wbopendata: Fifteen Years of Programmatic Access to World Bank Open Data

João Pedro Azevedo
jpazvd.github.io

Abstract. This article reflects on fifteen years of `wbopendata`, the first Stata command to provide programmatic access to an international development data repository. Today, `wbopendata` provides access to over 29,000 indicators from 51 databases spanning 296 countries and aggregates, with features including five download modes, multilingual metadata, and publication-ready graph formatting. Its broader contribution lies in treating data acquisition as code: indicator selections, country filters, and time ranges become explicit parameters in analysis scripts rather than undocumented manual downloads—addressing the data provenance dimension of the reproducibility crisis that statistical reforms alone cannot fix. Yet the command has never been more relevant. As AI tools accelerate analytical workflows while enabling plausible fabrication of statistics and citations, anchoring research to authoritative, version-controlled data sources becomes essential infrastructure—not legacy convenience. I document the command's latest syntax and stored results, demonstrate workflows from basic queries to choropleth mapping, and present a 44-scenario test suite. I also discuss risks that frictionless data access can obscure—provenance opacity, coverage gaps masked by convenient defaults, and sustainability pressures—alongside mitigation strategies. I distill three design principles from fifteen years of sustained use: backward compatibility builds trust, domain-specific syntax lowers barriers, and scripted data access makes reproducibility the default rather than the exception.

Keywords: Stata, Open Data, World Bank, API, reproducibility, WDI, development indicators

1 Introduction

In April 2010, the World Bank launched its Open Data Initiative (World Bank 2010), marking a structural shift in how official development statistics are disseminated and used. By removing paywalls and releasing a public data portal alongside a programmatic Application Programming Interface (API), the institution reframed decades of development statistics as a global public good.

Within a year of this launch, in February 2011, `wbopendata` (Azevedo 2011) was released as a Stata module to translate the World Bank's newly opened API into a command-based interface familiar to applied researchers. Over the subsequent fifteen years, both the API and the command evolved together: the World Bank expanded from flagship databases to over 50 thematic collections serving 29,000+ indicators; the API retired legacy endpoints (2018–2020) and redesigned its data catalog (2021); `wbopendata` maintained backward compatibility while adding features for metadata inspection, mul-

tilingual support, and publication-ready output formatting.

This fifteen-year trajectory illustrates a broader transformation in statistical computing: the shift from data as downloadable files to data as services—queryable, versioned, and increasingly embedded in analytical pipelines. `wbopendata` addresses this by encapsulating complexity behind a stable Stata command. Its design reflects a specific methodological commitment: **data acquisition should be scripted, parameterized, and version-controlled**, making data provenance explicit and enabling analyses to be reproduced exactly or systematically updated as new data become available.

Despite its age, `wbopendata` has never been more critical. The rapid adoption of AI in scientific workflows has dramatically accelerated how researchers analyze information, write code, and construct narratives. Yet this acceleration makes **analytical guardrails more important, not less**. When AI tools can generate plausible statistics, fabricate citations, and produce synthetic datasets, anchoring analyses to a **single source of truth**—verified institutional data with documented methodology—becomes essential. Tools like `wbopendata` provide that anchor: every data point traces to the World Bank API, every indicator carries transparent provenance, and every query is reproducible. In an era where the bottleneck has shifted from computation to credibility, programmatic access to authoritative data is not a legacy convenience—it is foundational infrastructure.

The remainder of this article proceeds as follows. Section 2 documents the command's syntax, options, and stored results. Section 3 provides reproducible examples ranging from basic downloads to choropleth mapping. Section 4 describes the technical implementation. Section 5 presents the test suite and error handling. Section 6 discusses broader implications for reproducible research, and Section 7 concludes.

2 The wbopendata command

The databases accessible through `wbopendata` include World Development Indicators (WDI), Doing Business, Worldwide Governance Indicators, International Debt Statistics, Africa Development Indicators, Education Statistics, Enterprise Surveys, Gender Statistics, Health Nutrition and Population Statistics, Global Financial Inclusion (Findex), Poverty and Equity, Human Capital Index, Sustainable Development Goals, and many more. Table 1 summarizes the current scope.

Table 1: World Bank Open Data coverage

Dimension	Coverage
Indicators	29,000+
Data sources	51 databases
Topic categories	21
Countries & regions	296
Country attributes	17
Time coverage	1960–present
Languages	3 (English, Spanish, French)

`wbopendata` talks directly to the World Bank API (JSON over HTTP), returns tidy Stata datasets, and caches results to minimize repeat downloads. Five pull modes cover country, topic, single-indicator (all countries), single-indicator (selected countries), and multi-indicator requests. Output may be wide or long; `latest` works in long mode and returns ready-to-use scalars for titles and subtitles. Metadata is always fetched; v17.7.1 adds basic country context (region, admin region, income level, lending type) by default.

2.1 Syntax

`wbopendata, { indicator(`*string*`) | country(`*string*`) | topics(`*string*`) }`
`[ options ]`

Data selection

Exactly one of the following is required:

`indicator(`*string*`)` specifies one or more World Bank indicator codes. Multiple indicators can be requested by separating codes with semicolons, for example, `indicator(SP.POP.TOTL;NY.GDP.PCAP`

`country(`*string*`)` specifies ISO3 country codes or World Bank region codes to retrieve all available indicators for selected countries. Multiple codes can be separated by semicolons.

`topics(#)` specifies a topic ID (1–21) to retrieve all indicators within a thematic category such as education, health, or environment.

Time and language

`year(`*string*`)` restricts the time interval. For example, `year(2000:2020)` returns data only for years 2000 through 2020.

`language(`*string*`)` sets the language for metadata display. Valid codes are `en` (English, default), `es` (Spanish), and `fr` (French).

`projection` accesses population estimates and projections from the Health Nutrition and Population Statistics database rather than actual census data.

Output format

`long` returns data in long format with one row per country-year. The default is wide format with year-specific columns (`yr1960`, `yr1961`, ...).

`clear` replaces any data currently in memory. Required if data are already loaded.

`latest` keeps only the most recent non-missing observation per country. Requires the `long` option. When multiple indicators are requested, retains only observations where *all* indicators have non-missing values in the same year.

`describe` displays indicator metadata without downloading data. Useful for exploring indicator definitions and sources before committing to a full download.

`nometadata` suppresses the metadata display that normally appears after data retrieval.

