Best Practices for Getting Past the 'Janitor Work' Data Wrangling Before Harmonization:*

Yue Hu[1] Yuehong Cassandra Tai[2] Frederick Solt[3]

This article focuses on a preliminary step in any ex-post data harmonization project—wrangling the pre-harmonized data—and suggests best practices for helping scholars avoid errors in this often-tedious work. To provide illustrations of these best practices, the article uses the examples of pre-harmonizing procedures used to produce the Standardized World Income Inequality Database (SWIID), a widely used database that uses Gini indices from multiple sources to create comparable estimates, and the Dynamic Comparative Public Opinion (DCPO) project, which creates a workflow for harmonizing aggregate public opinion data.

[1] Department of Political Science, Tsinghua University, Beijing, China

[2] Center for Social Data Analytics, Pennsylvania State University, University Park, USA

[3] Department of Political Science, University of Iowa, Iowa City, USA

[*] Corresponding author: yhcasstai@psu.edu. Current version: May 12, 2025. Replication materials and complete revision history may be found at https://github.com/fsolt/wrangling_data. The authors contributed equally to this work. Yue Hu appreciates the funding support from the National Natural Science Foundation of China (72374116) and Tsinghua University Initiative Scientific Research Program (2024THZWJC01).

1 A Wrangling Issue of Data Harmonization

Empowered by the spreading Internet and advancing computational power, researchers have entered an unprecedented age of data availability. A growing volume of social science research aims to take the benefit to extend the generality: they employ large quantities of data drawn from different sources. However, ensuring the quality of harmonized datasets remains a significant challenge in handling fitness to use and raw data quality monitoring among others (Slomczynsi, Tomescu-Dubrow, and Wysmulek 2025). Beyond the focus on the harmonization process itself, we argue that quality assurance must begin *earlier* to the data-wrangling step where raw inputs are selected, processed, and prepared for harmonization.

The wrangling step determines the quality of data in the harmonization process, and the challenges is how to properly and transparently clean the increasing amount and diversity of data. The conventional approach usually involves a notorious bulk of manual work on indicator identification, data merging, data scaling, and so on (see, e.g., Lohr 2014). The tiresome task is easy to introduce errors in data collection procedure. Manual wrangling make a full reproducibility of research pipeline more difficulty and undermine the transparency (Liu and Salganik 2019).

These challenges are amplified when raw data comes from heterogeneous sources and have been processed using various software environments over time. This is a common scenario in secondary data collection. For example, older survey files stored in SPSS's ASCII or portable formats often require extensive restructuring before they can be merged with new format of data. Such undocumented transformations make it difficult to track changes and undermine transparency.

Finally, even meticulous documentation cannot eliminate the influence of human discretion embedded in manual processing. Such discretion leaves behind few traces, making it difficult for collaborators or reviewers to verify the wrangling process or trace sources of error.

In short, poor source data quality, the absence of reproducibility, and untrackable human discretion in manual janitor work have collectively became the largest obstacle on the way to data harmonization, which yet have thus far gained little attention.

In this article, we provide a practical routine taken the advantage of automatic programming and team work to reduce such data-entry errors and improve the reproducibility and transparency of the wrangling process for researchers and reviewers to check the errors. The routine includes three steps: data selection, data entry, and opening. We illustrate how researchers use this routine on statistical (*hard*) and opinion (*soft*) data with two ongoing harmonization efforts, the Standardized World Income Inequality Database (SWIID) and the Dynamic Comparative Public Opinion (DCPO) project.

2 A 3-Step Routine for Data Harmonization

Our routine aims to helping researchers reach three goals for scientific research:

1. To incorporate as much available data as possible to provide base for comparable data and increase generality of the inferences;
2. To reduce the manual entry errors to improve the accuracy of the harmonized data and analytic data; and
3. To improve the reproducibility of data wrangling process for the sake of transparency.

