demscore

Working Paper Series

DEMSCORE:

One Infrastructure to Merge Them All

Melina Liethmann, University of Gothenburg
Lisa Gastaldi, University of Gothenburg
Staffan I. Lindberg, University of Gothenburg
Steven Wilson, Brandeis University

June 2025

Working Paper
No. 1: 2025

Demscore is a collaborative infrastructure of many of the world's most prominent social science research institutes, including V-Dem, UCDP/VIEWS, QoG, COMPLAB, REPDEM, and H-DATA. Demscore solves the persistent problem of inconsistent and incompatible datasets in the social sciences with an innovative strategy for open access and real-time customized data merging between all member datasets. As a collaborative effort between leading Swedish universities, Demscore elevates the scale of social science data to a new level, offering unprecedented opportunities for interdisciplinary research and knowledge advancement.

Please address comments and/or queries to:

Demscore
Department of Political Science
University of Gothenburg
Sprängkullsgatan 19, Box 711
405 30 Gothenburg
Sweden
E-mail: contact@demscore.se

Demscore Working Papers are available in electronic format at https://www.demscore.se

DEMSCORE: One Infrastructure to Merge Them All *

Melina Liethmann
University of Gothenburg

Lisa Gastaldi
University of Gothenburg

Staffan I. Lindberg
University of Gothenburg

Steven Wilson
Brandeis University

*Acknowledgments: This paper builds on the collective efforts of the team behind the national e-infrastructure Demscore. It would not have been possible without the valuable contributions of Josefine Pernes (University of Gothenburg), Johannes von Römer (Former Demscore Data Manager and Research Software Engineer), Torbjörn Bergman (Umeå University), Andreas Duit (Stockholm University), Johan Hellström (Umeå University), Håvard Hegre (Uppsala University), Kenneth Nelson (Stockholm University), Magnus Öberg (Uppsala University), Joakim Palme (Uppsala University), Aksel Sundström (University of Gothenburg), and Jan Teorell (Stockholm University). We are also grateful for the joint efforts of data managers, program managers, and research coordinators at V-Dem, QoG, UCDP/VIEWS, COMPLAB, REPDEM, and H-DATA.

DEMSCORE: One Infrastructure to Merge Them All

Introduction

The potential of the ever-growing availability of open source data is accompanied by the challenge of its diversity: interdisciplinary data requires data linkage through harmonization, integration, and merging. The DEMSCORE[1] research infrastructure has been engineered to enable universal linkage of varying types of data from diverse sources, collected for different kinds of units, stored in various formats, and conceptualized in multiple ways. Traditionally, this process is notorious for being, at best, time-consuming and complex and, at worst, prone to errors that can lead to flawed results. We solve this challenge by introducing a newly constructed infrastructure built on a generalized framework for conceptualizing the data linkage problem, presenting a universal and scalable solution, i.e., one that can be relatively effortlessly extended to additional datasets of almost any type. Providing linkage across 161 datasets after its inception phase, DEMSCORE already facilitates new types of rigorous analyses and thereby ultimately help provide more informed decision-making (Kiszkiel et al. 2024; Cheng et al. 2024). One of the key engineering innovations of DEMSCORE is that, by design, only one dataset needs to be linked to one of the existing ones manually, and all other possible linkages are then generated automatically.

DEMSCORE is a free and public research infrastructure resource that thus offers harmonized, quality-controlled data. It enhances replicability, supports new research, and simplifies data merging across different sources. Designed to expand beyond the social sciences, DEMSCORE supports a more inclusive research environment, lowering barriers for users and promoting diverse perspectives in data-driven research.

The Data Linkage Triangle

The amount of available data is growing rapidly, yet researchers often only need a small subset of variables for their analyses. This calls for a solution that on the one hand handles large amounts of available data, and on the other hand gives scholars *instant access* to download whatever combination of variables, across datasets and varying units such as countries, parties, individuals, and text or speech, that are required for each specific project. The most persistent problems researchers face when combining data from different sources fall into the three key categories of linkage: harmonization, integration, and merging.

First, in addition to the challenges of harmonization that arise from concepts such as equality, religiosity, and others not being identically conceptualized and operationalized across time and datasets, definitions of units of analysis, such as geographic borders, political parties, and so on, also vary significantly. For example, country-year datasets vary on whether the pre-1991 U.S.S.R. and the post-1991 Russian Federation are the same country, requiring careful *harmonization*. Even when the units are the same, variations in identifiers can create significant challenges. Inconsistencies across data sources, along with differences in naming conventions, coding schemes, and levels of detail, often lead to mismatches, complicating the process. As a result, data *merging* becomes more time-consuming and error-prone than it needs to be.

