Documentation

Demscore Handbook

User Support

June 2023

demscore

100+
DATASETS

250+
COUNTRY UNITS

500+
MERGE OPTIONS

20,000+
VARIABLES

PARTNERING RESEARCH INSTITUTES

Varieties of Democracy (V-Dem)
The world's largest data collection project on democracy.

Quality of Government (QoG)
Providing some of the largest and most comprehensive datasets on governance.

Uppsala Conflict Data Program/ Violence Early Warning System
The world's leading provider of data on armed conflicts and organized violence.

Comparative Policy Laboratory (COMPLAB)
COMPLAB brings together high quality data on social, environmental, and migration policies.

Representative Democracy
REPDEM combines leading datasets on national governments across Europe/OECD.

Historical Data Archive (H-DATA)
A hub for collecting, integrating and curating historical data for all other Demscore partners.

DEMSCORE (Democracy, Environment, Migration, Social policy, Conflict, and Representation) is building a world-leading hub for contextual data with user-friendly access to documentation, and online visualization tools. The data, provided by some of the world's leading research infrastructures, is open access and free to use worldwide.

DEMSCORE facilitates large-scale comparative analyses on the grand challenges of today's societies, including those caused by population aging, rapidly changing migration patterns, increased social inequalities, accelerating globalization, recurrent financial crises, political deadlocks, violent conflict, and the rise of populism. The interdisciplinary nature of DEMSCORE data is essential to advance adequate policy responses to such complex, and interrelated societal challenges facing the world today.

Users are offered harmonized social science data of the highest quality merged using an innovative and systematic data harmonization methodology. It maximizes usability in the measurement of contextual data with over 20,000+ variables across nearly all countries in the world, from 1789 to the present.

A fully normalized, joint PostgreSQL database, sophisticated programming, and a web-based interface make it possible to select a series of variables from all six partnering data-infrastructures, and get a custom-designed dataset and codebook generated automatically in a matter of seconds. This makes DEMSCORE an enormously time efficient resource compared to merging several datasets by hand.

To find out more, and download your own dataset, visit:
http://www.demscore.se

DEMSCORE IS FUNDED BY

Swedish Research Council

UNIVERSITY OF GOTHENBURG

UPPSALA UNIVERSITET

Stockholms universitet

UMEÅ UNIVERSITY

HOW TO USE THE DOWNLOAD INTERFACE

Demscore data can be retrieved from the download interface either by variable, by codebook section, by downloader ID, or by Output Unit. We recommend users to thoroughly read the Explanatory Notes section of the Methodology document before downloading data through Demscore.

1. Select how to retrieve data

Users interested in specific variables can download data by variable, and users interested in variables related to a specific topic can download data by codebook section. All variables from all datasets are thematically grouped into codebook sections in a PostgreSQL database. This organization enables the user to easily select all variables across all modules related to their topic of interest available in their Output Unit of interest.

If the purpose of the download is to replicate data, users can download data by downloader ID. Each download through the DEMSCORE web interface is assigned a unique downloader ID, allowing exact replicability when shared with other users.

If you are new to DEMSCORE data, and need to learn more about available datasets and Output Units in order to find the right format and sources for your individual dataset, we advice you to consult the list of Output Units here:
https://www.demscore.se/output-unit-selection/.
The page includes information about Output Units available in Demscore, and links to the download interface with the selected Output Unit filled in advance.

With the current setup, a user can generate a customized dataset and accompanying codebook in a matter of seconds. For example, creating a customized dataset with a tailored codebook encompassing 20 variables from ten datasets takes approximately 25 seconds, which is enormously time efficient compared to merging ten datasets by hand.

2. Select main dataset of interest

This step is only applicable when using the option retrieve by variable.

Depending on the selected dataset of interest, you will be suggested the original Output Unit of the dataset, i.e. the unit in which you can keep variables from the chosen dataset in their original form. You can however set the Output Unit to any unit you prefer.

