collaboration between WHO, the London School of Hygiene and Tropical Medicine (LSHTM) and the organizations providing datasets: WHO (Geneva, Switzerland), ACAPS (Geneva, Switzerland), University of Oxford (Oxford, UK), the Global Public Health Intelligence Network (GPHIN) of Public Health Agency of Canada (Ottawa, Canada), the Complexity Science Hub Vienna (Vienna, Austria), the United States Centers for Disease Control and Prevention (CDC, Atlanta, USA) and Johns Hopkins University (JHU, Maryland, USA).[1] LSHTM has developed the pipelines for importing and formatting the six datasets to create the global dataset.

This dataset is open content and available for any use. Users should be sure to familiarize themselves with the disclaimer provided on the landing page for the dataset.

This project has two stages and different databases are associated with each stage: in stage 1, the datasets from each of the seven providers were formatted and coded but not cleaned or altered, apart from standardising country, territory or area names to those used by the WHO. This database has duplicate records if more than one dataset reported the same measure. In stage 2, duplicates were removed, PHSM coding was validated and verified beyond what the dataset providers have already done, additional variables were coded for each measure, and records will be closed if the measure is no longer in place.

## Stage 1

This section describes the stage 1 data flow for each dataset provider into the format required for the harmonised, global PHSM dataset.

Data flow (conducted by LSHTM):

1. Permission obtained from the providers to include the dataset and agree on how the dataset will be supplied by the provider.
2. Datasets placed in Slack folder and in group GITHUB folder.
3. Data dictionaries produced for each providers' dataset.
4. Unique provider ID added for each record where not included in the providers' original datasets.
5. Data transformed, where needed, from country per row to country-measure per row and coding of the measure checked with another member of the LSHTM team.
6. Country names verified against the official WHO list of Member States and country codes added.
7. Provider variables linked to the WHO global dataset taxonomy.
8. Provider dataset converted into the WHO global dataset format by code.

[1] We are working with other organizations to incorporate their datasets and will update this document whenever a new group is added.

9. Full coding checked and coding structure cleaned and standardised.
10. Unique "category, sub-category, measure" list produced and linked to WHO taxonomy. Where a WHO category does not exist for a measure, the record is categorized as 'Other measures'.
11. Final WHO dataset table produced.

When permission to use a dataset was obtained, the method to access or supply the dataset was determined with the provider (Table 1). In most cases an initial dataset in excel or csv format was supplied by email to WHO or LSHTM and this was used for establishing the formatting pipeline. Each dataset was assigned to one LSHTM volunteer coder with a senior member checking their code at the end of each stage. Once this code was checked and any issues corrected, an update pipeline for the dataset was established (i.e., sending via email or downloading from online site) (Table 1).

Table 1: Original and update pipeline supply method

| **Provider** | **Original data** | **Updates** |
|---|---|---|
| ACAPS | Emailed file | Website |
| Oxford | Website | Website |
| GPHIN | Emailed files | Email updates |
| CSH Vienna | Emailed CSV file | Website |
| WHO IHR | Emailed file | Email updates |
| CDC | Emailed file | Email updates |
| JHU | Website | Website |

**Conversion to WHO global dataset structure**

The type and amount of formatting required to format data from each provider into the WHO global dataset structure varied. There were four main areas of formatting and coding:

- Development of unique record IDs
- Harmonising variable names
- Standardising names for country, territory or area
- Conversion and coding from country per row to measure per row

Table 2: Formatting and coding required for each dataset

| Provider | Unique ID | Formatting | |
|---|---|---|---|
| | | Conversion required | Extra coding required |
| ACAPS | Provided | No | NA |
| Oxford | Generated | Yes | Yes |
| GPHIN | Generated | Yes | No |
| CSH Vienna | Provided | No | NA |
| WHO IHR | Generated | Yes | Yes |
| CDC | Provided | No | NA |
| JHU | Provided | Yes | No |

Here we describe the types of formatting and coding required. Full descriptions of the dataset provided by each provider, along with more details about the coding and formatting required, are supplied at the end.