Country attributes

`basic` adds region, administrative region, income level, and lending type variables to the downloaded data. This is the default behavior in v17.7.1+.

`nobasic` suppresses the default country attribute variables.

`full` adds all 17 country attributes including geographic coordinates and capital city. See Table 4 for the complete list.

`geo`, `capital`, `latitude`, `longitude` add specific geographic fields without the full set of attributes.

`match(`*varname*`)` merges country attributes into an existing dataset. The variable *varname* must contain World Bank country codes (ISO3 format).

Metadata management

`update query` displays the vintage dates of locally cached indicator and country metadata.

`update check` compares local metadata against the remote repository and reports whether updates are available.

`update all` downloads fresh metadata from the repository, replacing the local cache.

Graph metadata (v17.7.1+)

`linewrap(`*string*`)` wraps metadata text for use in graph titles. The argument specifies which metadata to wrap: `name`, `description`, `note`, `source`, `topic`, or `all`.

maxlength(#) sets the maximum characters per line for wrapped text. The default is 50.

linewrapformat(*string*) controls the output format: `stack` (stacked lines), `newline` (newline-separated), `nlines` (returns line count), `lines` (returns individual lines), or `all` (returns all formats).

2.2 Stored results

`wbopendata` is an r-class command that stores results in `r()`. These stored results are critical for automation: they allow downstream code to programmatically access indicator metadata, construct dynamic graph titles, and build reproducible pipelines without manual intervention.

Indicator codes and variable names. World Bank indicator codes like `SI.POV.DDAY` contain periods, which Stata does not allow in variable names. The command automatically converts indicator codes to Stata-safe variable names by replacing periods with underscores and converting to lowercase: `SI.POV.DDAY` becomes `si_pov_dday`. Both forms are stored: `r(indicator#)` preserves the original API code for documentation and re-querying, while `r(varname#)` provides the Stata variable name for use in analysis commands.

Indexed versus aggregate returns. Results come in two forms. Indexed returns (`r(varname1)`, `r(varname2)`, ...) store metadata for each indicator separately, enabling indicator-specific labeling and citation. Aggregate returns store combined information: `r(indicator)` contains the full semicolon-separated query string as entered, while `r(name)` contains all variable names as a space-separated list suitable for `foreach` loops or variable lists.

For each requested indicator (indexed by # = 1, 2, ...), the command returns:

Table 2: Stored results

Result	Type	Description
Aggregate returns (always)		
`r(indicator)`	local	Full query string (semicolon-separated)
`r(name)`	local	All Stata variable names (space-separated)
Indexed returns (per indicator, always)		
`r(indicator#)`	local	Original API indicator code (e.g., SI.POV.DDAY)
`r(varname#)`	local	Stata-safe variable name (e.g., si_pov_dday)
`r(varlabel#)`	local	Indicator label from API
`r(source#)`	local	Source database identifier
`r(time#)`	local	Time dimension name
`r(sourcecite#)`	local	Clean organization name (when Note is non-empty)
With `year()` option		
`r(year#)`	local	Year or year range requested
With `latest` option		
`r(latest)`	local	Formatted subtitle string for graphs
`r(latest_ncountries)`	local	Number of countries with data
`r(latest_avgyear)`	local	Average year of observations
`r(latest_year)`	local	Maximum year retained
With `linewrap()` option		
`r(name#_stack)`	local	Wrapped name for `title()`
`r(description#_stack)`	local	Wrapped description for captions
`r(note#_stack)`	local	Wrapped methodological notes
`r(source#_stack)`	local	Wrapped source text
`r(topic#_stack)`	local	Wrapped topic name
`r(*#_nlines)`	scalar	Line count for each field
`r(*#_line1)`, ...	local	Individual wrapped lines

These returns enable fully automated workflows: a script can download data, extract `r(name1_stack)` for the graph title, `r(sourcecite1)` for the source note, and `r(latest)` for a coverage subtitle—all without hardcoding any metadata.

3 Examples

Basic data retrieval

The core operation is downloading indicator data. These examples show the two fundamental output shapes: wide format (one row per country, year columns) and long format (one row per country-year).

Basic data download. The simplest use case retrieves a single indicator for all

countries. By default, data are returned in wide format with year-specific columns (`yr1960`, `yr1961`, ...):

```
. wbopendata, indicator(NY.GDP.MKTP.CD) clear linewrap(name note) maxlength(35 7
      0)

    Metadata for indicator NY.GDP.MKTP.CD
-------------------------------------------------------------------------------
    Name: GDP (current US$)
-------------------------------------------------------------------------------
    Collection: 2 World Development Indicators
-------------------------------------------------------------------------------
    Description: Gross domestic product is the total income earned through
    the production of goods and services in an economic territory during an
    accounting period. It can be measured in three different ways: using
    either the expenditure approach, the income approach, or the production
    approach. This indicator is expressed in current prices, meaning no
    adjustment has been made to account for price changes over time. This
    indicator is expressed in United States dollars.
-------------------------------------------------------------------------------
    Note: Country official statistics, National Statistical Organizations
    and or Central Banks;
-------------------------------------------------------------------------------
    Topic(s):  ; 3 Economy and Growth
```

Metadata—indicator name, definition, source, and topic—appears automatically, following the SDMX standard (ISO 17369:2013) that ensures consistent definitions across international statistical organizations.