Dataset Unit

A Dataset Unit, e.g., Country-Year, describes the level at which observations for a dataset are collected. Observations are stored as rows in a table. In order to find a specific observation, e.g., information on a specific country for a given year, special table columns are needed as identifiers. Comparable to page numbers in a book, these columns help us find the location of the table row that contains the values for each variable of interest for a given observation i.e., a given country and year. The information necessary to identify these rows may be stored in a single or in a combination of several identifier columns. In the most common example for datasets that have the Country-Year Dataset Unit, the country and year information is each stored in a separate column.

Please note that you do not automatically download the whole dataset when selecting a main dataset of interest. You still need to select single variables (see step 6).

3. Select Output Unit

The selected Output Unit determines the identifiers on which your chosen variables are merged. The Output Unit is recommended based on the chosen main dataset of interest. This is however just a recommendation, it can be freely chosen by the user.

What is the suggestion based on?

If you select the QoG Standard TS dataset as the main dataset of interest, we assume that most of the variables you want to download are from that dataset. Hence, you are recommended to select the QoG Country-Year Output Unit as this is the original unit of this dataset and all its variables will be available in their original form when downloaded in this unit. Variables from other datasets are merged based on the country and year identifiers in the QoG Country-Year unit.

Which Output Unit should I choose?

To choose the right Output Unit, you first need to decide in which format you want to retrieve the data. Demscore offers several formats which includes, but are not limited to, the following:

- Country-Year
- Cabinet
- Country/Regional
- Conflict
- Other

Output Unit

An Output Unit, e.g., QoG Country-Year, is defined as an output format in which variables can be retrieved from one or more datasets through a strictly defined output grid. A unit table defining this output grid contains unit identifier columns with u_ prefixes and the table is sorted based on these unit identifier columns and has a fixed number of rows. An Output Unit has specific definitions for the level at which observations are presented, e.g., country definitions. For example, variables from a QoG dataset may have been collected under QoG country definitions, but in Demscore can also be retrieved through a V-Dem Output Unit which follows V-Dem country definitions.

Unit Columns

Unit columns are based on the columns that constitute a Dataset Unit. They are added to the original dataset and marked by a unit prefix (consisting of a u_ and the dataset unit name) before the original variable name. Unit columns can contain slightly modified data, e.g., NAs are replaced by a default value. Sometimes we add additional columns to the unit table, for instance if a dataset includes both a country_id column with a numeric country code, we add the variable storing the full country name to the unit table as well for better readability.

3. Include Unit Columns

We recommend including unit columns as they identify rows in the customized dataset you download. You can select additional identifier variables (e.g., country) from the original datasets you download variables from, however, these dataset-identifiers might not cover all rows in the Output Unit.

4. Exclude empty rows

You can choose to include or exclude the rows for which no non-missing observation from the chosen variables has a match in the selected Output Unit, i.e., rows that only have missing observations. Excluding empty rows might be a good option if you choose variables that only have very few observations in your chosen Output Unit. It however means that you might not be able to column-bind your dataset with another dataset downloaded in the same Output Unit at a later point, as columns might then be of different length. We therefore recommend to exclude empty rows manually at a later point.

5. Select variables

We display the variable label as well as the DEMSCORE internal long tag for each variable (in parenthesis). The first part of the tag in parenthesis indicate which dataset the variable comes from, e.g., from qog_ei_ccci_em you can derive that the variable originated from the dataset that has the tag “qog_ei”, which corresponds to the QoG Environmental Indicators Dataset.

