Aletheia
RESEARCH BRIEFINGS

Constructing the BLOCS 4.0 Integrated Panel Database: A Technical Methodology

ALEJANDRA BIANKA DELGADILLO CANQUI
CLARK N. BANACH
IGNACIO SILVA NEIRA
FELIPE VALENZUELA

August 2026

BLOCS Lab Briefing No. 07/2026
Aletheia Research Institution

I. OVERVIEW

This document provides a detailed technical description of the construction methodology used to develop Bilateral Longitudinal Observations and Country Statistics (BLOCS) database, version 4.0. BLOCS is an integrated panel database that systematically combines bilateral trade, investment, institutional, and macroeconomic data between 1948 - 2023 from 16 international sources.

The briefing is structured to reflect the construction workflow, beginning with the foundational country code standardization ("Rosetta Stone" system), proceeding through individual dataset processing for trade (BACI, WTF), investment (UNCTAD, IMF) and institutional factors (Polity IV, WGI). It serves as a technical guide for users interpreting BLOCS variables and as a reproducibility framework for researchers interested in integrated economic data construction, addressing challenges like historical country transitions, identifier reconciliation, and scalable merging architectures for large-scale panel data. When reviewing the replication file some stages will appear out of order as some data has been grouped by category and historic tariff data was moved to the end.

In addition to our standard codebook, this documentation examines the procedural architecture, sequential data processing, standardization protocols, and merging methodologies that transform disparate datasets into a coherent analytical resource.

BLOCS has been developed to aggregate bilateral observations related to trade, international investment, and characteristics of agreements between pairs of countries. To explore these relationships within the context of international institutional changes, macroeconomic and historic institutional information about the countries is included, allowing for a broader analysis of trade. References for all sources and methods can be found in the accompanying codebook and spreadsheet.

Database	General Information	Source	Dates
Direction of Trade Statistics	Bilateral trade data	International Monetary Fund. (2020a)	1948 - 2023
Bilateral Product Trade Flows (BACI)	Bilateral trade data	CEPII (2020a); Lall (2000)	1995 - 2023
World Trade Flows	Bilateral trade data; *historic control	Feenstra, R., & Romalis, J. (2016).	1984 - 2016
Regional Trade Agreements	Trade agreements	Egger, P., & Larch, M. (2008).	1950 - 2023
Design of International Trade Agreement (DESTA)	Trade agreements	Dür, A. B., & Elsig, M. (2014)	1948 - 2023
Deep Trade Agreements	Trade agreements	Mattoo, A., Rocha, N. & Ruta, M. (2020)	1958 - 2023
FDI/TNC	Foreign Direct Investment	UNCTAD (2020)	2000 - 2023
Coordinated Direct Investment Survey	Foreign Direct Investment	IMF (2020b)	2009 - 2023
Balance of Payments and International Investment Position Statistics (BOP/IIP)	Balance of Payment data	IMF (2020d)	1948 - 2023
International Financial Statistics	Macroeconomic attributes and financial characteristics	IMF (2020c)	1948 - 2023
Worldwide Governance Indicators	Governance	World Bank (2021)	1996 - 2019
Regime Authority Characteristics and Transitions Dataset	Democratic and Autocratic patterns of authority	Center for Systemic Peace (2018)	1948 - 2018
Meta Analysis EMU Trade Effects	Geographic and monetary information.	Rose A. (2017)	1948 - 2022
Gravity characteristics	Trade, geographical and cultural information	CEPII (2020b)	1948 - 2021
Countries or areas / geographical regions	Standard country or area codes (M49)	United Nations (2026)	1948 - 2023
Tariff Data	Tariff *historic control	Fuceri & Rose (2019)	1960 - 2014

Table 1: Sources of Observations in BLOCS

II. BLOCS ROSETTA STONE and DOTS

The construction of BLOCS begins with the Direction of Trade Statistics (DOTS) dataset, which provides the core bilateral trade observations. However, DOTS uses only IMF country codes (countrycode), while other databases employ different identification systems. To create a universally mergeable foundation, we perform two sequential merges with our novel BLOCS Rosetta Stone.