Second, social science data are often constructed from fundamentally different units of analysis, such as dyad-year, country-year, party-election-year, event-date, individual-survey-round, or speech-act. Utilizing such diverse units in the same research project requires meticulous *integration* of datasets.

Third, key variables are often scattered across a plethora of datasets. For example, variables on migration, violent conflict, and democracy are rarely collected in the same dataset, meaning that even when the previous two problems are solved, the process of *merging* remains complex and time-consuming even for experienced researchers.

Meeting all these three challenges of linkage at the same time in a generalized solution that is easily scalable at low cost is what DEMSCORE does.

Finally, a research infrastructure handling many millions of data points and thousands of variables must enable users to select their preferred set of variables, and download that unique combination without much waiting time. In addition, each uniquely generated output dataset should be replicable. This requires an advanced engineering solution, optimizing computational power to prepare massive amounts of data to be made available for users within seconds. DEMSCORE meets this challenge as well.

Pioneering Data Linkage

Given that data harmonization practices are still evolving, DEMSCORE stands as a significant innovation with the potential to set methodological standards across the social sciences and beyond. By addressing key challenges in harmonization, integration, and merging, and by providing instant access to download the data along with a unique identifier for replication, the infrastructure facilitates data linkage on all levels. It enables the creation of customized datasets and reference documents from 161 datasets and more than 25,000 variables, generated in seconds, without requiring users to handle common data linkage issues themselves.[2]

This paper pursues two main goals: 1) describe DEMSCORE's innovative contributions to deterministic data linkage, and 2) emphasize both the limitations and opportunities for future contributions of the DEMSCORE infrastructure to the field of data linkage.

Innovations in Data Harmonization and Integration

To understand how DEMSCORE approaches this challenge, we define the concept of data linkage as the intersection of harmonization, integration, and merge efforts.

DEMSCORE is a framework that solves distinct problems on all levels of data linkage. The following section outlines our understanding of these challenges and illustrates how DEMSCORE approaches and solves the issues related to data harmonization, integration, and merging.

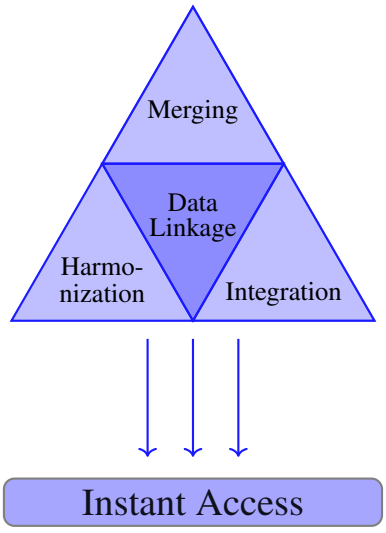

Figure 1: Data linkage illustrated as the intersection of harmonization, integration, and merging, processes required to enable instant access.

The Data Processing Pipeline

A data processing pipeline scripted in Bash, SQL, and R, facilitates seamless data linkage. While the DEMSCORE data processing pipeline can theoretically be run from just one line of code, the underlying structure is a complex interplay of a Relational PostgreSQL database (RDB), 40.000+ lines of R code, and three web servers. The core of DEMSCORE's

data linkage process is the programmatic generation of primary and secondary unit data.

Data Harmonization

DEMSCORE solves issues of ex-post, or retrospective harmonization, combining pre-existing datasets (Cheng et al. 2024; Kiszkiel et al. 2024; Wysmułek et al. 2022) all applying their own methodology (e.g., (Pemstein et al. 2024; Davies et al. 2023). This is different from ex-ante (or prospective) harmonization, i.e., the harmonization of research methodologies to collect data. DEMSCORE also enables flexible harmonization (different from stringent harmonization, i.e., the application of identical measures across different studies): Datasets, despite not being identical in how their observations are measured, are still inferentially equivalent and thus transformable into a common format Cheng et al. (2024).

Unlike common harmonization efforts, such as those outlined in de Sio et al. in this issue, and other theory- or data-driven frameworks, DEMSCORE is not oriented toward creating *the one* dataset for analysis. Instead, DEMSCORE enables flexibility and modularity. Harmonization in DEMSCORE is engineered to prepare a broad array of variables and datasets for selective harmonization according to each researcher's specific analytical goal and distinct research project. Our contribution to harmonization lies primarily in ensuring structural consistency across datasets: DEMSCORE's infrastructure aligns and translates identifiers, establishes exchangeable common output units, and standardizes key metadata elements to enable automated integration across data types. In this way, DEMSCORE supports both immediate data use and user-defined harmonization, transferring the ultimate decisions about scope and structure for each unique dataset creation out of a vast universe of possibilities to the researcher.