The routine decomposes a data-wrangling process into three steps:

1. Team-based concept construct and data selection;
2. Data entry automation; and
3. "Second-order" opening.

To illustrate the above routine, we use two data harmonization projects as examples, SWIID and DCPO. SWIID is a long-running project that seeks to provide harmonized income inequality statistics for the broadest possible coverage of countries and years (Solt 2009, 2015, 2016, 2020). As of its most recent update at the time of this writing, its source data consists of some 27,000 observations of the Gini coefficient of income distribution in nearly 200 countries over as many as 65 years, collected from over 400 separate sources including international organizations, national statistics bureaus, and academic studies.

DCPO is both a method and a database. Scholarship on comparative public opinion only rarely benefits from relevant items asked annually by the same survey in many countries (see, e.g., Hagemann, Hobolt, and Wratil 2017). To address the lack of cross-national and longitudinal data on many topics, a number of works have presented latent variable models that harmonize available but incomparable survey items (see e.g., Caughey, O'Grady, and Warshaw 2019; Claassen 2019; Kołczyńska et al. 2024). Along this line, DCPO not only provides latent variable measurements but also automatized and reproducible data collection (Solt 2020), which has been applied in a complete pipeline for a variety of topics such as gender egalitarianism (Woo, Goldberg, and Solt 2023), political interest (Hu and Solt 2024), and support for gay rights (Woo et al. 2024), among other aspects of public opinion and open it freely for global researchers (see more updated data collections at https://dcpo.org/).

In the following sections we first address the common challenges for the phases of data wrangling and explain how our routine can help deal with it illustrated with the data wrangling processes of the SWIID and DCPO projects.

2.1 Step 1: Team-Based Construct Building and Data Selection

Large scale of data selection and cleaning is almost always tedious, as something to be delegated to research assistants, to someone—indeed anyone, but usually research assistants (RA)—else (see Torres 2017). This manual procedure is easy to make mistakes and errors. Haegemans, Snoeck, and Lemahieu (2019, 1) has demonstrated examples of misrouted financial transactions and airline flights. In a more systematic examination, Barchard and Pace (2011) found that RA assigned in an experiment to carefully enter data manually, even those instructed to double-check their entries against the original, had error rates approaching 1% in just a single roughly half-hour session. The consequences of such errors can be pernicious.

Our antidote for this issue is a combination of team work and automation. We will focus more on the team work and discuss the latter in OSM 2.2. The goal here is to have consistent understanding on conceptualized construct, select valid data for later measurement and/or analyses, and reduce biases caused by inconsistent human judgment. A team work framework for this end requires a deliberative set and a dual-entry process.

A deliberative set requires the members in a research team—regardless several coauthors or a primary author with one or two RAs—to have a clear and coherent understanding of the reseach questions and associated data goals. These understandings will help the team members identify the right data to collect and discover extra useful data sources that are not in the initial plan.

In the SWIID program, for example, we told RAs that the goal of the research is to generate comparable statistics of country-level economic inequality. We provide a list of sources mainly from national statistic bureaus for them to start, but we also told them that update statistics for some countries may come from academic papers, published documents, and other sources, and they are free to add them in while making sure a valid link of the new sources are also recorded.

Ensuring team members to understand how the data would use later is also important, as they could have a better sense of what data are analyticable and a forward perspective of how many situations would the later entry part need to take care. In the SWIID project, we told the RAs that the inequality statistics be recorded in four formats: Gini index in disposable (post-tax, post-transfer) income, Gini in market (pre-tax, pre-transfer) income, absolute redistribution (market-income inequality minus net-income inequality), or relative redistribution (market-income inequality minus net-income inequality, divided by market-income inequality). So, for later unification work, they need not only to record the digits but also seek documents to explain the methods of the statistics.

The SWIID project requires update for almost every year and we also often hire new RAs. Therefore, the cross-check is done in a rolling basis usually by the rookies who are in charge of checking the old data and updating malfunctional links. This is both a learning process and a way to improve data accuracy.