**Unique identification numbers**

To allow feedback and comparison to the original dataset, a unique identifier (ID) was required for each record in the original dataset. Where IDs were not provided simple ID numbers were assigned to each record based on the row number. It is recognised this simple ID number will not allow linkage back to the original dataset, particularly if data are added retrospectively or if substantial changes in format are made to the original data. For the purposes of the WHO global dataset, this unique ID is most important to allow us to check that re-formatting of the data works correctly.

**Conversion to official WHO Member State names**

Country, territory and area names must follow the WHO guidelines for Member States; correction of country, territory and area names was the only part of the original data from each provider that had to be edited. All providers supplied a column relating to the country, territory or area the row related to; this name had to be standardised. A column called country_territory_area contains the name of the country, territory or area the measure relates to. Where a country code was provided, this code was used to fill the field with the official WHO Member State name. Where only text was provided, a combination of methods was used related to packages in R that standardises names. In some cases, e.g. where the name was the Diamond Princess, decisions were made as to whether this record should be removed because it was not suitable for the global dataset being produced, or that it should be assigned to the country, territory or area with the original name being placed in a second variable used to report where the measure related to sub-national area.

**Conversion into a harmonised format**

The format for the WHO global table is that each measure introduced is represented as a row in the table. Several of the datasets were structured with rows representing countries, websites or reports, with information on measures included as different variables. These datasets had to be converted into the required format. The amount of formatting required varied for each dataset is shown in Table 2.

There were two forms of conversion of provider datasets:

- Straight conversion: the original dataset contained a row per country with a set of variables representing different measures. An X in a field indicated the implementation of that measure in the country represented by the row. These datasets were converted to one row per country-measure.
- Conversion and coding: In some datasets, not only were multiple measures included in the same row, but the measures needed to be coded as well. The exact coding depended on the data set but for some involved combining columns to generate the WHO category, or coding level of enforcement.

**Harmonising taxonomy**

The final step was to harmonise the categories used to describe the PHSM. WHO produced a taxonomy of PHSM which contained a coded list of measures split into categories and in some cases sub-categories. This list contained only PHSM with potential direct effects on COVID-19 transmission, and did not include policies designed to promote PHSM, such as economic incentives, teleworking policies or actions to scale up response elements. An additional category was included to cover these measures so that records were not lost from the original datasets. Two final categories were also added: “coding required”, where it was not possible to assign to a WHO category without examining every record, and a second one for “Not required”. This final category relates to the fact that some datasets, especially the one automatically extracting information from social media and the internet, pick up new reports that do not relate to an intervention or actual measure; e.g. 20 people being repatriated, or someone making a donation to a country. In stage 2, when records will be validated, this category will be used to identify records that will be excluded.

For each dataset a list of unique category-sub-category-measure was produced and the LSHTM team lead produced a Vlookup sheet from the provider information to the WHO taxonomy. In the case of the CSH Vienna data for which there were many records, a small group of LSHTM volunteers each coded a section of the data, with 3 people coding each measure. The LSHTM team lead then went through and selected the agreed coding for that measure based on the majority choices. In some cases, it was not possible to code down to the exact measure; e.g. a measure in the original dataset may be coded as ‘travel restriction’, which might be related to several different measures in the WHO taxonomy. In these cases, the measure in the original dataset was coded to the lowest level possible; in stage 2, further coding will be done.

This Vlookup table was then transferred in R to allow automated production of the final whole table for that dataset.

**Full dataset and updates**

The steps described above for the formatting and coding of the Stage 1 dataset for each provider is completely automated. Each time an updated dataset is provided, the dataset is run through the pipeline and a table in the required format is output and can be combined with data from other providers. A simple data dictionary for the Stage 1 dataset is provided in Table 3. Each data provider was sent their reformatted dataset for comment should they wish to. Where formatting was required, this was explained along with the coding used. The Stage 1 dataset contains duplicate records but, in an effort to provide all information available, this dataset is preserved with duplicates.

## Stage 2

To generate a de-duplicated database where extensions or curtailments of measures are linked, the Stage 1 database was subjected to a series of cleaning functions to verify information, correct records, supplement the source data and link records relating to the same measure. The database was divided by country into separate Excel files that were assigned to a team of volunteers consisting of students, staff and alumni of LSHTM. There was an attempt to match volunteers by language and familiarity with the country whose data they cleaned.