Multiple indicators in long format. The `long` option reshapes data to one row per country-year—the preferred structure for most Stata routines including `regress`, `tabulate`, `summarize` with `bysort`, and panel estimation commands. Multiple indicators can be requested by separating codes with semicolons:

```
. wbopendata, indicator(SI.POV.DDAY;NY.GDP.PCAP.PP.KD) clear long ///
      linewrap(name note) maxlength(35 70)

    Metadata for indicator SI.POV.DDAY
-------------------------------------------------------------------------------
    Name: Poverty headcount ratio at $3.00 a day (2021 PPP) (% of
    population)
-------------------------------------------------------------------------------
    Collection: 2 World Development Indicators
-------------------------------------------------------------------------------
    Description: Poverty headcount ratio at $3.00 a day is the percentage of
    the population living on less than $3.00 a day at 2021 purchasing power
    adjusted prices. As a result of revisions in PPP exchange rates, poverty
    rates for individual countries cannot be compared with poverty rates
    reported in earlier editions.
-------------------------------------------------------------------------------
    Note: World Bank, Poverty and Inequality Platform. Data are based on
    primary household survey data obtained from government statistical
    agencies and World Bank country departments. Data for high-income
    economies are mostly from the Luxembourg Income Study database. For more
    information and methodology, please see http://pip.worldbank.org., World
    Bank (WB), uri: http://pip.worldbank.org, note: Data are based on
    primary household survey data obtained from government statistical
    agencies and World Bank country departments. Data for high-income
```

```
    economies are mostly from the Luxembourg Income Study database.
-------------------------------------------------------------------------------
    Topic(s): 11 Poverty ; 2 Aid Effectiveness ; 19 Climate Change
-------------------------------------------------------------------------------

    Metadata for indicator NY.GDP.PCAP.PP.KD
-------------------------------------------------------------------------------
    Name: GDP per capita, PPP (constant 2021 international $)
-------------------------------------------------------------------------------
    Collection: 2 World Development Indicators
-------------------------------------------------------------------------------
    Description: This indicator provides values for gross domestic product
    (GDP) expressed in constant international dollars, converted by
    purchasing power parities (PPPs). PPPs account for the different price
    levels across countries and thus PPP-based comparisons of economic
    output are more appropriate for comparing the output of economies and
    the average material well-being of their inhabitants than exchange-rate
    based comparisons.
-------------------------------------------------------------------------------
    Note: .
-------------------------------------------------------------------------------
    Topic(s):
-------------------------------------------------------------------------------

. describe

 Observations:        17,290
    Variables:            13                  5 Jan 2026 14:06
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
countrycode     str3    %9s                   Country Code
countryname     str75   %75s                  Country Name
region          str3    %9s                   Region Code
regionname      str51   %51s                  Region Name
adminregion     str3    %9s                   Administrative Region Code
adminregionname str75   %75s                  Administrative Region Name
incomelevel     str3    %9s                   Income Level Code
incomelevelname str19   %19s                  Income Level Name
lendingtype     str3    %9s                   Lending Type Code
lendingtypename str14   %14s                  Lending Type Name
year            int     %9.0g                 Year
si_pov_dday     float   %9.0g                 SI.POV.DDAY
ny_gdp_pcap_p~d float   %9.0g                 NY.GDP.PCAP.PP.KD
-------------------------------------------------------------------------------
Sorted by: countrycode  year
     Note: Dataset has changed since last saved.
```

Filtering and annotation

Once data are retrieved, several options refine outputs for specific purposes—filtering to the most recent observations or formatting metadata for publication-ready graphs.

Latest available data. The `latest` option keeps only the most recent non-missing observation per country. When multiple indicators are requested, the algorithm retains only observations where *all* indicators have non-missing values in the same year—

prioritizing comparability over recency:

```
. wbopendata, indicator(SI.POV.DDAY) clear long latest ///
      linewrap(name note) maxlength(35 70)

    Metadata for indicator SI.POV.DDAY
-------------------------------------------------------------------------------
    Name: Poverty headcount ratio at $3.00 a day (2021 PPP) (% of
    population)
-------------------------------------------------------------------------------
    Collection: 2 World Development Indicators
-------------------------------------------------------------------------------
    Description: Poverty headcount ratio at $3.00 a day is the percentage of
    the population living on less than $3.00 a day at 2021 purchasing power
    adjusted prices. As a result of revisions in PPP exchange rates, poverty
    rates for individual countries cannot be compared with poverty rates
    reported in earlier editions.
-------------------------------------------------------------------------------
    Note: World Bank, Poverty and Inequality Platform. Data are based on
    primary household survey data obtained from government statistical
    agencies and World Bank country departments. Data for high-income
    economies are mostly from the Luxembourg Income Study database. For more
    information and methodology, please see http://pip.worldbank.org., World
    Bank (WB), uri: http://pip.worldbank.org, note: Data are based on
    primary household survey data obtained from government statistical
    agencies and World Bank country departments. Data for high-income
    economies are mostly from the Luxembourg Income Study database.
-------------------------------------------------------------------------------
    Topic(s): 11 Poverty ; 2 Aid Effectiveness ; 19 Climate Change
-------------------------------------------------------------------------------

. di as text "latest year: "
latest year: 2024

. di as text "countries: "
countries: 186

. di as text "avg year: "
avg year:    2019.8
```

The stored scalars `r(latest)`, `r(latest_ncountries)`, and `r(latest_avgyear)` enable automated construction of figure subtitles that transparently document coverage.