DEMSCORE METHODOLOGY

The following section includes a short summary of the DEMSCORE Methodology, including a description of the general approach, and dataset- and Output Unit combinations. To read and download the full version of the DEMSCORE Methodology, please visit www.demscore.se

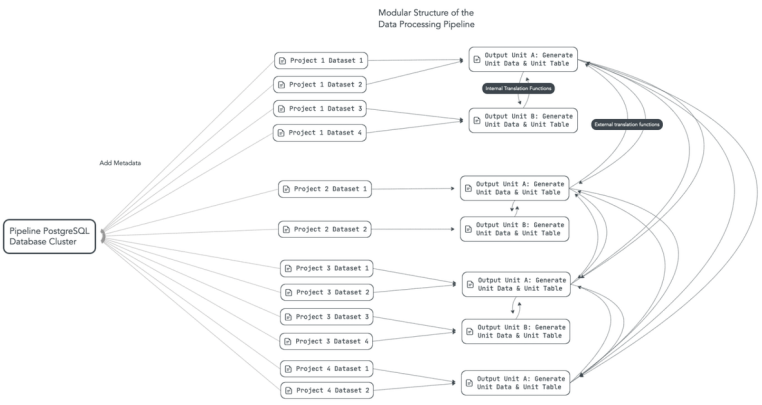
The potential that lies in the interdisciplinarity of DEMSCORE data is accompanied by the challenge of its diversity. On this page, we outline a summary of the methodological considerations, choices, and procedures guiding the development of the DEMSCORE project, stretching across the construction of a common e-infrastructure to fully harmonized data from all sources.

We continually review our methodology—and occasionally adjust it—with the goal of improving the quality of DEMSCORE indicators and indices. We report the challenges of the harmonization process and present the solutions provided by the DEMSCORE Team and its member projects.

In order to present data merged from different sources in a consistent way, we have developed an innovative and systematic data harmonization process. Each dataset in Demscore has its own dataset unit, i.e., unit of analysis. Based on their units of analysis, datasets are grouped into Output Units and finally merged and translated through these Output Units.

Several datasets within the same module have very similar units. That is, if more than one dataset within a module has the combination of columns containing e.g., the country and year as their identifying unit, we take the union of the country-year combinations from each dataset with similar units and define the result as an Output Unit. This is illustrated in figure 1, which also demonstrates that we use these Output Units to translate and merge data from different units and modules together, i.e., we translate between Output Units both within and across projects.

Figure 1



While Dataset A of Project X for example has observations for European countries from 1944 to 2015, Dataset B for Project X has observations for all countries from 1990 to 2021. The Country-Year unit table for Project X will have rows for European countries from 1944 to 1989 as well as rows for all countries from 1990 to 2021. In other words, a unit table always covers all possible combinations of unit values in the identifier columns of datasets with the same dataset unit. This is illustrated using the following example based on a subset of the data above: Member Project X has two datasets available; Dataset A and Dataset B. Dataset A is a table with three columns, called country, year and var_1 with corresponding values in five rows. Dataset A is a table with three columns, called country, year and var_1 with corresponding values in five rows. Dataset A has the Unit Country-Year, meaning each row in the dataset is identified by a unique combination of countries and years.

Table 1: Project X: Dataset A

country	year	var_1
ITA	1988	1
ITA	1989	2
ITA	1990	3
DEU	2013	4
DEU	2014	5
DEU	2015	6

Dataset B is a table with three columns, country_id, year and var_2, with corresponding values in eight rows. It also has Country-Year as its unit, meaning each row in the dataset is identified by a unique combination of countries and years. Please note that Dataset B contains some country and year combinations that can be found in Dataset A as well, but also country and year combinations that are not included in Dataset A.

Table 2: Project X: Dataset B

Country_id	Year	var_2
ITA	1990	a
ITA	1991	b
ITA	1992	c
DEU	2015	d
DEU	2016	e
DEU	2017	f
NPL	2005	g
NPL	2006	h
NPL	2007	i

We first identify a unit grid, i.e., unit table for datasets with common dataset units, in this case the Country-Year unit for Dataset A and Dataset B.

The Country-Year Unit Table for Project X contains all Country-Year combinations from Dataset A as well as all Country-Year combinations from Dataset B without duplicates.