Combining Similar Identifiers to Create Output Units

DEMSCORE harmonizes and integrates variables using common units of analysis, hereby referred to as "Output Units", instead of merging datasets directly. This generalized solution then lays the foundation for unique, researcher-defined dataset creations.

In a first step, we derive and create Output Units from the original data sources by taking the union of unique identifiers (e.g., country-year, conflict-year, individual-survey-round) from datasets with similar units and define the result as an Output Unit type "X".

Project X: Creating an Output Unit from two Country-Year datasets

Dataset A: Country-Year Format

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

Dataset B: Country-Year Format

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

Unit Table: Unique Country-Year Combinations from Datasets A and B

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

Figure 2: Creating an Output Unit of the type Country-Year from two datasets

Figure 2 illustrates that identifiers determining an Output Unit are stored in a sorted table with a fixed number of rows, equivalent to the number of distinct identifier combinations within the union of included datasets. Output Units are a vital step for flexible data harmonization in DEMSCORE, as they account for the fact that despite data being collected and measured differently, they are still structurally similar and therefore linkable.

Output Units Resolve Merge Conflicts

When a researcher eventually chooses a specific Output Unit as the desired unit of analysis, any other selected data with other Output Unit types will follow the selected unit's set of identifiers when merging. We want to illustrate this using deviating country definitions as an example. Country definitions are chosen for simplicity, as they are an easily relatable concept, but the principle for handling mismatches is similar across all unit identifiers.

Choosing the V-Dem Country-Year Output Unit means that selected variables are merged

based on the country and year identifiers used in V-Dem. Variables from datasets that constitute this Output Unit can be retrieved through DEMSCORE in their original form and without any information loss. Variables from other datasets and Output Units will only be merged to country and year combinations that exist in the V-Dem Country-Year Output Unit. This resolves several merge issues:

One of the most sensitive cases are merges between the country units for Israel and Palestine. V-Dem includes Palestine/British Mandate, Palestine/Gaza, and Palestine/West Bank as separate country units. UCDP/VIEWS does not include a distinct country unit for Palestine, as the territory does not fall under the country definition used in their data collection. Instead, data on events/conflicts happening on Palestinian territories are collected under the location variable for the country unit Israel. As a result, this case does not allow the merging of country identifiers from V-Dem and UCDP/VIEWS due to the mismatch in country definitions. If a user would choose to download a UCDP variable together with several V-Dem variables in the V-Dem Country-Year Output Unit, the UCDP variable would have missing values for the three Palestinian territories included in the V-Dem Output Unit. This may lead to the false conclusion that UCDP does not collect data on Palestine, when, in fact, events in Palestinian territories are recorded under the country unit for Israel. Whilst the definition of Israel as a country is too different to allow for merging the data, it is important to emphasize that data is, in fact, collected for the UCDP variable but could not be matched. As a result, observations for Palestine in the UCDP variable are marked in the merged datasets as ”missing from mismatch” (code: -22222).

A similar example is that UCDP defines Eritrea as its own country only from 1993 onward (Gleditsch and Ward 2017) while V-Dem codes Eritrea as a separate unit even during periods of rule by Italy and Ethiopia (Coppedge et al. 2024). Cases for Eritrea pre-1993 are marked in the merged datasets as ”missing from mismatch”.

First Step: Generating Primary Unit Data as Basic Building Blocks

The first basic building block is then generated: we call it generating *primary unit data*. Primary unit data is the result each variable matched with its Output Unit identifiers, stored in one separate file per variable. Hence, single variables are extracted from their original dataset structure while preserving the data's original form, and functions as the basis for the next step: data integration.

Data Integration

Data integration refers to the process of combining data from different sources into a new, unique dataset (Hansen and Delgadillo 2024; Reis and Housley 2022). The result of successful data integration is a multidimensional dataset programmatically assembled from conceptually different datasets (Cheng et al. 2024). This can involve preparing datasets that collect data on different levels of analysis so they can be integrated into a common format, or shaping completely different types of data, such as text and survey data, into a common format. Currently, DEMSCORE's efforts are focused on the former.