Publication-ready metadata. The `linewrap()` option prepares metadata for graph titles and annotations. The command returns wrapped text in `r(name#_stack)`, `r(description#_stack)`, and source citations in `r(sourcecite#)`:

```
. * Download indicators with linewrap option for graph-ready metadata
. * Returns r(name#_stack), r(description#_stack), r(sourcecite#), r(latest)
. wbopendata, indicator(SI.POV.DDAY; SH.DYN.MORT) clear long latest ///
      linewrap(name description note) maxlength(40 160)

    Metadata for indicator SI.POV.DDAY
-------------------------------------------------------------------------------
    Name: Poverty headcount ratio at $3.00 a day (2021 PPP) (% of
    population)
-------------------------------------------------------------------------------
    Collection: 2 World Development Indicators
-------------------------------------------------------------------------------
```

```
    Description: Poverty headcount ratio at $3.00 a day is the percentage of
    the population living on less than $3.00 a day at 2021 purchasing power
    adjusted prices. As a result of revisions in PPP exchange rates, poverty
    rates for individual countries cannot be compared with poverty rates
    reported in earlier editions.
-------------------------------------------------------------------------------
    Note: World Bank, Poverty and Inequality Platform. Data are based on
    primary household survey data obtained from government statistical
    agencies and World Bank country departments. Data for high-income
    economies are mostly from the Luxembourg Income Study database. For more
    information and methodology, please see http://pip.worldbank.org., World
    Bank (WB), uri: http://pip.worldbank.org, note: Data are based on
    primary household survey data obtained from government statistical
    agencies and World Bank country departments. Data for high-income
    economies are mostly from the Luxembourg Income Study database.
-------------------------------------------------------------------------------
    Topic(s): 11 Poverty ; 2 Aid Effectiveness ; 19 Climate Change
-------------------------------------------------------------------------------

    Metadata for indicator SH.DYN.MORT
-------------------------------------------------------------------------------
    Name: Mortality rate, under-5 (per 1,000 live births)
-------------------------------------------------------------------------------
    Collection: 2 World Development Indicators
-------------------------------------------------------------------------------
    Description: Under-five mortality rate is the probability per 1,000 that
    a newborn baby will die before reaching age five, if subject to
    age-specific mortality rates of the specified year.
-------------------------------------------------------------------------------
    Note: UN Inter-agency Group for Child Mortality Estimation, UN
    Children´s Fund (UNICEF), uri: http://www.childmortality.org, publisher:
    UNICEF, WHO, World Bank, United Nations Population Division;
-------------------------------------------------------------------------------
    Topic(s):
-------------------------------------------------------------------------------

. * Display wrapped metadata available for graph annotations
. return list

macros:
              r(name) : "si_pov_dday  sh_dyn_mort"
       r(latest_year) : "2023"
    r(latest_avgyear) : "   2019.6"
  r(latest_ncountries) : "186"
           r(latest) : "Latest Available Year, 186 countries (avg year    .."
         r(indicator) : "SI.POV.DDAY; SH.DYN.MORT"
             r(time2) : "year"
         r(varlabel2) : "Mortality rate, under-5 (per 1,000 live births)"
           r(source2) : "2 World Development Indicators"
        r(indicator2) : "SH.DYN.MORT"
          r(varname2) : "sh_dyn_mort"
       r(sourcecite2) : "UN Inter-agency Group for Child Mortality Estimati.."
       r(note2_stack) : ""UN Inter-agency Group for Child Mortality Estimat.."
  r(description2_stac
    k)                : ""Under-five mortality rate is the probability per .."
       r(name2_stack) : ""Mortality rate, under-5 (per 1,000" "live births)""
             r(time1) : "year"
         r(varlabel1) : "Poverty headcount ratio at $3.00 a day (2021 PPP) .."
           r(source1) : "2 World Development Indicators"
        r(indicator1) : "SI.POV.DDAY"
```

```
          r(varname1) : "si_pov_dday"
       r(sourcecite1) : "World Bank"
       r(note1_stack) : ""World Bank, Poverty and Inequality Platform. Data.."
r(description1_stac
  k)                  : ""Poverty headcount ratio at $3.00 a day is the per.."
       r(name1_stack) : ""Poverty headcount ratio at $3.00 a day" "(2021 PP.."
```

These stored results can then be used directly in graph commands. Figure 1 demonstrates how the returned metadata populates axis titles, captions, and source notes:

```
. * Graph with wrapped axis titles, subtitle, definitions, and sources
. set scheme sj

. twoway (scatter sh_dyn_mort si_pov_dday, msize(small) mcolor(blue%50)), ///
     xtitle("Poverty headcount ratio at $3.00 a day" "(2021 PPP) (% of population)", size(small)) ///
     ytitle("Mortality rate, under-5 (per 1,000" "live births)", size(small)) ///
     title("Poverty and Child Mortality", size(medium)) ///
     subtitle("Latest Available Year, 186 countries (avg year    2019.6)", size(small)) ///
     caption("bf:Definitions:" ///
     "bf:X-axis: " "Poverty headcount ratio at $3.00 a day is the percentage of the population living on less
     "bf:Y-axis: " "Under-five mortality rate is the probability per 1,000 that a newborn baby will die before
     note("bf:Data Sources:" ///
     "bf:X (Poverty): World Bank" ///
     "bf:Y (Mortality): UN Inter-agency Group for Child Mortality Estimation", size(vsmall)) name(tmp1, replac
```

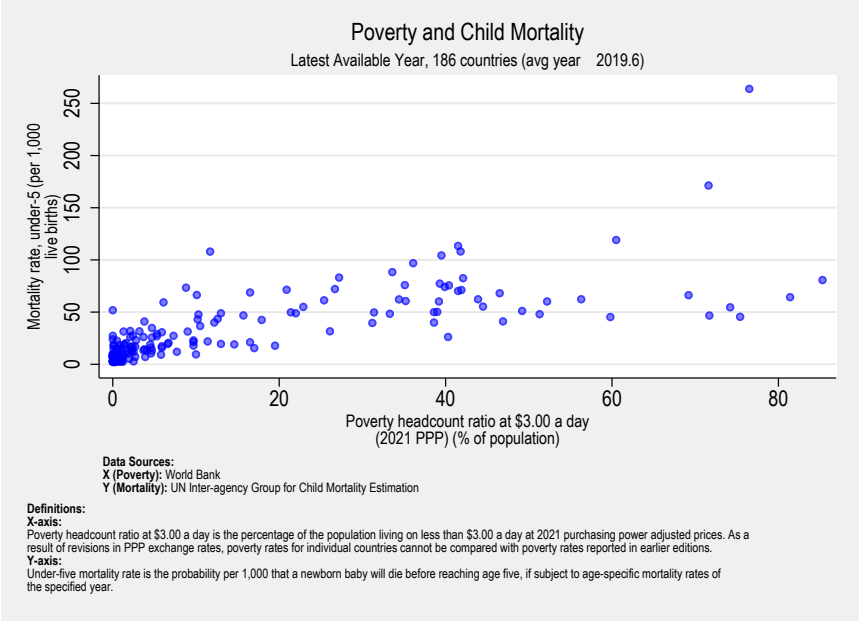

Figure 1: Poverty and child mortality scatter plot with automatic metadata annotation. Axis titles, definitions, and source citations are populated directly from `wbopendata`'s stored results using the `linewrap()` option.

Country attributes and mapping

The `full` option enriches data with 17 country attributes—regional classifications, income levels, and geographic coordinates—enabling merges, stratified analysis, and spatial visualization.