If a user now selects the Output Unit Project X Country-Year Unit, and also selects the variables var_1 from Dataset A and var_2 from Dataset B, or even both at the same time to be expressed in this Output Unit, they receive a new dataset with twelve rows sorted the same as the unit table and containing the selected variables.

Table 3: Unit Table for Project X Country-Year Unit

u_x_cy_country	u_x_cy_year
ITA	1988
ITA	1989
ITA	1990
ITA	1991
ITA	1992
DEU	2013
DEU	2014
DEU	2015
DEU	2016
DEU	2017
NPL	2005
NPL	2006
NPL	2007

As Dataset A does not include all country-years included in Dataset B and vice versa, but the unit table encompasses all combinations from both datasets, var_1 and var_2 have missing values (NA = Not Available) for those Country-Year combinations not included in their original dataset when retrieved through the Country-Year Output Unit of Project X.

Table 4: New Dataset from Selection of Variables from Datasets A and B

u_x_cy_country	u_x_cy_year	var_1	var_2
ITA	1988	1	-11111
ITA	1989	2	-11111
ITA	1990	3	a
ITA	1991	-11111	b
ITA	1992	-11111	c
DEU	2013	4	-11111
DEU	2014	5	-11111
DEU	2015	6	d
DEU	2016	-11111	e
DEU	2017	-11111	f
NPL	2016	-11111	g
NPL	2017	-11111	h
NPL	2018	-11111	i

Note: var_1 and var_2 combined in the Project X Country-Year Unit

At this point we assume that every dataset has a Primary Unit and an attached unit table. Note that many Output Units may have only been constructed using a single dataset. Also note that Output Units are constructed across datasets within the same member project, but never across datasets from different member projects. The Country-Year unit table for Project X will thus not include Country-Year observations from datasets in Project Y if these Country-Years only exist in project Y.

Datasets that belong to the Country-Year Unit of Project X can now be translated to Other Units, e.g., the hypothetical Country-Year Unit of Project Y. Through aggregation and disaggregation, datasets that initially collect information on a Country-Year level also become available in other Output Units such as Cabinet-Date or Agency-Year.

MERGE INFORMATION

For more detailed information, we refer users to the explanatory notes of the Methodology Document, available on demscore.se

Merge Scores

We provide three merge scores in Demscore:

1. The number of non-missing observations in the original Output Unit of the variable.
2. The number of non-missing observations in the chosen end Output Unit for that variable.
3. In direct translations, the number of lost observations between the variable in its original unit and the variable in the chosen end Output Unit.

Here are a few general guidelines on how to read and interpret the merge scores offered in Demscore:

- If the score for a variable is very high in the original Output Unit but very low in the end Output Unit, and at the same time the score for lost observations is very high, this means that overlap in identifier combinations between these two Output Units is low.
- If the merge score is high in the original Output unit, but low in the end Output Unit and at the same time, the number of lost observations is low, this means that you have probably chosen a variable that is available only in very few identifier combinations, compared to the identifier combinations in the end Output Unit. However, the end Output Unit covers a lot of the observations from the original Output Unit of the variable.

The merge scores in Demscore v1 can thus provide the user with a hint of what to expect from the downloaded data. However, we recommend users to also investigate their customised dataset and which observations that matched, before deciding on whether or not to use the chosen variable for their analysis.

Information on which identifier combinations "get lost" during a translation from one Output Unit to another Output unit, will soon be listed below.

Please note that this information is currently only available for selected combinations of Output Units. We add files continuously and aim to include the lost observations for each variable to the chosen end Output Unit in the download file in the future.

Differences in Country Identifiers

A commonly used identifier across datasets in DEMSCORE modules are countries, but country definitions and names often vary. This poses a challenge for merging data. As a general rule, resolving merge conflicts depends on the chosen Output Unit. Merge scripts always prioritize preserving the data quality of the units of that Output Unit. This means that we follow the country definitions of the dataset chosen as Output Unit. When merging from other datasets we only keep combinations that match the chosen Output Unit.