Second Step: Generating Secondary Unit Data Allows Customization

The process of generating *secondary unit data* means matching variables from primary unit data to different sets of identifiers, i.e., making them available in different Output Units. For data downloads through DEMSCORE's web interface, users can customize a dataset from all variables available in a chosen Output Unit.

While generating secondary unit data can refer to simply matching variables collected at country-year level to a different set of country and year identifiers, the matching is typically much more complex and requires a deep understanding of the collected data, advanced skills in data processing, as well as the availability of technical hardware to transform large sets of data. Relatively straightforward examples of this is provided below.

Example 1: Integrating Data Through New Output Units

The following example illustrates how DEMSCORE integrates data from different levels of analysis. All transformation steps can be thought of as individual tasks in the data processing pipeline. The point of departure is two datasets: One collects data at the dyad-year level (Table 1(a)), another at the country-year level (Table 1(b)).

Table 1: Integrating country-year data into a dyad-year-location structure

(a) Dyad-Year Output Unit

dyad_id	year	location	var1
1	1996	A,B	a
2	1996	C,D	b
3	1996	B,C	c
4	1996	A,D	d

Note: A variable 'var1' available in its primary Output Unit Dyad-Year.

(b) Country-Year Output Unit

country	year	var2
A	1996	x
A	1997	y
B	1996	x
B	1997	x
C	1996	y
C	1997	z

Note: A variable 'var2' available in its primary Output Unit Country-Year.

(c) Dyad-Year-Location Output Unit

dyad_id	year	location
1	1996	A
1	1997	B
2	1996	C
2	1996	D
3	1996	B
3	1996	C
4	1996	A
4	1996	D

Note: Transformed Output structure with one row per per Dyad-Year-Location.

(d) Merged Output

dyad_id	year	location	var1	var2
1	1996	A	a	x
1	1997	B	a	z
2	1996	C	b	y
2	1996	D	b	-11111
3	1996	B	c	z
3	1996	C	c	y
4	1996	A	d	x
4	1996	D	d	-11111

Note: Final output dataset integrating both var 1 and var2.

Note: Each subfigure shows a key step in transforming country-year and dyad-year data into an integrated dyad-year-location structure.

No common set of identifiers exists to directly merge these two variables in their original Output Unit formats. However, we can include a location variable as an identifier in the Dyad-Year Output Unit, with comma-separated values similar to the country variable in the country-year Output Unit. We use this location variable in combination with the dyad- and year identifiers to transform the Output Unit identifier grid into a Dyad-Year-Location format, meaning one row per Dyad-Year-Location combination.

This transformation allows both *var1* of Table 1(a), and *var2* of Table 1(b) to be merged to a common set of identifiers (Table 1(c)).

Identifier combinations that did not match a value from either of the datasets, are indicated as ”missing from merge (-11111)” in the final dataset (Table 1(d)).

Example 2: Integrating Data Through Variable Transformation

Another way to integrate data from different levels of analysis, i.e., different Output Units, is to create new variables storing similar information in another format. This is implemented by taking a variable in its primary Output Unit format, and aggregating it to a different level of analysis before it is generated as secondary data in another Output Unit. This can include the means, sums or weighted averages of values in a group of identifiers. New variables are marked as such in the metadata database. The following example explains this process: The H-DATA Diplomatic Representation Dataset (Teorell 2022) collects information on which country *country_2* is represented by another country *country_1* at five different levels of diplomatic representation (0, 1, 2, 3, 9) in a given year, the original level of analysis being the dyad of *country_1* and *country_2* and year. The variable for the level of diplomatic representation is called *diprep_dr*. In its primary Output Unit, the *diprep_dr* variable takes the values 0, 1, 2, 3, or 9, each number standing for one level of diplomatic representation. A value ”2” for instance means representation at the Minister level, DEMSCORE's engineering aggregates the data to a Country-Year format by counting the countries that a *country_1* represents at each level of diplomatic representation in a given year. The new variable, i.e., the secondary unit data available in the Output Unit it is translated to, indicates how many countries are represented by a country 1 in a given year at the *diprep_dr* level 01239. For level 2, this new variable is named *count_diprep_level_2* and can be any positive integer or 0.

Merging

In the context of DEMSCORE, the data linkage triangle is completed by adding the component of data merging. Merging means bringing together information from two or more data sources following the objective of consolidating information on an object of interest (OECD 2006). This requires the development of a single taxonomy able to encompass disparate taxonomies or ontologies in datasets (Cheng et al. 2024). In DEMSCORE, the Output Unit type provides the single taxonomy. The generation of one file for each variable as primary unit data grouped by Output Unit types, and then making each variable available for each Output Unit type as secondary unit data, provides an ontology for merging any combination of variables regardless of original Output Unit group.