Country attributes. The `full` option attaches all 17 country attributes—regional classifications, income levels, lending types, and geographic coordinates—enabling merges and mapping without separate lookup steps:

```
. wbopendata, indicator(NY.GDP.MKTP.CD) country(BRA) clear full
```

Table 3 shows the country attributes returned for Brazil:

Table 3: Country attributes returned with `full` option

Category	**Code**	**ISO2**	**Name**
Region	LCN	ZJ	Latin America and Caribbean
Income Level	UMC	XT	Upper middle income
Admin Region	LAC	XJ	LAC (excl. high income)
Lending Type	IBD	XF	IBRD
Geographic: Capital: Brasilia, Lat: −15.78, Long: −47.93			

Choropleth mapping. Figure 2 demonstrates how `wbopendata` combines with `spmap` for geographic visualization. The workflow downloads indicator data, merges with shape file coordinates, and renders a choropleth map:

```
. * Download indicator data
. tempfile wdi_data
. wbopendata, indicator(it.cel.sets.p2) long clear latest

    Metadata for indicator IT.CEL.SETS.P2
-------------------------------------------------------------------------------
    Name: Mobile cellular subscriptions (per 100 people)
-------------------------------------------------------------------------------
    Collection: 2 World Development Indicators
-------------------------------------------------------------------------------
    Description: Mobile cellular telephone subscriptions are subscriptions
    to a public mobile telephone service that provide access to the PSTN
    using cellular technology. The indicator includes (and is split into)
    the number of postpaid subscriptions, and the number of active prepaid
    accounts (i.e. that have been used during the last three months). The
    indicator applies to all mobile cellular subscriptions that offer voice
    communications. It excludes subscriptions via data cards or USB modems,
    subscriptions to public mobile data services, private trunked mobile
    radio, telepoint, radio paging and telemetry services.
-------------------------------------------------------------------------------
    Note: World Telecommunication ICT Indicators Database, International
    Telecommunication Union (ITU)
-------------------------------------------------------------------------------
    Topic(s): 9 Infrastructure
-------------------------------------------------------------------------------
```

```
. sort countrycode
. * Merge with shapefile coordinates
. use "C:/Users/jpazevedo/ado/plus/w/world-d.dta", clear
. * Create choropleth map
. set scheme sj
. sum year

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
        year |        262    2022.584    2.400234       2004       2024
. local avg = string(, "%16.1f")
. spmap it_cel_sets_p2 using "C:/Users/jpazevedo/ado/plus/w/world-c.dta", id(_ID) ///
     clnumber(20) fcolor(Reds2) ocolor(none ..) ///
     title("Mobile cellular subscriptions (per 100 people)", size(*1.2)) ///
     legstyle(3) legend(ring(1) position(3)) ///
     note("Source: World Telecommunication ICT Indicators Database (latest: 2022.6)")
```

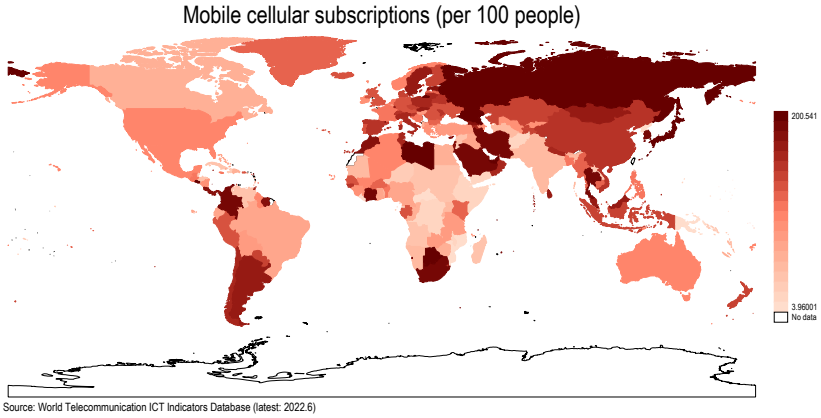

Figure 2: Mobile cellular subscriptions per 100 people (latest available year). Created by merging `wbopendata` output with geographic shape data and visualizing using the `spmap` command (Pisati 2007).

Multi-indicator scatter plot. Figure 3 shows a scatter plot of poverty headcount ratios against GDP per capita, demonstrating multi-indicator downloads with a lowess smoother and labeled regional aggregates:

```
. * Scatter plot: Poverty vs GDP per capita with lowess smoother
. set scheme sj

. graph twoway ///
     (scatter si_pov_dday ny_gdp_pcap_pp_kd, msize(*.3)) ///
     (scatter si_pov_dday ny_gdp_pcap_pp_kd if regionname == "Aggregates", ///
     msize(*.8) mlabel(countryname) mlabsize(*.8) mlabangle(25)) ///
     (lowess si_pov_dday ny_gdp_pcap_pp_kd), ///
     legend(off) ///
     ytitle("Poverty headcount ratio at $2.15 a day", size(small)) ///
     xtitle("GDP per capita, PPP (constant intl $)", size(small)) ///
     note("Source: WDI (latest as of  5 Jan 2026 14:07)")
```

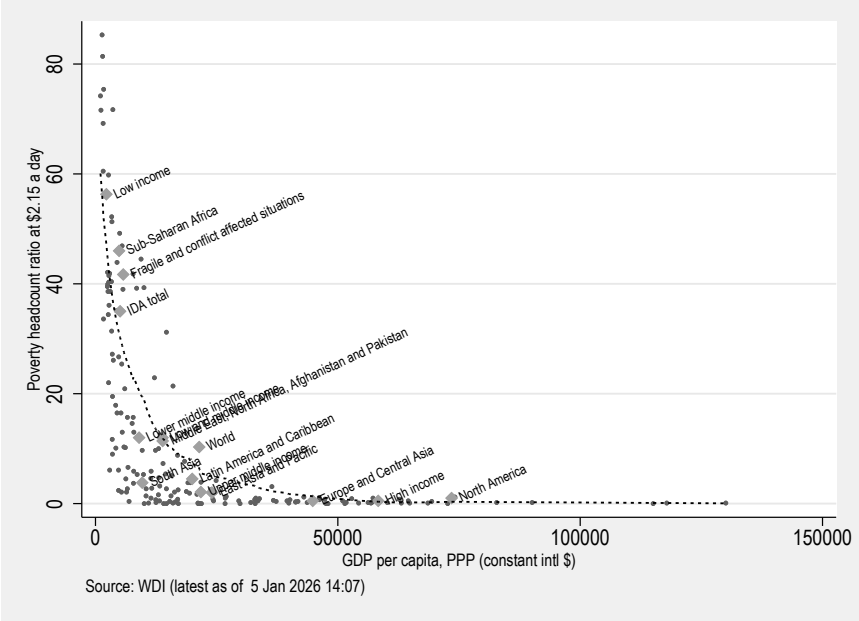

Figure 3: Poverty headcount ratio versus GDP per capita (PPP, constant international $). Regional aggregates are labeled; lowess smoother shows the cross-country relationship.

User-written extensions

The open architecture of `wbopendata` has enabled a small ecosystem of user-written extensions. One notable example is `worldstat` (Clarke 2012), which produces geographic and temporal visualizations of World Bank indicators with minimal user effort. The command retrieves shape files remotely and calls `wbopendata` internally to fetch indicator data.