In the DCPO project, clearly defining and agreeing upon the latent construct among team members is a critical first step for ensuring theoretical comparability across countries and over time (Koc and Kołczyńska 2025). This process begins with a shared conceptual foundation established through literature review and corresponding pre-defined potential dimensions of the latent opinion. Each team member is then assigned survey datasets from specific geographic regions and tasked with identifying potentially relevant items and potential dimensions based on both general theoretical guidance and region-specific knowledge. This structure ensures that the construct is informed by both global theory and local context.

Before data selection begins, team members undergo hands-on training on how the method work and what type of data and detail they need to collect, such as data format and weighting types, which provide a valuable help of later build the automative data preparation software.

Following the initial round of item selection and collection, the dural-entry section comes in. In this stage, each team member reviews and re-codes the survey data originally handled by another member. The independently coded versions are then compared to detect discrepancies, which may arise from misinterpretations of the construct, ambiguous item wording, or common entry errors.

Disputed cases are flagged for group discussion. Some mismatches may indicate items that may not be conceptually equivalent across cultures or regions, and others suggest multidimensionality that requires theoretical disaggregation. For the latter, we either categorize such items into pre-defined dimensions and/or revise the codebook accordingly to add new dimensions—an iterative process aimed at improving construct validity, intercoder reliability, and reducing oversimplification of target variable (Slomczynsi, Tomescu-Dubrow, and Wysmulek 2025).

Therefore, we broke down the cross-check step into several lab meetings interspersed during the data selection to collect new insights from each members' selection works and make sure everyone were on the same page through the whole process. The process ends with a systemic cross-check of the final selected data among members.

2.2 Step 2: Data Entry Automation

Formatting data is arguably the easiest step to involve manual errors and controversies. The best solution is to automate the entry process taken the advantages of the programming languages and application programming interfaces (APIs) of the data source.

In the DCPO case, data entry is fully automated through the R-based software, `DCPOtools` (Solt, Hu, and Tai 2018). This software processes raw survey files directly, ensuring reproducible data entry. It converts various file formats to R-readable objects, extracts variables of interest, reorders response values, applies survey weights, and aggregates weighted respondents by country and year based on actual fieldwork dates.

To address theoretical comparability concerns, DCPO employs conservative filtering, removing items appearing in fewer than five country-years in countries surveyed at least three times, minimizing the risk of sacrificing comparability for coverage (Koc and Kołczyńska 2025). `DCPOtools` standardizes country names using Arel-Bundock, Enevoldsen, and Yetman (2018)'s `countrycode` and ensures years reflect actual fieldwork dates, creating aggregated respondent data for the latent variable model.

While coding datasets and items into structured spreadsheets facilitates automation, an

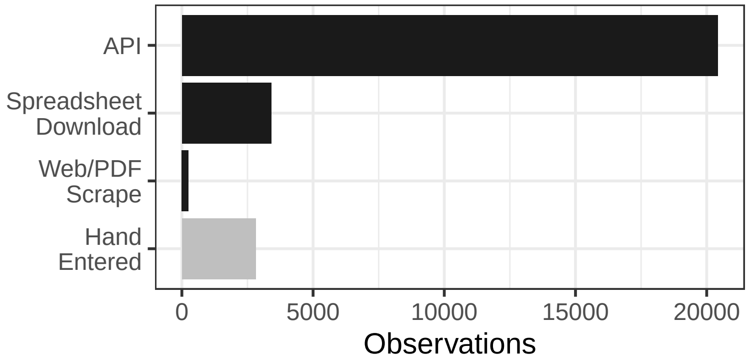

Figure 1: Income Inequality Observations by Method of Collection

even better version starts the automation since the data selection step via programming and APIs. As shown in Figure 1, the current version of SWIID grapes 76% of the observations through API. When no API is available, the automation script downloads and reads any available spreadsheets (see Wickham 2016). In the absence of a spreadsheet, the process of scraping the data either directly from the web or, preferably, from a pdf file (see Sepulveda 2024) is automated. Together the collection of 90% of the source data is scripted. This means not only that the possibility of errors introduced by hand entry for a vast majority of observations is eliminated but also that the updates and revisions that are frequent in these data are automatically incorporated as they become available.[1]

For data sources, such as those from academic articles or books, that have to be entered in hand, there is still rooms for automation. For the remaining 10% of the SWIID observations, for instance, we collected them using Sepulveda's `tabulapdf` R package to avoid data-entry errors as long as they are in pdf (Sepulveda 2024). The advanced Optical Character Recognition (OCR) can extend this method on data sources even in hard copies.