The major contributions of DEMSCORE in the field of merging are speed and amount. The Output Unit approach as well as automated data integration during the process of generating secondary unit data allow users combine large amounts of data collected on different topics in almost no time.

From Primary to Secondary Data: Catalyzing Merges

Aggregating, transforming, and reshaping variables from their primary Output Unit format to a secondary Output Unit format is an automated part of the data processing pipeline. The final puzzle piece to fully explain this process are variable merges. DEMSCORE merges data for both primary and secondary unit data generation, in the case of primary unit data generation, variables form datasets constituting an Output Unit are merged to their ”own” Output Unit table. In the case of generating secondary unit data, variables in their primary Output Unit format are merged to ”foreign” Output Unit tables.

Output Units allow us to treat larger groups of variables similarly within the pipeline compared to keeping the original dataset structure.

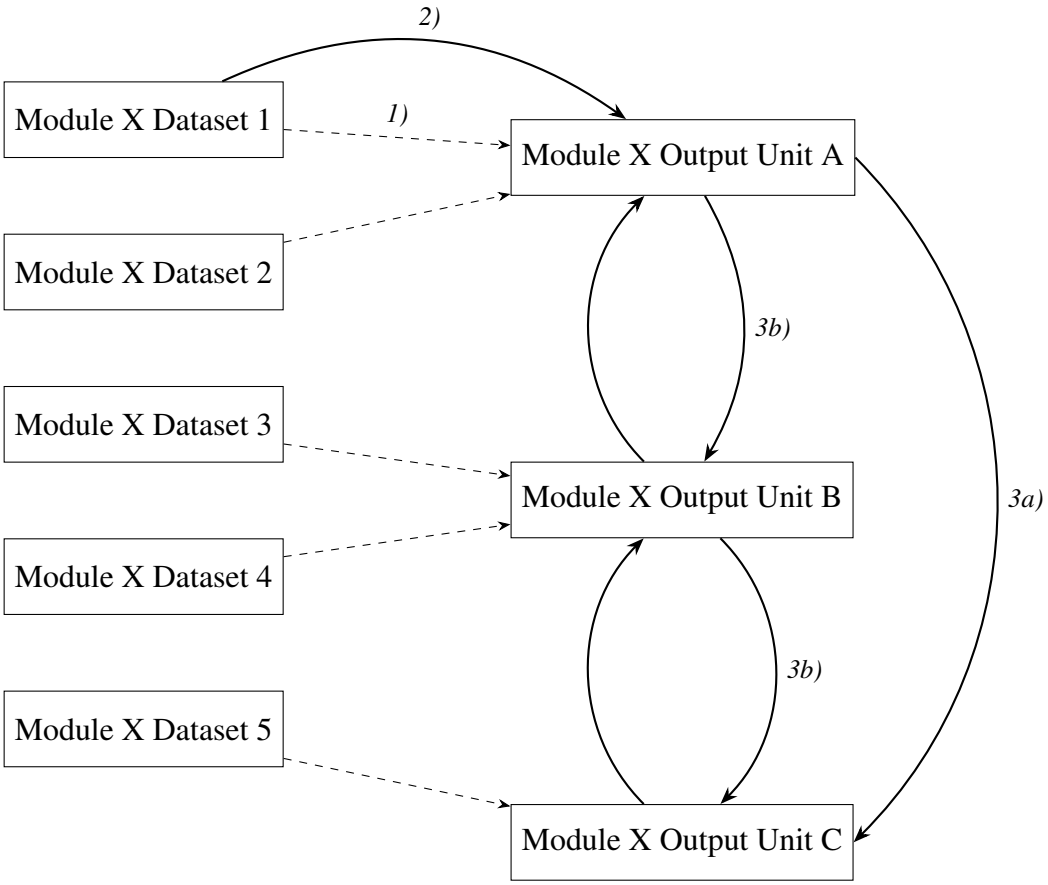

Figure 3: ***Data Harmonization, Integration and Merging in DEMSCORE*** *1) Identifier variables from datasets with common units of analysis are grouped into Output Units by storing the union of identifier variables in a unit table. 2) Primary unit data generation: Each variable is matched with the union of identifiers from its own Output Unit. 3) Secondary Unit data across units within a Module is generated using a) direct and b) indirect translations*

Figure 3 illustrates that "indirect translations" further reduce necessary translation functions. The arrow between *Module X Output Unit A* and *Module X Output Unit C* indicates that variables belonging to Output Unit A are directly merged to Output Unit C by a customized function between identifiers of the respective units. However, there is no direct function merging variables primarily available in *Module X Output Unit C* (in this case, all variables from *Module X Dataset 5*) to *Module X Output Unit A*. Yet, using indirect translation functions, i.e., translating variables originally available in *Output Unit C* first to *Output Unit B*, and then to *Output Unit B* to *Output Unit A*, the variables from Dataset 5 are still available in Output Unit A.