While this resolves most merge issues, some conflicts remain due to differences in what territories are included in e.g. "France" or "Russia" over time.

The result of this process is a set of versions of indicators of democratic institutions and concepts, which allow academics and policymakers alike to understand the different features of a polity.

We list the most notable differences in country definitions that users should pay attention to when using data merged on country identifiers in the Methodology Document and on demscore.se.

For differences in the full country names across data sources and how they are adjusted for merges and translations between Output Units in DEMSCORE, see the original DEMSCORE translation functions in our code which is publicly available on: https://github.com/demscore/

We want to emphasize that DEMSCORE expresses no opinions on sovereign claims to disputed territories neither through the descriptions in the Methodology Document nor through any merge decisions. Nor does DEMSCORE make any judgments concerning which territories qualify as countries.

MERGE INFORMATION

Missing Data

DEMSCORE indicates different types of missingness for observations in the customized datasets:

Missing in original data = Whenever an observation in the original variable is a missing (NA, missing code such as 7777, blank cell), we preserve this missing value. When the original source has special codes for various types of missing, those are preserved.

Missing code: -11111 = DEMSCORE code for observation is missing due to the translation/merge, i.e., missing data due to no data being included for this combination of identifiers in the end Output Unit.

Missing code: -22222 = No observation is merged/translated, but the original data contains information for these identifier combinations elsewhere. For these cases, we use a different code. The user needs to consult the reference documents (Methodology Document Section 5.1. or the Demscore Handbook) to clarify why the translation to the identifier combinations in the end Output Unit was not possible.

Citations

The DEMSCORE project does not have a formal citation of its own. Hence, when using DEMSCORE, we encourage that you cite the respective projects and datasets.

We indicate how every dataset is to be cited in the autogenerated codebook you receive with your data download, both in the dataset description and the codebook entry for each variable.

Most often it is sufficient to cite the dataset a variable originates from, but sometimes there is a variable specific citation listed in the codebook entry in addition. For these cases, please also add the variable specific citation to the reference list of your publication. Full references are linked in the codebook entries of the variables and listed in the codebook's bibliography.

We also encourage you to cite the DEMSCORE Methodology Document when using data retrieved through DEMSCORE.

Download ID

The download ID allows the user to share the ID with other users for replication purposes.

A user can type the download ID into the download interface on demscore.se and retrieve the same download selection and files as the original user.

A unique download ID is autogenerated for each download from the DEMSCORE website and will always retrieve the same data, even if the DEMSCORE version was updated in the meantime.

You can find the downloader ID in your autogenerated customized codebook which is included in the zip file retrieved from the download interface.

Unit Identifier Variables

An Output Unit is defined as an output format in which variables can be retrieved from one or more datasets through a strictly defined output grid. A unit table defining this output grid contains unit identifier columns with u_ prefixes and the table is sorted based on these unit identifier columns and has a fixed number of rows.

Unit columns are based on the columns that constitute a the unit of analysis in a dataset. They are added to the original dataset and marked by a unit prefix (consisting of a u_ and the dataset unit name) before the original variable name.

Unit columns can contain slightly modified data, e.g., missing values are replaced by a default value. Sometimes we add additional columns to the unit table, for instance if a dataset includes both a country_id column with a numeric country code, we add the variable storing the full country name to the unit table as well for better readability.

DEMSCORE IN NUMBERS

Number of Datasets **101**

23,650 **Number of Variables**

Number of Output units **43**

309,537 **Number of variables accessible through Output Units**

Number of non missing observations accessible through Output Units **9,199,891,315**

44,072 **Number of lines of R code for Infrastructure Construction**

Number of available dataset to Output Unit combinations **798**

10,559 **Demutils R-package utilities for Infrastructure Construction**

Number of translation functions **118**

demscore