Regional map. Figure 4 shows GDP per capita across Africa:

```
. * Regional map: GDP per capita in Africa (2009)
. * Options: stat(GDP), year(2009), cname displays country names
. worldstat Africa, stat(GDP) year(2009) cname
worldstat is built using the functionality of the module wbopendata.
checking wbopendata consistency and verifying not already installed...
Accessing shape file for Africa to create geographical visualisation
Accessing shape files for map output remotely
(prefix now "http://damianclarke.net/stata/worldstat")
Importing GDP from World Bank database

    Metadata for indicator NY.GDP.PCAP.KD
--------------------------------------------------------------------------------
    Name: GDP per capita (constant 2015 US$)
--------------------------------------------------------------------------------
    Collection: 2 World Development Indicators
```

```
-------------------------------------------------------------------------------
    Description: Gross domestic product is the total income earned through
    the production of goods and services in an economic territory during an
    accounting period. It can be measured in three different ways: using
    either the expenditure approach, the income approach, or the production
    approach. The core indicator has been divided by the general population
    to achieve a per capita estimate.This indicator is expressed in constant
    prices, meaning the series has been adjusted to account for price
    changes over time. The reference year for this adjustment is 2015. This
    indicator is expressed in United States dollars.
-------------------------------------------------------------------------------
    Note: Country official statistics, National Statistical Organizations
    and or Central Banks;
-------------------------------------------------------------------------------
    Topic(s):  ; 3 Economy and Growth
-------------------------------------------------------------------------------

Visualising data
```

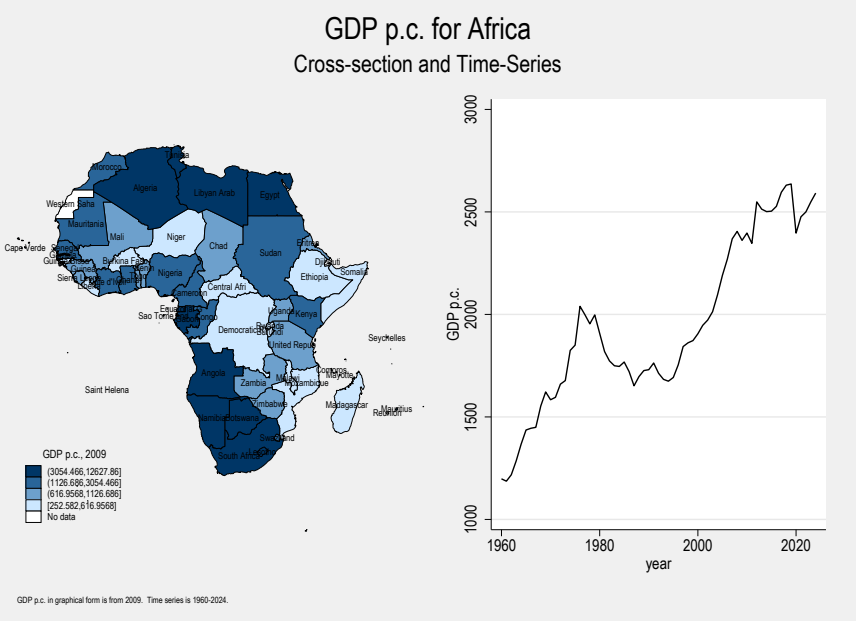

Figure 4: GDP per capita (constant 2015 US$) in Africa, 2009. Generated using `worldstat`, which calls `wbopendata` internally.

Global map. Figure 5 displays global fertility rates with an alternative color scheme:

```
. * Global map: Fertility rate with Pastel2 color scheme
. worldstat world, stat(FERT) fcolor(Pastel2)
worldstat is built using the functionality of the module wbopendata.
checking wbopendata consistency and verifying not already installed...
Accessing shape file for world to create geographical visualisation
```

```
Accessing shape files for map output remotely
(prefix now "http://damianclarke.net/stata/worldstat")
Importing FERT from World Bank database

    Metadata for indicator SP.DYN.TFRT.IN
-------------------------------------------------------------------------------
    Name: Fertility rate, total (births per woman)
-------------------------------------------------------------------------------
    Collection: 2 World Development Indicators
-------------------------------------------------------------------------------
    Description: Total fertility rate represents the number of children that
    would be born to a woman if she were to live to the end of her
    childbearing years and bear children in accordance with age-specific
    fertility rates of the specified year.
-------------------------------------------------------------------------------
    Note: World Population Prospects, United Nations (UN), publisher: UN
    Population Division;
-------------------------------------------------------------------------------
    Topic(s):  ; 8 Health
-------------------------------------------------------------------------------

Visualising data
```

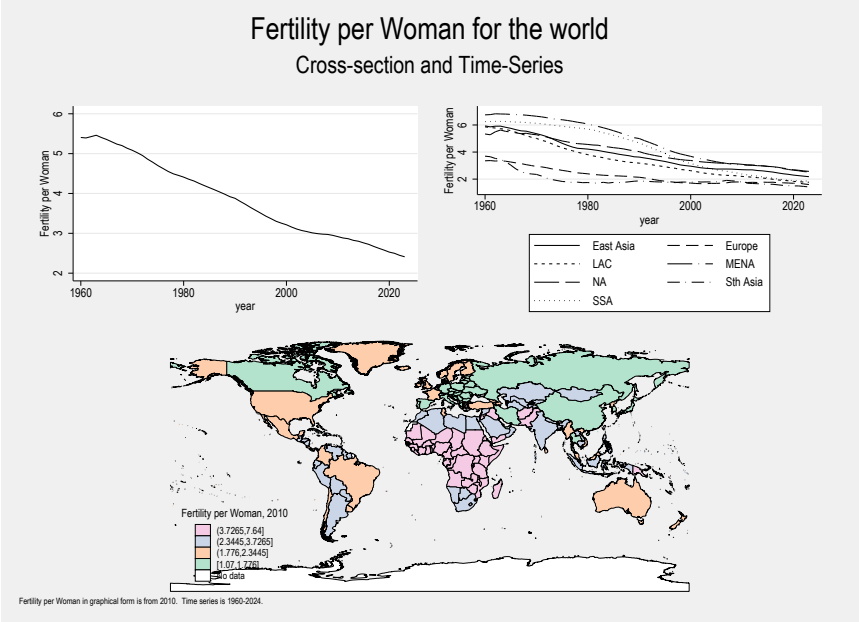

Figure 5: Fertility rate (births per woman) worldwide. The `worldstat` command uses `wbopendata` internally to retrieve World Bank indicators.