First, we match origin countries to obtain standardized codes (UN and ISO Alpha-3). Then, we repeat the process for partner countries. This dual merge creates three comprehensive pair identifiers: ifs_pairid (IMF), un_pairid (UN), and alpha_pairid (ISO). Historical cases like Yemen require special handling, with code adjustments for pre- and post-unification periods (1991 boundary). This foundational dataset establishes the complete country-pair panel that will receive all subsequent data. The procedure harmonizes IMF IFS, UN, and ISO Alpha-3 codes for dataset merging.

Rosetta Stone Merge:

```
1. Drop specific codes from DOTS: Andorra (AND), Liechtenstein (LIE), Monaco
(MCO), Romania old code (ROM), Byelorussia old (BYS), Rhodesia (RHO), South
Vietnam (VDR), Zaire (ZAR), East Timor old (TMP), Burma old (BUR), USSR (SUN),
and empty codes.

2. Standardize codes: Replaces "YDM" with "YDR" for Democratic Yemen

3. Create partner code variables (p_alpha_3, p_ifs_num, p_un_num) for pair IDs

4. Modify Rosetta Stone for FDI:
   - Loads country code table for FDI data
   - Standardizes variable name: o_country → countryname
   - Saves cleaned mapping table

5. Add Continents Data:
   - Loads continent classification from UNSD (United Nations)
   - Standardizes UN code: m49code → un_num (3 digits)
   - Creates partner variable: p_un_num
   - Encodes least developed countries: "x" → "1", others → "0"
   - Converts to numeric: leastdevelopedcountriesldc
```

III. BACI INTEGRATION: PRODUCT-LEVEL TRADE DETAIL

The BACI database provides highly disaggregated trade data. Due to its size, BACI is stored as 28 separate annual files (1996-2023). The integration uses a sequential merge approach. We first merge the most recent year (2023), then loop through previous years using the update option. This method ensures the most recent data takes precedence. Each merge matches on un_pairid and year.

After the sequential merge, disaggregated product-level bilateral trade data is used to construct measures of technological classification. The technological sophistication panel integrates through a 1:1 exact match merge using un_pairid and year. Unmatched records from the baseline database are filled with zeroes, while unmatched observations from the using dataset are dropped. This step renames variables value1 through value5 to trade_lall_1 through trade_lall_5 to store the final technological trade tiers.

BACI (Bilateral Product Trade Flows) Preparation:

1. Create output directory: BACI separated

2. Loop through years 1995-2023 loading annual BACI CSV files

3. Standardize 3-digit country codes: France: 251→250 USA: 842→840 (includes Puerto Rico & Virgin Islands) Belgium: 058→056 Italy: 381→380 India: 699→356 Switzerland: 757→756 Norway: 579→578 Taiwan: 490→158

4. Create unique pair ID: un_pairid = i+j (origin+partner)

5. Reshape data from long to wide format by product categories

BACI HS2 Technological Classification (SITC-Lall Concordance) Merge:

1. Loop through years 2007–2023 loading annual BACI CSV files

2. Standardize 3-digit country codes: France (251→250), USA (842→840), Belgium (058→056), Italy (381→380), India (699→356), Switzerland (757→756), Norway (579→578), Taiwan (490→158)

3. Transform product categories by merging with the SITC Rev. 3 classification table (JobID-38_Concordance_H2_to_S3.dta)

III. WTF INTEGRATION: HISTORIC TRADE FLOW VALIDATION

With the standardized foundation established, we integrate World Trade Flows (WTF) data through a simple 1:1 exact match merge using alpha_pairid and year. WTF serves as an independent validation source for variation between DOTS and BACI trade totals. The merge uses the nogenerate option since WTF contains only country pairs already present in our foundation dataset. This step adds the wtf_tot_trade variable, enabling direct comparison with DOTS trade values for robustness checks. Since the dataset was discontinued in 2016 it acts as a historic control and additional measure for economic history. The procedure processes bilateral trade flow data to validate and supplement DOTS and BACI trade statistics.