Within each country blocks, volunteers verified that the WHO code (who_code) given to the measure was correct by reading the comment field and checking the links. If there was insufficient information to verify the code, or the comment suggested the link referred to general information rather than an intervention, the who_code was set to either 'Not enough to code' or 'Not of interest'. Where the who_code was believed to be incorrect, it was changed to the correct code. The fields link_live and link_eng were populated to indicate if the link was live, and if the link was in English. The comment field was amended where the comment was excessively long or contained information that was not relevant.

Next, the comment field was used either to verify or, in some cases, populate the admin_level and area_covered fields, if the record referred to a subnational level. An additional permitted value for admin_level (state) was added for the largest administrative level within the country (admin level 1). 'State' was only used where the measure related to a single relevant area: where a measure related to two or more admin level 1 areas, the admin_level was set to 'Other'. Where the measure related to any other area (e.g., city, small geographical area, multiple areas) the admin_level was set to 'Other'. Area_covered was left blank if admin_level was 'national'; where area_covered was 'state' or 'other', area_covered contains the names of the relevant area(s).

Following or concurrent with the above step, the comment and link fields were used either to verify or populate the target and enforcement fields where possible. 'Target' was populated with the specific number, place or population referred to; for example, type of curfew, the number of persons gatherings are restricted to, country names for restricted entry, types of business or schools closed, etc.

The ‘Enforcement’ variable indicates the level of enforcement that authorities were applying to the measure. ‘Monitored’ is the most stringent of enforcement levels for measures like quarantining at home and their continued presence is monitored, or for cordon sanitaires. Other values were set to ‘required’ or ‘recommended’, with ‘not applicable’ (e.g. public awareness campaigns) or ‘not known’ where there was insufficient information in the comments and link.

Each volunteer then processed the data in their file, looking for duplicate information either within or between datasets; i.e., looking for interventions recorded multiple. They accomplished identification of duplicates by comparing dates and text descriptions.

Care was taken to ensure that the volunteers understood that measures could be changed on a daily basis, and that changes constituted new records. Therefore, volunteers were instructed to check if similar records within a small date range were simply dates being recorded differently or if they were actual changes in the measure. Where they found presumed duplicates, they linked these records to each other by taking the unique record ID number for a specific record and entering it into the appropriate column created in Stage 1 for all other records related to that intervention. This process was repeated for each of the linked interventions. To decide on the ‘best’ source (link) to carry forward, the volunteers used a crib sheet (see Table 2) for a sample of suggested source ranking.

**Table 2. Suggested ranking of original sources of data**

- Group 1
    - Ministry of health websites (MOH)
    - Federal or territorial government websites
    - Regional offices of WHO
    - Press releases from federal or territorial government websites or MOH
- Group 2
    - Trusted news agencies
- Group 3
    - Direct quotes from government ministers found in news articles
    - Twitter posts from federal or territorial government websites or MOH
    - Facebook posts from federal or territorial government websites or MOH
    - You Tube posts from federal or territorial government websites or MOH
    - Instagram posts from federal or territorial government websites or MOH
- Group 4
    - General social media

The measure_stage field indicates whether the measure is ‘new’ or if a change has been made to an existing measure. Changes are either an ‘extension’ when the time has been extended or a ‘modification’ where the measure has changed (or been both changed and extended). Where a measure is being reduced, the value was set to ‘phase-out’, while measures that ended were set to ‘finish’. Where a measure is changed over time which resulting in multiple rows in this way, the fields prev_measure_number and following_measure_number were populated to indicate the

sequence of the changes.  Date_end was populated by a coding routine derived from the measure_stage and date_start fields.

Once the volunteer has manually gone through the whole file, it was checked by a member of the coding admin team.  This team consisted of people that routinely checked files, and fully understood the taxonomy and how records were linked. Feedback was provided to volunteers so they could learn from the corrections. These checked files were returned to LSHTM where it is combined together into one large file.

This combined file is run through R to population the date_end, and reason for ending fields and to remove unwanted columns to produce the dataset for release. Finally, the clean, verified data is ready to be shared with WHO and other researchers.