When adding a new dataset with a new type of Output Unit, the only requirement is to establish one new link to an already existing Output Unit. From there, the pipeline provides an indirect path specified in the RDB, and the necessary merge functions are automatically applied to get from *Output Unit C* to *Output Unit A* for variables from *Dataset 5*.

Theoretically, there is no limit to how many in-between Output Units a variable can pass through until it is translated to the desired format. However, each additional step comes at the cost of data quality loss: Indirect translation functions can only be used if the in-between Output Unit has similar identifiers to those in the desired end Output Unit. Otherwise, too many observations "get lost", as only matches with the in-between unit carry through. Hence, it is sometimes necessary to add a direct translation function between two units (such as the translation function between *Output Unit A and C*).

This process is currently applied to all datasets from all of DEMSCORE's Partner Modules, both within and across partner Modules. Figure 4 exemplifies the complexity of the infras-

tructure, once we begin to merge variables to Output Units from different Partner Modules. These ”external” translations follow the same logic of direct and indirect translations.

The relevant aggregation and merge functions, translation paths between primary and secondary units, as well as standardized descriptions are documented in a Relational Database (RDB) cluster, and automatically processed once DEMSCORE's data generation process is started.

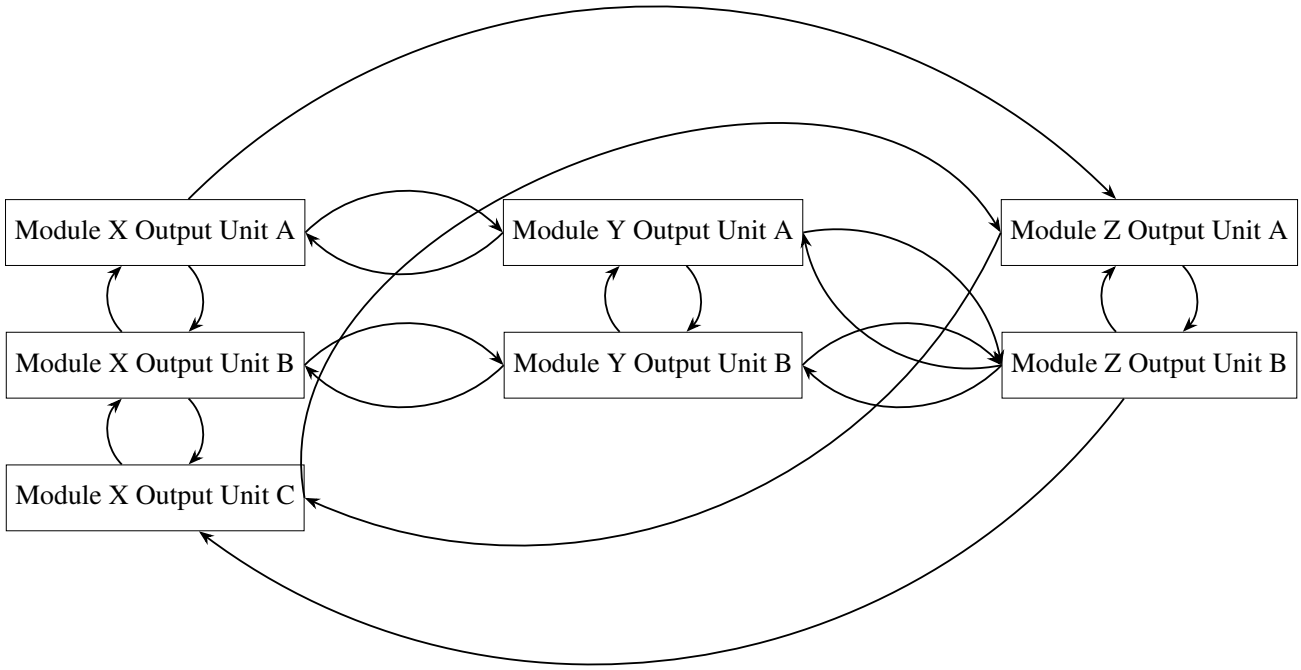

Figure 4: *Integrating data from all levels of analysis through merges between Output Units of Modules X, Y, and Z. The process illustrated in Figure 3 and Figure 4 is illustrated with a fictive example in Appendix A.*

The Online Construction Kit

All merging and data generation occur locally. Interconnecting with the RDB cluster, the data processing pipeline generates all data, which are transferred to the production PostgreSQL RDB cluster and finally uploaded to the joint web-portal as pre-processed files. This engineering solution enables users to choose an Output Unit in which they can select from all variables regardless of their original Output Unit. DEMSCORE's download interface automatically combines the chosen variables into a single data frame structure, and the unique dataset is immediately made available for analysis.