WTF (World Trade Flow) Preparation:

1. Loop through 33 country files (1984-2016)

2. Standardize country codes: Yemen: YDM→YDR, France: 251→250, USA: 842→840, Belgium: 058→056, Italy: 381→380, India: 699→356, Switzerland: 757→756, Norway: 579→578, Taiwan: 490→158

3. Rename variables to match BLOCS standards

4. Append all country files into single dataset

5. Merge with Rosetta Stone for complete country coding

6. Creates multiple pair IDs: alpha_pairid, un_pairid, ifs_pairid

7. Remove duplicates by country pair and year

V. LARCH TRADE AGREEMENTS: BASIC AGREEMENT INDICATORS

The Larch Regional Trade Agreements database integrates with minimal transformation. We primarily correct historical country codes (Yemen and Belgium-Luxembourg Economic Union) before performing a 1:1 exact match merge using alpha_pairid and year. This adds seven dummy variables indicating agreement types (rta, cu, fta, psa, eia, cueia, ftaeia) and creates the composite variable RTA_larch_noEIA that identifies any active trade agreement. The procedure processes trade agreement data to identify different types of regional trade agreements between country pairs.

Larch (Regional Trade Agreements Database) Preparation:

1. Standardize country codes: Yemen: YDM→YDR

2. Modify Belgium-Luxembourg historical union: BEL→BLX for years ≤1996

3. Create alpha_pairid from exporter+importer

4. Generate composite RTA variable (RTA_larch_noEIA) for CU, FTA, and PSA agreements

V. DESTA INTEGRATION: DETAILED AGREEMENT CHARACTERISTICS

The Design of International Trade Agreements (DESTA) database demands the most complex integration. After extensive preprocessing to convert treaty-level data to bilateral format, we perform two directional merges. We first match origin-partner pairs, then partner-origin pairs to ensure bidirectional coverage.

The post-merge process includes forward-filling of missing agreement characteristic**s** within country pairs, propagating values from earlier to later years when agreements remain active. Finally, we merge pre-created dummy variables (gstp, RTA_DESTA_wPTA, PTA_DESTA) that simplify analysis of specific agreement types. The procedure processes data on trade agreements with agreement characteristics and creates specialized dummy variables for pairwise trade agreement analysis using DESTA.

DESTA (Design of International Trade Agreements) Merge:

1. Merge treaty lists with agreement characteristics

2. Modify treaty numbering and year standardization

3. Drop specific territories: Cook Islands (184), Niue (570), Western Sahara (732), Liechtenstein (438), Andorra (20), Monaco (492)*****

4. Standardize 3-digit UN country codes

5. Create un_pairid from iso1+iso2

6. Remove 91 empty variables (c1-c91, mc1-mc91)

7. Exclude non-PTA agreements: African Economic Community, Cotonou, GSTP, SPARTECA, Lomé agreements

8. Duplicate and maintain unique country pair-year observations

9. Create mirrored dataset for bidirectional pair analysis (A-B y B-A)

DESTA Dummy Variables Creation:

1. Load previously saved DESTA dataset

2. Create 3 key dummy variables: gstp (Global System of Trade Preferences -base_treaty 0461/461b) RTA_DESTA_wPTA(All trade agreements except African Economic Community) PTA_DESTA (Specific partner trade agreements :African EC, Cotonou, GSTP, SPARTECA, Lomé)

3. Modify duplicate country pair-year observations

4. Drop specific territories: Cook Islands (184), Niue (570), Western Sahara (737), Liechtenstein (438), Andorra (020), Monaco (492)

5. Create mirrored dataset for bidirectional analysis

VI. DEEP TRADE AGREEMENTS: ROBUST TRADE AGREEMENT MEASURE

Information on the type of agreements, as well as their corresponding IDs (WBID), comes preprocessed as bilateral data with one observation per year. Thus, a simple 1:1 merge is possible using alpha_pairid and year. We follow this by dropping the unmatched observations from the *using* data (observations not available in BLOCS). This provides a robust control to the Larch and DESTA trade agreement measures by adding variables for each type of agreement that is active during each observed year (FTA, PSA, EIA and CU). The procedure processes bilateral agreement data to validate and supplement Larch and DESTA agreement observations.