Beyond single-command wrappers, `wbopendata` serves as infrastructure for larger analytical pipelines. The World Bank's Learning Poverty repository (World Bank 2019) demonstrates this pattern: a fully reproducible Stata codebase that calculates global

learning poverty indicators—the share of 10-year-olds unable to read and understand a short text—by combining harmonized learning assessment data with enrollment statistics. The pipeline uses `wbopendata` to retrieve population weights, enrollment rates, and other World Development Indicators needed to construct the composite measure. Released under an MIT license with 7 versioned releases (v1.0–v4.0), the repository exemplifies how programmatic data access enables transparent, auditable research workflows that can be independently replicated and extended.

4 Technical implementation

Installation

The recommended installation method is from SSC:

```
. ssc install wbopendata, replace
```

For the latest development version with graph metadata features:

```
. net install wbopendata, ///
     from("https://raw.githubusercontent.com/jpazvd/wbopendata/main/src") replace
```

Architecture

The command is written entirely in Stata's ado-file language and relies on the World Bank API's REST endpoints (World Bank 2024). Key implementation features include automatic pagination (the command retrieves all pages and assembles complete datasets), informative error handling (HTTP status codes surface diagnostics for missing indicators or connectivity issues), caching with hashed request parameters for deterministic reuse, and variable naming that normalizes indicator codes to legal Stata identifiers. String tokenization and metadata line-wrapping adapt community-contributed utilities `tknz` (Elliott 2002), `linewrap` (Over and Azevedo 2000), and `_pecats` (Long and Freese 2001).

The implementation avoids third-party binaries, ensuring compatibility with institutional computing environments that restrict software installation. Users operating behind corporate proxies can configure Stata's HTTP proxy settings (`set httpproxy on`, etc.), which `wbopendata` respects for all API requests.

Country attributes. When using the `full` option, `wbopendata` returns 17 country attributes that enable rich classification and merging. Table 4 summarizes these variables.

Table 4: Country attributes returned with `full` option

Variable	Description
`countrycode` / `countryname`	ISO3 code and name
`region` / `regionname`	Region code and name
`adminregion` / `adminregionname`	Administrative region
`incomelevel` / `incomelevelname`	Income classification
`lendingtype` / `lendingtypename`	Lending type (IBRD, IDA, Blend)
`capital`	Capital city name
`latitude` / `longitude`	Capital coordinates

The `match(varname)` option merges these attributes into an existing dataset containing WDI country codes, enabling enrichment without separate data downloads.

Filtering aggregates. The `region` variable identifies whether an observation is a country or regional aggregate. Researchers can filter using `keep if region != "NA"` (individual countries only) or `keep if regionname == "Aggregates"` (aggregates only).

5 Reproducibility, testing, and error handling

`wbopendata` ships with a live integration test harness designed to validate actual user workflows: indicator downloads, pagination, caching, option handling, and fixes for historical issues. The harness runs end-to-end against the live World Bank API, exercising the same code paths users encounter in practice.

Because Stata lacks the mature testing and CI infrastructure found in R and Python, we chose a self-contained live-API approach rather than mocked unit tests. This design prioritizes detection of upstream regressions—schema changes, pagination failures, and API shifts—that directly impact users.

The test suite comprises 44 integrated tests distributed across 13 categories (Table 5). QA materials in `qa/` include the test driver (`run_tests.do`), protocol documentation (`test_protocol.md`, `TESTING_GUIDE.md`), and timestamped result logs.

Table 5: Test suite composition (44 tests across 13 categories)

Abbr.	Category	Tests	Focus
ENV	Environment	2	Network connectivity, API availability
DL	Download modes	8	Indicator, country, topic, multi-indicator
FMT	Output format	5	Wide, long, reshape, latest
CTRY	Country options	4	basic, full, geo, iso
LW	Linewrap	6	Metadata wrapping, formats
REG	Regression	7	Historical bug fixes
ADV	Advanced	12	Edge cases, error handling

Error handling. When an indicator code is misspelled or does not exist, `wbopendata` returns an informative error message with guidance on verifying the indicator list, testing connectivity, and contacting support:

```
. wbopendata, language(en) indicator(platypus) long clear

    Sorry... No data was downloaded for indicator platypus.

    (1) Please check your internet connection by clicking here, if does not
    work please check with your internet provider or IT support, otherwise...
    (2) Please check your access to the World Bank API by clicking here, if
    does not work please check with your firewall settings or internet
    provider or IT support, otherwise...
    (3) Please check the availability of your indicator or topic by clicking
    here. If the paramater value is not valid...
    (4) Please check the list of available indictator(s) or topic(s) in the
    help wbopendata or by visiting the API query builder, if all the above
    seems fine...
    (5) Please consider ajusting your Stata timeout parameters. For more
    details see netio.
    (6) Please send us an email to report this error by clicking here or
    writing to:
            email: data@worldbank.org
            subject: wbopendata query error at 5 Jan 2026 14:07:20:
    https://api.worldbank.org/v2/en/countries/all/Indicators/platypus?download
    > format=CSV&HREQ=N&filetype=data

. di as text "Captured return code (expected nonzero): "
Captured return code (expected nonzero):
```

Some historical indicators have been moved to the World Bank Database Archives. When requesting a deprecated indicator such as `AG.AGR.TRAC.NO`, the command returns a deprecation notice with the archived source location:

```
. wbopendata, language(en) indicator(AG.AGR.TRAC.NO) clear

    Sorry... but indicator AG.AGR.TRAC.NO has been moved to 57 WDI Database
    Archives.

    Please send us an email to obtain more information clicking here or
    writing to:
            email: data@worldbank.org
            subject: wbopendata query error 23 [AG.AGR.TRAC.NO - Agricultural
    machinery, tractors] at 5 Jan 2026 14:07:20:
    https://api.worldbank.org/v2/Indicators/AG.AGR.TRAC.NO

. di as text "Captured return code (expected r(23) archive notice): "
Captured return code (expected r(23) archive notice):
```

Keeping metadata current. The `update` options manage the local metadata cache. Use `update query` to check the current vintage, `update check` to compare against the remote repository, and `update all` to refresh the local cache:

```
. wbopendata, update query

-------------------------------------------------------------------------------
```

```
Indicators update status

   Existing Number of Indicators:  29323
   Last check for updates:         4 Jan 2026 13:05:24
   New update available:           none      (as of 4 Jan 2026 13:05:24)
   Current update level:           4 Jan 2026 13:05:24

   Country metadata:               296
   Last country check:             4 Jan 2026 13:05:24
   Current country update level:   4 Jan 2026 13:06:46

Possible actions

   Check for available updates   (or type -wbopendata, update check detail -)

   See current documentation on indicators list, Regions,
   Administrative Regions, Income Levels, and Lending Types
```

Regular metadata updates ensure that new indicators, revised country classifications, and corrected definitions are available locally. This is particularly important before using the `match()` option to merge country attributes, as stale metadata may contain outdated income or regional classifications or even country names.