Tables 3: Data dictionary for the Stage 1 WHO global PHSM dataset

| Variable name | Description | Data type | Notes / coding |
|---|---|---|---|
| Number | Unique record number | | |
| who_id | Record number in original dataset | Text | |
| Dataset | Original dataset | Coded | ACAPS, CDC_ITF, GPHIN, IHR, OXGCRT, CSH_Vienna, JHU |
| prop_id | | String | |
| who_region | WHO regional designation | Coded | AFRO, AMRO, EMRO, EURO, SEARO, WPRO |
| country_territory_area | Country or territory name | String | WHO Member States list and, residually, UN list provided by UNTERM |
| Country_code | Three-letter country or area code | Coded | https://www.iso.org/obp/ui/#search |
| Country_code_numeric | Three-digit country or area code | Coded | https://www.iso.org/obp/ui/#search |
| admin_level | The level measure applies to | Coded | national, other |
| area_covered | Name of area covered where not who country / territory | String | |
| prov_category | The category provided by provider | String | |
| prov_subcategory | The sub-category provided by the provider | String | |
| prov_measure | The measure provided by the provider | String | |
| who_code | PHSM code given in the WHO taxonomy | Coded | Link to taxonomy |
| who_category | PHSM category given in the WHO taxonomy | Coded | Link to taxonomy |
| who_subcategory | PHSM subcategory given in the WHO taxonomy | Coded | Link to taxonomy |
| who_measure | PHSM measure given in the WHO taxosecondmy | Coded | Link to taxonomy |
| targeted | Extent of targeting of the measure | Coded | General, no, targeted, yes |
| value_usd | The value of financial stimulus | Number | Only currently used by Oxford |
| percent_interest | The change interest rates | Number | Only currently used by Oxford |
| comments | The notes or comments filled out by the provider | String | For some groups this also contains links. |
| non_compliance | Consequences for non-compliance with the measure | String | |
| response_type | Indicating whether measure is being implemented or lifted | Coded | Impose, Introduction / extension of measures, Lift, Phase-out measure |
| source | Entity providing information on the measure | String | |

| | | | |
|---|---|---|---|
| source_type | Classification of the entity providing information on the measure | Coded | Government, Media, Official, Other, Other organisations, Social media, UN |
| link | Link to the source of the information on the measure | String | |
| link_alt | An alternative link to the measure | String | |
| date_start | Date for the start of the measure | Date | |
| date_end | Date for the end of the measure | Date | |
| date_entry | Date of entry of this measure into the original dataset | Date | |

Tables 4: Data dictionary for the Stage 2 WHO global PHSM dataset

| Variable name | Description | Data type | Notes / coding |
|---|---|---|---|
| who_id | Record number in original dataset | Text | |
| Dataset | Original dataset | Coded | ACAPS, CDC_ITF, GPHIN, IHR, OXGCRT, CSH_Vienna, JHU |
| prop_id | | String | |
| who_region | WHO regional designation | Coded | AFRO, AMRO, EMRO, EURO, SEARO, WPRO |
| country_territory_area | Country or territory name | String | WHO Member States list and, residually, UN list provided by UNTERM |
| Country_code | Three-letter country or area code | Coded | https://www.iso.org/obp/ui/#search |
| Country_code_numeric | Three-digit country or area code | Coded | https://www.iso.org/obp/ui/#search |
| admin_level | The level measure applies to | Coded | national, other |
| area_covered | Name of area covered where not who country / territory | String | |
| who_code | PHSM code given in the WHO taxonomy | Coded | Link to taxonomy |
| who_category | PHSM category given in the WHO taxonomy | Coded | Link to taxonomy |
| who_subcategory | PHSM subcategory given in the WHO taxonomy | Coded | Link to taxonomy |
| who_measure | PHSM measure given in the WHO taxosecondmy | Coded | Link to taxonomy |
| comments | The notes or comments filled out by the provider | String | For some groups this also contains links. |
| date_start | Date for the start of the measure | Date | |
| Measure_stage | Indicating whether measure is being implemented or lifted | Coded | Impose, Introduction / extension of measures, Lift, Phase-out measure |
| Prev_measure | The who_id of the previous measure, if there is one. | String | |
| Following_measure | The who_id of the follow on measure, if there is one | String | |
| date_end | Date for the end of the measure | Date | |
| Reason_end | The reason the measure ended | Coded | New, Modification, extensions, phase-out, finish, time limited |
| targeted | Extent of targeting of the measure | Coded | General, no, targeted, yes |
| enforcement | If the measure is enforced or not | Coded | Recommended, required, monitored, not applicable, not known |
| non_compliance | Consequences for non-compliance with the measure | Coded | |