Deep Trade Agreements (World Bank) Preparation:

1. Open the "bilateral information" sheet provided in the XLSX file

2. Create a variable for each type of trade agreement, equaling 0 in all rows (WB_FTA, WB_PSA, WB_EIA, WB_CU)

3. Replace these variables with 1 when a country pair has an agreement in place for the given year (similar to the structure used by Mario Larch)

4. Create a deep_withdraw binary variable that equals 1 if an agreement was withdrawn during this year

5. Whenever more than one agreement is in place during a year, the types of agreements as well as their corresponding ID's (WBID) are merged into one per row (as a string and separated by "&")

VII. UNCTAD FDI: INVESTMENT FLOW DATA

UNCTAD's Foreign Direct Investment data is modified into a bilateral panel format using appended files of stocks and flows obtained directly from the Investment and Enterprise office at UNCTAD. Files for each year are transformed from wide to long and missing data are identified and coded for robustness.

Integration involves a straightforward 1:1 exact match merge using un_pairid and year. This adds eight FDI variables covering investment stocks and flows in both directions, completing the investment dimension of BLOCS. The procedure processes Foreign Direct Investment data from UNCTAD, of multiple metrics and time periods, into a single operational bilateral panel.

UNCTAD FDI Panel Creation:

1. Process 8 FDI variables across 257 country files: inflow, inflow-M, outflow, outflow-M, instock, instock-M, outstock, outstock-M

2. Transform data structure from wide to long format by country and year

3. Modify missing sheets/files with error capture

4. Append all countries for each FDI variable type

5. Merge all 8 FDI variables into single dataset

6. Integrate with Rosetta Stone for country code standardization

7. Drop obsolete countries codes after Rosetta Stone integration: Vietnam, USSR, East Germany, Czechoslovakia, Iran, Swaziland

8. Create un_pairid for country pair identification

9. Modify duplicate country pairs (Democratic Republic of Congo vs Zaire)

VIII. IMF INVESTMENT DATA: COORDINATED DIRECT INVESTMENT SURVEY

The IMF CDIS data integrates through a 1:1 exact match merge using ifs_pairid and year. This adds two key variables: inward_investment_IMF and outflows_investment_IMF, providing an alternative source of investment position data that complements UNCTAD's figures. The procedure processes Foreign Direct Investment position data from the International Monetary Fund.

IMF CDIS (Coordinated Direct Investment Survey) Preparation:

```
1. Create ifs_pairid from countrycode and counterpartcountrycode

2. Rename variables: timeperiod → year, inwarddirectinvestmentpositionsu →
inward_investment_IMF, v80 → outflows_investment_IMF
```

IX. NATIONAL CHARACTERISTICS: SEQUENTIAL INTEGRATION

Four databases containing country-level characteristics follow an identical integration pattern. Each undergoes dual many-to-one merges: first attaching origin-country variables using origin codes and year, then partner-country variables using partner codes and year. This efficient pattern attaches national characteristics to both sides of each trading relationship. The national characteristics data are composed of the following: 1. BOP: Balance of payments components, 2. IFS: Macroeconomic and financial statistics, 3. WGI: Governance quality indicators, 4. Regime Authority Characteristics and Transitions (Polity IV): Democracy and autocracy indicators. The procedure processes these four sources to provide macroeconomic and financial statistics, as well as governance quality indicators and political regime characteristics.

BOP (Balance of Payments) Preparation:

```
1. Loop for origin (o) and partner (p) countries

2. Keep  BOP account components:

   - Capital Account: transfers, acquisitions, total net
   - Current Account: goods services, primary income, secondary income
   - Financial Account: derivatives, net lending, portfolio investment
   - Supplementary Items: errors-omissions, reserves, other investment

3. Rename variables with o_/p_ prefixes and standardized abbreviations

4. Drop 88 generic variables (v1-v88)

5. Save separate datasets for origin and partner countries
```

IFS (International Financial Statistics) Preparation:

```
1. Loop for origin (o) and partner (p) countries

2. Keep key macroeconomic variables: Gross fixed capital formation, Exchange
rates, Financial interest rates (average, deposit, lending), GDP (nominal and
real), Unemployment (levels and rates), Labor force and population, Consumer
price index (CPI)

3. Rename variables with o_/p_ prefixes for origin/partner countries

4. Save separate datasets for origin and partner variables
```

WGI (Worldwide Governance Indicators) Preparation:

1. Convert year to string format

2. Rename code to alpha_3 and drops countryname

3. Create two datasets: Origin country version with "o_" prefix for all WGI variables and Partner country version with "p_" prefix and p_alpha_3