Figures 3 and 4 emphasize that despite its complexity, the DEMSCORE infrastructure is built in a way that is sophisticated yet straightforward, allowing for the creation of a large hub constructed from smaller dataset and Output Unit *bricks* connected through merge functions, and the possibility to extend this hub with as many new facets as desired in a time-efficient manner.

As of version 5, DEMSCORE encompasses 161 datasets, grouped into 53 Output Units. 159 translation functions and 131 additional indirect translation paths scale the original 25.299 variables up to 402.839 secondary unit variable files users can choose from to customize datasets.

Usability of Harmonized and Integrated Data

DEMSCORE is a generalized solution to time efficient merging and data integration, that can be applied to datasets across all scientific disciplines. Once we have successfully translated a new dataset to only one other Output Unit, it can relatively effortlessly be merged with variables from the other 161 datasets.

Harmonization is often defined as the process of aligning data directly into rows. In this sense, DEMSCORE's limitation lies in not automating the harmonization of similar variables (e.g., measures related to gender equality) into common indices. However, DEMSCORE does not (and does not aim to) modify or calculate new values from the original data, except in cases of transparent aggregations where necessary. Nor does DEMSCORE seek to develop new measures or variables of its own by employing common methods for data harmonization, such as Natural Language Processing, Machine Learning, etc., on, for instance, text data (Cheng et al. 2024). Ultimately, DEMSCORE does not yet offer solutions for probabilistic linkage, i.e., options for combining data when common unique identifiers are missing (Harron et al. 2016). Instead, the e-infrastructure contributes by laying the foundations for further harmonization (as described in de Sio et al. in this issue) and by automating data linkage through joining columns across different datasets. Ultimately, DEMSCORE focuses entirely on making existing data more accessible, customizable, and interoperable, thereby minimizing the burden of repetitive data wrangling for users.

Several important challenges do, of course, remain. There is a need to expand geographic and temporal coverage, refine metadata standards, and develop tools that automatically detect inconsistencies and provide clearer information and options regarding merge decisions, data loss through merging and integration, particularly for especially ambiguous or contested concepts.

To address some of these challenges, we are improving the handling of lost observations and developing more advanced linkage strategies using individual-level data. These updates are in progress and will be included in an upcoming release of the DEMSCORE infrastructure. There is also considerable potential to solve these issues by expanding ongoing and new collaborations with complementary initiatives to pursue shared harmonization goals and a more inclusive and methodologically rigorous data foundation for comparative researchers within the social sciences.

Nonetheless, DEMSCORE's innovations in data linkage represent a noteworthy contribution to methodological advancements in the social sciences. Users can access and merge data in real-time with unmatched flexibility. The platform offers tools to solve common merge issues, even across datasets not included in the infrastructure[3], enabling researchers to explore time-consuming correlations that may yield significant findings.

Users can filter data by countries, years, and other factors. Through a data processing pipeline connected to a relational PostgreSQL metadata database, each dataset includes a tailored codebook with details on selected variables, minimizing the need for consulting original references.

Every dataset customized in the download interface receives a unique download ID. By sharing this ID, supervisors and peer-reviewers can retrieve the exact same dataset via the DEMSCORE interface, reducing the need to submit data cleaning and preparation documentation.

Conclusion

DEMSCORE represents a significant advancement in data linkage methodology, offering a generalized and scalable solution to the time-consuming challenges of data harmonization, integration, and merging within the social sciences and beyond.

This contribution highlights the innovative processing of the included data and the significant improvement of the merge process. By constructing 'Output Units' from common identifiers, DEMSCORE reduces the number of required translation functions, facilitating the

seamless integration of variables across different datasets. In addition, DEMSCORE maintains high data quality through rigorous checks, defensive programming, and ongoing collaboration with expert researchers from its Partner Modules. The e-infrastructure incorporates versioning and reproducibility features, enabling users to share and recreate any dataset via a unique download ID. Merge reports provide transparency into the success of individual merges, detailing the effects of missing values and dropped observations. Each dataset download is also accompanied by comprehensive documentation, including detailed information on harmonization choices and methodological approaches (Gastaldi et al. 2025). Making this solution instantly accessible through our web-portal, DEMSCORE saves users significant time and effort in the process.