| | | | |
|---|---|---|---|
| link | Link to the source of the information on the measure | String | |
| Link_live | Does the link still show the same information as when the provided coded it | Coded | Yes, No, unknown |
| Link_eng | Is the link in English | Coded | Yes, No, unknown |
| source | Entity providing information on the measure | String | |
| source_type | Classification of the entity providing information on the measure | Coded | Government, Media, Official, Other, Other organisations, Social media, UN |
| Alt_link | An alternative link to the measure | String | |
| Alt_link_live | Does the link still show the same information as when the provided coded it | Coded | Yes, No, unknown |
| Alt_link_eng | Is the link in English | Coded | Yes, No, unknown |
| date_entry | Date of entry of this measure into the original dataset | Date | |

## Appendix

### Description of source datasets

**WHO International Health Regulations**

The International Health Regulations, or IHR (2005), represent an agreement between 196 countries including all WHO Member States to work together for global health security.

Through IHR, countries have agreed to build their capacities to detect, assess and report public health events. WHO plays the coordinating role in IHR and, together with its partners, helps countries to build capacity.

IHR also includes specific measures at ports, airports and ground crossings to limit the spread of health risks to neighboring countries, and to prevent unnecessary travel and trade restrictions so that international traffic disruption is kept to a minimum.

Under IHR Article 43, countries implementing additional health measures that significantly interfere with international traffic are required to provide to WHO the public health rationale and relevant scientific information for the measure. These data are collected in a standardized format.

**ACAPS**

In the ACAPS Government Measures dataset, each action taken by governments in response to COVID-19 falling into the taxonomy corresponds to a record, with changes to that measure noted through additional records. Data collectors were trained on the taxonomy of PHSM and in the dataset structure. The coverage of the dataset is intended to be global, with data available for 193 countries. The information comes from a variety of publicly available sources obtained through the internet. Analysts navigate the web looking for information on governments measures, utilising sources from: governments (official sites, embassies), media, United Nations agencies and other organizations. Priority is given to official/governmental sources. Some measures may not have been recorded and the exact date of implementation may not be accurate in some cases due to the different methods of reporting used by the primary data sources consulted.

Contact: info@acaps.org

Link to dataset: https://www.acaps.org/covid19-government-measures-dataset

**University of Oxford**

The Oxford COVID-19 Government Response Tracker (OxCGRT) collects real-time information on governmental responses to COVID-19 across 13+ indicators. These indicators include: school closings, travel restrictions, testing regimes, economic measures, and other types of responses. Government actions are scored on an ordinal scale to create comparable cross-national measures, while contextual notes and source materials are also recorded to create a rich qualitative record. Our objective is to be accurate, up-to-date, and global. OxCGRT collect data via a team of 100+ Oxford students, staff, and alumni from around the world. These trained volunteers collect data from government sources, media, and other publicly available materials. Data collectors work continuously in 3-day cycles to ensure data are updated, expanded, and checked for accuracy. We

make all of our data instantly accessible downloads, visualization, and an API feed on our project website:

Link to dataset: https://www.bsg.ox.ac.uk/research/research-projects/oxford-covid-19-government-response-tracker

**Global Public Health Intelligence Network**

The Public Health Agency of Canada's Global Public Health Intelligence Network (GPHIN) is an early-warning and situational awareness system for potential chemical, biological, radiological and nuclear public health threats worldwide—including outbreaks of infectious disease. Formed in the late 1990s by the Government of Canada (Health Canada) in collaboration with the WHO, GPHIN is headquartered at the Public Health Agency of Canada in Ottawa, Ontario, Canada. GPHIN users include non-governmental agencies and organizations, as well as government authorities who conduct public health surveillance. GPHIN is a significant contributor to the World Health Organization's Epidemic Intelligence from Open Sources.

For each row in the original file, there is either a GPHIN ID for an article in our system or a link to the source website at the comment column.