And finally, one source of SWIID contains crucial information encoded in the typeface of its tables (see Mitra and Yemtsiv 2006, 6); this information would be lost if the tables were read directly into R. We reapplied the approach from the data selection here to enter them twice into separate spreadsheets.[2] The dual-entry process allows for automated cross-checks of the newly entered data that increase the chances that errors are identified and corrected (see Barchard and Pace 2011).

2.3 Step 3: "Second-Order" Opening

Since the replication crisis, replication files for analytical results in academic articles has become a standard requirement for top-tier journals in political science (Chang and Li 2015; Open Science Collaboration 2015). Nevertheless, the continual raising controversies on the researcher degrees of freedom indicated that current open is still not adequate.[3] Especially in relation with data harmonization, we eager researchers to conduct a, what we called, the "second-order" opening. That is, not only opening analytical steps (the "first-order") but also the data generation process (the "second-order"), including data collection, data cleaning, and data wrangling, as mentioned above.

If researchers applied our suggestions of team-based construct building, systematic data selection, and automated data entry, the second-order opening becomes both feasible and efficient. Along with a clearly conceptualized theoretical framework, researchers can simply share their programming scripts for data downloading, formatting, and wrangling, ensuring that the full pipeline is documented and reproducible.

With developed scientific and technical publishing system, such as Quarto or R markdown, and version control platforms (e.g., Github) and open collaboration platforms (e.g., Open Science Framework, OSF), researchers can integrate the entire workflow—from raw data collection to final analysis—within a single, publicly trackable archive. We reached at this step for all the DCPO projects so far. Readers can find a Github repo for the research from scratch, and every wave of data update in the corresponding OSF project.[4]

3 Discussion

Implementing these open-science practices requires effort (see Engzell and Rohrer 2021). Though labor-intensive, the double-entry method reduces error rates thirty-fold (Barchard and Pace 2011, 1837), justifying the investment. Teamwork distributes tasks, reducing fatigue-related errors, while allowing discrepancies to be resolved through discussion.

Social scientists now benefit from standardized harmonization workflows (Slomczynsi, Tomescu-Dubrow, and Wysmulek 2025) and automated data processing (Kritzinger, Lutz, and Boomgaarden 2025). Researchers can reuse high-quality harmonized datasets, enhancing efficiency and comparability. Open-source software packages like those used

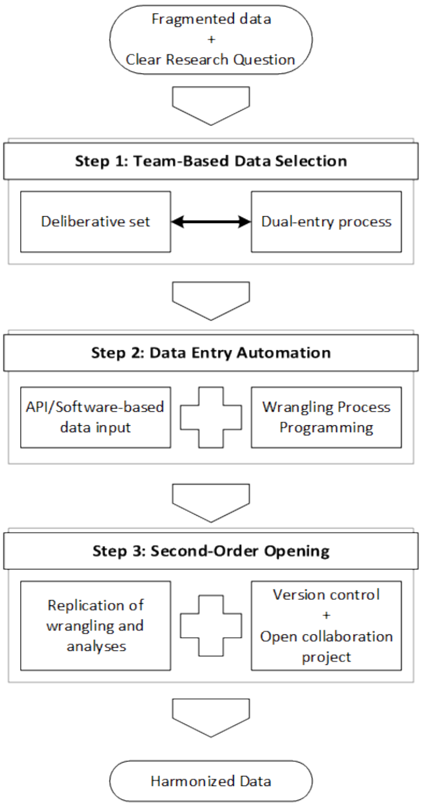
Fragmented data
+
Clear Research Question
Step 1: Team-Based Data Selection
Deliberative set
Dual-entry process
Step 2: Data Entry Automation
API/Software-based data input
Wrangling Process Programming
Step 3: Second-Order Opening
Replication of wrangling and analyses
Version control
+
Open collaboration project
Harmonized Data

Figure 2

by the SWIID and `DCPOtools` have already automated many data preparation tasks. With large language models emerging, intelligent agents may soon handle parts of these routines, potentially advancing automation to new levels (Kritzinger, Lutz, and Boomgaarden 2025).