DEMSCORE is not designed as a static system but as an evolving platform. By remaining transparent about its accomplishments and limitations, we aim to enhance methodological clarity and empower researchers to pursue more ambitious and diverse questions across datasets and contexts. Offered as a free resource, DEMSCORE lowers the entry barrier for combining data from different sources and formats within seconds, supporting a more inclusive research environment and broadening the diversity of ideas in data-driven research.

Notes

[1]DEMSCORE is a collaboration between the University of Gothenburg, Stockholm, Uppsala, and Umeå University, jointly funded by the collaborating universities, and the Swedish Research Council (Grant number 2019-00187, 2021-00162, and 2023-00160). The data is provided by the partnering research institutes: V-Dem, QoG, UCDP/VIEWS, COMPLAB, REPDEM, and H-DATA. See `https://www.DEMSCORE.se/partners`

[2]Numbers based on DEMSCORE version 5.0, released in March 2025.

[3]DEMSCORE can be used to create country code translation tables for numeric and alphabetical country identifiers. See `https://www.DEMSCORE.se/news/using-DEMSCORE-to-combine-external-data/`

References

Cheng, C., L. Messerschmidt, I. Bravo, M. Waldbauer, R. Bhavikatti, C. Schenk, V. Grujic, T. Model, R. Kubinec, and J. Barceló (2024). A General Primer for Data Harmonization. *Scientific Data 11*(1), 1–41.

Coppedge, M., J. Gerring, C. H. Knutsen, S. I. Lindberg, J. Teorell, L. Gastaldi, A. Good God, and S. Grahn (2024). V-Dem Country Coding Units v14 (April 3, 2024) V-Dem Dataset. Available at SSRN: https://ssrn.com/abstract=4782741 or http://dx.doi.org/10.2139/ssrn.4782741.

Davies, S., T. Pettersson, and M. Öberg (2023). Organized Violence 1989-2022 and the Return of Conflicts Between States? *Journal of Peace Research 60*(4).

Gastaldi, L., M. Liethmann, J. von Römer, A. E. Owing, L. Morini, S. Sjögren, S. Wilson, and S. I. Lindberg (2025). Demscore Methodology v5. Demscore National Research Infrastructure.

Gleditsch, K. S. and M. D. Ward (2017). A revised list of independent states since the congress of Vienna. *International Interactions 25*(4), 393–413.

Hansen, J. and S. Delgadillo (2024). Publishing Survey Results. In D. Myers and J. Hansen (Eds.), *Inventories and Surveys for Heritage Management*. Getty Publications, Getty Conservation Institute.

Harron, K., H. Goldstein, and C. Dibben (2016). Introduction. In K. Harron, H. Goldstein, and C. Dibben (Eds.), *Methodological Developments in Data Linkage*, pp. 1–7. Wiley.

Kiszkiel, L., P. P. Laskowski, D. Voas, R. J. Bacon, W. J. Wildman, I. Puga-Gonzalez, F. L. Shults, and K. Talmont-Kaminski (2024). Dataset of Integrated Measures of Religion (DIM-R). Harmonization of Religiosity Data from Selected International Multiwave Surveys. *Religion, Brain & Behavior 0*(0), 1–41.

OECD (2006). Organisation for Economic Co-operation and Development. Glossary of statistical terms. accessed: 20 January 2025.

Pemstein, D., K. L. Marquardt, E. Tzelgov, Y.-T. Wang, J. Medzihorsky, J. Krusell, F. Miri, and J. von Römer (2024). The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data. *Varieties of Democracy Institute Working Paper 21*(9th Ed).

Reis, J. and M. Housley (2022). *Fundamentals of Data Engineering: Plan and Build Robust Data Systems*. O'Reilly: Beijing, Boston, Farnham, Sebastopol, Tokyo.

Teorell, J. (2022). Rules of recognition? Explaining Diplomatic Representation Since the Congress of Vienna. *Cooperation and Conflict 58*(2).

Wysmułek, I., I. Tomescu-Dubrow, and J. Kwak (2022). Ex-post Harmonization of Cross-national Survey Data: Advances in Methodological and Substantive Inquiries. *Quality & Quantity 56*, 1701–1708.