Regime Authority Characteristics and Transitions (Polity IV) Preparation:

1. Drop variables: prior-regtrans, cyear, ccode, country

2. Convert year to string format

3. Standardize country codes: GMY → DEU (Germany) SPN → ESP (Spain) AUS → AUT (Austria) SWD → SWE (Sweden) POR → PRT (Portugal) FRN → FRA (France) ALG → DZA (Algeria)

4. Create two datasets: Origin country version with "o_" prefix variables (similar as before) and Partner country version with "p_" prefix variables

X. GRAVITY AND EMU

The CEPII Gravity database integrates through a 1:1 merge using un_pairid and year, adding geographic, demographic, and cultural variables essential for gravity model estimation. The EMU data from Rose completes the dataset through a final 1:1 merge using ifs_pairid and year, adding variables on customs unions and geographic characteristics. The procedure processes bilateral geographic, demographic, and cultural data for gravity model estimation, as well as common currencies from Rose's EMU trade effects study with additional variables for landlocked and island countries as controls.

Gravity (CEPII Gravity Database) Preparation:

1. Load CEPII Gravity database

2. Standardize 3-digit country codes for origin and destination

3. Drop self-trade observations (country_id_o == country_id_d)

4. Remove Netherlands Antilles (ANT) from dataset

5. Modify historical country splits/unifications: Ethiopia: ETH.1 (pre-1993) vs ETH.2 (post-1992) Germany: DEU.1 (pre-1991) vs DEU.2 (post-1990) India: IND.1 (pre-2002) vs IND.2 (post-2001) Malaysia: MYS.1 (pre-1965) vs MYS.2 (post-1964) Pakistan: PAK.1 (pre-1971) vs PAK.2 (post-1970) Vietnam: VNM.1 (pre-1976) vs VNM.2 (post-1975)

6. Remove invalid country codes (00.)

7. Create un_pairid from standardized country codes

EMU (Economic and Monetary Union)

```
1. Convert cty1, cty2, and year to string format

2. Create ifs_pairid from cty1 + cty2

3. Keep key variables: custrict11 = Customs union/strict monetary policy,
island = Island status, landl = Landlocked status
```

XI. ROSE AND FURCERI: HISTORIC TARIFF DATA

Although only offering observations through 2014, this data provides historic controls for trade and investment beginning in 1960. First, we merge origin-country variables (o_tariff_rose, o_tb_rose, etc.) using ifs_num and year. Then, we merge partner-country variables (p_tariff_rose, etc.) using p_ifs_num and year. Both use many-to-one merges since each country's tariff data applies to all its trading relationships in a given year. The procedure processes tariff and macroeconomic data from the Rose & Furceri database.

Rose & Furceri Tariff Merge:

```
1. Convert ifscode to string format

2. Drop variables: wdicode, lur, ngdp_r, country

3. Rename variables with "o_" prefix for origin country: Tariff →
o_tariff_rose, Tb → o_tb_rose, le → o_le, reer →  o_reer, Gini_net_50 → o_gini

4. Save origin country version

5. Create partner country version by renaming with "p_" prefix

6. Saves partner country version
```

XII. FINAL MERGE

After joining all sources into the master panel, we sort the panel chronologically by country pairs and generate unique indexing keys. The primary pair identifiers across all the standardization systems, alongside time indicators, are prioritized at the front of the database before exporting the definitive release file.

Final Standardization and Output

```
1.- Sort observations by country-pair identifier (ifs_pairid) and year (year).

2.- Create a unique numerical observation index (id) and a composite lookup
identifier (id_ifs).

3.- Reorder key identification variables (id, id_ifs, ifs_pairid, un_pairid,
alpha_pairid, year, year_str) to the primary columns.

4.- Save the compiled master dataset as BLOCS 4.0.dta.
```

Disclosure Statement: This briefing was developed in collaboration with Aletheia's BLOCS Lab at the Berlin School of Economics and Law and received nonfinancial support from Aletheia Research Institution. GenAI was used for code generation, code validation and image creation for this project. Authors transcribed final written content, .do files and validated results of data aggregation with manual robustness checks. References for all sources can be found in the accompanying codebook and spreadsheet.