A final point we would like to clarify is that, in our three-step routine, researchers remain central to harmonization. As illustrated in the SWIID and DCPO examples, researchers are responsible for all critical decisions from clarifying research questions and building theoretical constructs to conducting version control and developing replication materials. Early and critical steps, such as construct development and codebook refinement, must be conducted iteratively to achieve high intercoder reliability. Even with automated data entry, human validation remains essential for verifying variable formats and value ranges. Computing environments should be documented to minimize system-related discrepancies (Liu and Salganik 2019).

For ex-post harmonization projects, careful attention to pre-harmonization stages substantially contributes to overall dataset quality. While some error is inevitable, with responsible researcher oversight, data-entry errors can be minimized while transparency, openness, and research credibility continue to grow.

Notes

[1] The R community has often built software to ease the access of APIs and make the batch work for multiple waves of data in a more comfortable and efficient way (see Blondel 2018; Lahti et al. 2017; Lugo 2017; Magnusson, Lahti, and Hansson 2014; Wickham, Hester, and Ooms 2018).

[2] Most often this has been done by two different investigators, but sometimes sequentially by a single researcher.

[3] See a summary of the "researcher degrees of freedom" literature in Hu, Tai, and Solt (2024).

[4] See a comprehensive example applied the second-order opening strategy in Tai, Hu, and Solt (2024).

References

Arel-Bundock, Vincent, Nils Enevoldsen, and C. J. Yetman. 2018. "countrycode: Convert Country Names and Country Codes." *Journal of Open Source Software* 3(28): 848–49.

Barchard, Kimberly A., and Larry A. Pace. 2011. "Preventing Human Error: The Impact of Data Entry Methods on Data Accuracy and Statistical Results." *Computers in Human Behavior* 27(5): 1834–39.

Blondel, Emmanuel. 2018. "rsdmx: Tools for Reading SDMX Data and Metadata."

Caughey, Devin, Tom O'Grady, and Christopher Warshaw. 2019. "Policy Ideology in European Mass Publics, 1981–2016." *American Political Science Review*: 1–20. doi:10.1017/S0003055419000157.

Chang, Andrew, and Phillip Li. 2015. "Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say 'Usually Not'." *Finance and Economics Discussion Series* 7: 1–25.

Claassen, Christopher. 2019. "Estimating Smooth Country–Year Panels of Public Opinion." *Political Analysis* 27(1): 1–20.

Engzell, Per, and Julia M. Rohrer. 2021. "Improving Social Science: Lessons from the Open Science Movement." *PS: Political Science & Politics* 54(2): 297–300.

doi:10.1017/S1049096520000967.

Haegemans, Tom, Monique Snoeck, and Wilfried Lemahieu. 2019. “A Theoretical Framework to Improve the Quality of Manually Acquired Data.” *Information & Management* 56(1): 1–14.

Hagemann, Sara, Sara B. Hobolt, and Christopher Wratil. 2017. “Government Responsiveness in the European Union: Evidence from Council Voting.” *Comparative Political Studies* 50(6): 850–76.

Hu, Yue, and Fredrick Solt. 2024. “Macrointerest Across Countries.” In Shenzhen.

Hu, Yue, Yuehong Cassandra Tai, and Frederick Solt. 2024. “Revisiting the Evidence on Thermostatic Response to Democratic Change: Degrees of Democratic Support or Researcher Degrees of Freedom?” *Political Science Research and Methods*: 1–7. doi:10.1017/psrm.2024.16.

Koc, Piotr, and Marta Kołczyńska. 2025. “Modeling Trends in Public Opinion: An Overview of Approaches, Assumptions and Trade-Offs.” *Working Paper*.