Contact: phac.gphin-rmisp.aspc@canada.ca

For more information: https://gphin.canada.ca/cepr/aboutgphin-rmispenbref.jsp?language=en_CA

**US Centers for Disease Control and Prevention (CDC)**

The US Centers for Disease Control and Prevention (CDC) is working closely with partners worldwide to respond to the coronavirus (COVID-19) pandemic. CDC is aggressively responding to COVID-19 through the development of pandemic preparedness and response plans and working on multiple fronts to prepare communities to respond to this public health threat. The CDC COVID-19 International Task Force (ITF) aims to limit human-to-human transmission and minimize the impact of COVID-19 in vulnerable countries with limited preparedness capacity. To this end, the ITF works to (1) strengthen capacity to prevent, detect, investigate and respond to local COVID-19 transmission and provide useful information to public health authorities so they may better plan and implement appropriate control and intervention measures; (2) Mitigate COVID-19 transmission in the community, across borders, and in health care facilities; (3) Support governments, nongovernmental organizations, and healthcare facilities to rapidly identify, triage and diagnose potential cases to improve patient care and minimize disruptions to essential health services; (4) Address crucial unknowns regarding clinical severity, extent of transmission and infection with support for special investigations and other forms of cooperation between CDC and country partners; and (5) Ensure readiness to implement vaccines and therapeutics when available. The mitigation data were provided from the CDC COVID-19 Response ITF Mitigation Team. Data were abstracted from several available public sources including US embassy websites, government and UN websites, CDC country offices, and media reports. Efforts including mitigation measures were categorized into seven overarching categories plus sub-categories. The team has also worked on determining whether restriction categories are focused on preventing importation or containing the spread of COVID-19

within the country. Furthermore, initial implementation of measures, extension or lifting of existing measures was tracked temporally for all countries.

More information about CDC's Global COVID-19 Response can be found on the CDC website: https://www.cdc.gov/coronavirus/2019-ncov/global-covid-19/index.html

For questions about the database, please contact the CDC International Task Force inbox: eocevent223@cdc.gov

**Complexity Science Hub Vienna**

On 19th March 2020, Complexity Science Hub Vienna (CSH Vienna) set up a platform for students, researchers, and volunteers to collect data on the public health and social measures implemented by governments to prevent and limit the spread of COVID-19, including the time schedules for implementation. A wide range of different public sources are used to populate, update and curate the dataset, including official government sources, peer-reviewed and non-peer-reviewed scientific papers, webpages of public health institutions, press releases, newspaper articles, and social media. Strategies that could provide assistance to the population (e.g., related to financial support or food supply) or that may encourage compliance with the measures (e.g. resource allocations, risk communication) were also included. The specific details and descriptions of each public health and social measure were subsequently standardized and coded.

Our project aims to assess the impact of these governmental actions on the spread of COVID-19.

Link to dataset: https://github.com/amel-github/covid19-interventionmeasures

Link to project: http://covid19-interventions.com/

For questions about the database, please see: Amélie Desvars-Larrive (desvars@csh.ac.at)

**Johns Hopkins University (JHU)**

The Health Intervention Tracking for COVID-19 (HIT-COVID) project tracks the implementation and relaxation of public health actions taken by governments to slow transmission of SARS-COV-2 globally. Hundreds of volunteer data contributors were trained, provided with standardized field definitions and access to an online forum for asking questions and sharing ideas. Each change in policy and corresponding date is documented at the first-level administrative unit (admin unit level 1, e.g., states, districts) and nationally for all countries with more detailed geographic resolution in some locations (e.g., counties in the US). Data are entered into a structured questionnaire with a source document(s) required for each record. Source documents from official government sources are prioritized, but other sources are permitted. For each intervention, HIT-COVID captures a suite of additional data including whether interventions are required or recommended and the particular subpopulation to which policies apply. To ensure data quality, contributors are asked to complete weekly self-audit reports, have the ability to submit corrections on past entries, and the management team performs geographic or intervention-specific audits as issues arise.

Contact: hit-covid@jhu.edu

Link to Dataset: https://github.com/HopkinsIDD/hit-covid

Visualizations: https://akuko.io/post/covid-intervention-tracking