

Research Data Cleaning, Integration & Validation

Case 01 Central Europe Political Atlas

Xuanhui Zhang | Budapest | sallaizhang@outlook.com

Prepared 8 October 2026

A research dataset is useful when its fields are consistent, its sources are traceable and its gaps are visible. I apply that approach in the Atlas, combining country observations, bilateral trade and monthly macro-driver records into a documented research product.

Client problem

A research team needs comparable information from several sources. Country identifiers, units, periods and missing-value conventions differ. The delivery must make those differences explicit and show which records meet the agreed comparison rules.

Data workflow



Normalize agreed keys, periods and units; retain source links and acquisition metadata; validate values and relationships; join coverage records before comparison; export the data with an exception log and concise handover. Annual and monthly layers remain distinct.

Verified project scale

Data layer	Verified snapshot
Core observations	980 record positions; 67 null-valued and 65 pending records; 980 unique observation IDs. Null and pending categories overlap.
Trade network	46,151 edge records; 220 reporter-year-flow groups; 208 pass and 12 fail the 95% eligible coverage rule.
Monthly macro drivers	14,873 stored records, including levels and transformations; definitions retain source, units, frequency, timing and scope.

Deliverables

Structured JSON or CSV tables, a field dictionary, source register, documented missing values, eligibility and exception summaries, and a handover explaining how the dataset can be used.

Quality control and client value

Quality-control example

Slovakia's 2024 import eligible coverage is 89.1495%, below the 95% threshold. Calculated shares remain in storage, but I withhold them from the comparison and retain the failed group in the exception log. This keeps a stored value from being mistaken for an approved comparison.

Verification checks

Check	Verified result
Eligible coverage and gate decisions	Recalculated from stored edges for all 220 groups; ratios and decisions match.
Eligible totals and partner HHI	Recalculated for all 220 metric rows; matches within floating-point tolerance.
Observation ID uniqueness	980 unique IDs across 980 records.
Source trail and missing records	Observation exports retain source fields and missing-value reasons; trade manifest retains query details and acquisition hashes.

Tools

Atlas uses JavaScript and TypeScript for acquisition, transformation and validation, with JSON and CSV exports. Python supports the independent snapshot checks in this case. My wider toolkit includes Excel and PowerPoint; R, Python and SQL remain developing skills.

What I can do for clients

I can structure multi-source research datasets, standardize identifiers and periods, preserve source trails, document missing or unresolved records, validate agreed rules and prepare concise handover documentation.

What this demonstrates

A repeatable research-data workflow and practical judgment about missingness and comparability. The result is a traceable dataset with explicit eligibility controls and a usable explanation of its limits.

Scope and evidence note

Independent-project case presented through an illustrative client problem. Verification covers the stored exports and the checks listed above; it does not certify freshly reacquired source data or a complete release build. Atlas snapshot 994d61fe1e56, reviewed 8 October 2026.

[public/research-data/observations.json](#)

[src/data/network/network_coverage.json](#)

[scripts/network/build-trade-network.mjs](#)

[public/research-data/macro_driver_observations.json](#)