Kołczyńska, Marta, Paul-Christian Bürkner, Lauren Kennedy, and Aki Vehtari. 2024. “Modeling Public Opinion over Time and Space: Trust in State Institutions in Europe, 1989-2019.” *Survey Research Methods* 18(1): 1–19.

Kritzinger, Sylvia, Georg Lutz, and Hajo Boomgaarden. 2025. “Visions for the Future: Challenges and Opportunities for Creating Sustainable Scholarly Infrastructures for Data Harmonization.” *Working Paper*.

Lahti, Leo, Janne Huovari, Markus Kainu, and Przemysław Biecek. 2017. “Retrieval and Analysis of Eurostat Open Data with the Eurostat Package.” *The R Journal* 9(1): 385–92.

Liu, David M, and Matthew J Salganik. 2019. "Successes and Struggles with Computational Reproducibility: Lessons from the Fragile Families Challenge." *Socius* 5: 2378023119849803.

Lohr, Steve. 2014. "For Data Scientists, 'Janitor Work' Is Hurdle to Insights." *New York Times*: B4.

Lugo, Marco. 2017. "CANSIM2R: Directly Extracts Complete CANSIM Data Tables."

Magnusson, Mans, Leo Lahti, and Love Hansson. 2014. "pxweb: R Tools for PX-WEB API."

Mitra, Pradeep, and Ruslan Yemtsiv. 2006. "Increasing Inequality in Transition Economies: Is There More to Come?"

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349(6251): aac4716.

Sepulveda, Mauricio Vargas. 2024. "tabulapdf: Extract Tables from PDF Documents."

Slomczynsi, Kazmierz M., Irina Tomescu-Dubrow, and Ilona Wysmulek. 2025. "Navigating Complexities of Ex-Post Harmonization of Cross-National Survey Data: Insights from the Survey Data Recycling, SDR, Project." *Working Paper*.

Solt, Frederick. 2009. "Standardizing the World Income Inequality Database." *Social Science Quarterly* 90(2): 231–42.

Solt, Frederick. 2015. "On the Assessment and Use of Cross-National Income Inequality Datasets." *Journal of Economic Inequality* 13(4): 683–91.

Solt, Frederick. 2016. "The Standardized World Income Inequality Database." *Social Science Quarterly* 97(5): 1267–81.

Solt, Frederick. 2020. “Measuring Income Inequality Across Countries and over Time: The Standardized World Income Inequality Database.” *Social Science Quarterly* 101(3, 3): 1183–99. doi:10.1111/ssqu.12795.

Solt, Frederick, Yue Hu, and Yuehong Tai. 2018. “DCPOtools: Tools for Dynamic Comparative Public Opinion.”

Tai, Yuehong ‘Cassandra’, Yue Hu, and Frederick Solt. 2024. “Democracy, Public Support, and Measurement Uncertainty.” *American Political Science Review* 118(1): 512–18. doi:10.1017/S0003055422000429.

Torres, Rachel. 2017. “Me: Shouldn’t There Be Someone in a Basement That We Just Pay to Do All This Awful Data Cleaning? Advisor: That’s Who You Are.”

Wickham, Hadley. 2016. “rvest: Easily Harvest (Scrape) Web Pages.”

Wickham, Hadley, James Hester, and Jeroen Ooms. 2018. “xml2: Parse XML.”

Woo, Byung-Deuk, Lindsey A. Goldberg, and Frederick Solt. 2023. “Public Gender Egalitarianism: A Dataset of Dynamic Comparative Public Opinion Toward Egalitarian Gender Roles in the Public Sphere.” *British Journal of Political Science* 53(2): 766–75. doi:10.1017/S0007123422000436.

Woo, Byung-Deuk, Hyein Ko, Yuehong Cassandra Tai, Yue Hu, and Frederick Solt. 2024. “Public Support for Gay Rights Across Countries and over Time.” *Social Science Quarterly* 106(1): e13478. doi:10.1111/ssqu.13478.