6 Discussion

Fifteen years after its initial release, `wbopendata` remains relevant because of the methodological principles embedded in its design. The core contribution lies not in any single technical feature, but in its role as **infrastructure for reproducible research**. By translating API endpoints into domain-specific commands, it lowers the barrier to scripted data access without requiring researchers to master HTTP protocols, JSON parsing, or pagination logic.

The command exemplifies what can be called **data acquisition as code**: indicator selections, country lists, time ranges, and filters are explicitly parameterized in analysis scripts rather than buried in manual downloads. This design supports reproducibility in two ways: analyses can be replicated exactly at a given point in time, and they can be systematically updated as new data become available.

Addressing the reproducibility crisis. The social sciences have faced a well-documented reproducibility crisis (Baker 2016; Open Science Collaboration 2015), with numerous high-profile failures to replicate published findings. While much attention has focused on statistical methods, pre-registration, and publication bias, a quieter but equally important dimension concerns *data provenance*. When researchers manually download spreadsheets from web portals, rename files, apply undocumented filters, and copy-paste into analysis software, the resulting workflow is effectively unreproducible—not because of statistical error, but because no complete record exists of how the analytical dataset was constructed. `wbopendata` addresses this gap directly. A single command line such as `wbopendata, indicator(SI.POV.DDAY) year(2000:2020) long clear` constitutes a complete, executable specification of data acquisition. Years later, another researcher can run the same command and obtain either identical data

(if the source has not been revised) or systematically updated data (if new observations have been added)—either outcome being preferable to the black box of manual data assembly.

Authoritative sources in the AI era. The rise of large language models has fundamentally changed the landscape of empirical research. AI tools can now generate plausible-looking code, synthetic datasets, and even fabricated citations. In this environment, programmatic access to *authoritative sources of truth*—official statistics maintained by accountable institutions—becomes more important than ever. When a researcher queries the World Bank API through `wbopendata`, the provenance is verifiable: the data come from a specific institutional source with documented methodology, revision history, and contact information for follow-up. This stands in contrast to data scraped from unattributed web sources or generated by statistical models that may hallucinate plausible but incorrect values. Tools like `wbopendata` thus serve as anchors of credibility in an increasingly noisy information environment—not because they prevent errors, but because they make data provenance explicit and auditable.

The broader ecosystem. `wbopendata` is part of a growing but still fragmented ecosystem of tools that translate complex data infrastructures into scriptable access layers. Within the World Bank ecosystem, `datalibweb` (World Bank 2018) provides structured access to harmonized microdata collections, enabling version-controlled retrieval of household survey datasets—though this infrastructure remains largely internal to the Bank, limiting external reproducibility. For Brazilian microdata, `datazoom_social_Stata` (PUC-Rio Department of Economics 2020) demonstrates how national statistical agencies' complex survey structures can be systematically documented and harmonized, though it still requires manual download of original files before processing. More recently, `unicefData` (Azevedo 2024) extends programmatic access to child-related indicators at global scale, providing unified R, Python, and Stata interfaces to UNICEF's SDMX data warehouse.

These efforts share a common pattern: encapsulating data access and preprocessing logic into reusable code, reducing ad-hoc workflows, and increasing transparency. Yet constraints remain. Microdata access continues to rely largely on institution-specific platforms and manual approval processes. No widely adopted open architecture yet exists to provide secure, standardized, API-enabled access to individual-level survey microdata across institutions and countries—a gap that continues to limit reproducibility in applied microeconomic research.

Lessons for open science. The fifteen-year trajectory offers concrete lessons. First, **infrastructure matters as much as principles**—transparency and reproducibility are widely endorsed, yet their implementation often lags due to tooling gaps. Second, **stability enables cumulation**—by maintaining backward compatibility across API changes, the command has built trust that encourages adoption. Third, **domain-specific design trumps generality**—while generic HTTP libraries could access the World Bank API, a command that speaks the language of development economics (indicators, countries, years, topics) reduces cognitive load and makes correct usage intuitive.

Risks and challenges

Provenance opacity. Most indicators accessible through `wbopendata` originate from national statistical offices, UN agencies, and interagency collaborations—not the World Bank itself. The Bank's value-added lies in compilation; the harder work of harmonizing definitions happens upstream. Key indicators depend on governance processes managed entirely outside the Bank: UN IGME produces under-five mortality rates; JME (UNICEF, WHO, World Bank) harmonizes child anthropometric data; JMP tracks water and sanitation access. *Mitigation*: Use `r(sourcecite#)` to identify original producers; cite upstream agencies in publications; consult methodology notes before analysis.

Coverage and timeliness. The `latest` option masks substantial heterogeneity: the "latest" poverty estimate for some countries may be a decade old. Missing values are not random—they correlate with state capacity, conflict, and political sensitivity. Methodology changes (PPP rebasing, poverty line revisions) compound these issues by creating discontinuities in panel data. *Mitigation*: Report the year distribution (mean, SD, frequency table); document coverage in appendices—countries and population, both overall and by region or income-grouping; test sensitivity to stale observations; check methodology notes for series breaks.

Sustainability risk. Open data is not free data. Tool infrastructure costs are trivial compared to what member states spend on household surveys or interagency groups invest in harmonized estimates. Researchers benefit from open APIs but bear none of the production costs—a classic free-rider problem. *Mitigation*: Acknowledge statistical infrastructure in publications; cite data producers (not aggregators); recognize that continued access depends on upstream investment requiring sustained advocacy.

7 Conclusion

Fifteen years after the World Bank Open Data Initiative transformed development statistics into a global public good, `wbopendata` continues to serve as essential infrastructure for reproducible research in Stata. The command provides seamless access to over 29,000 development indicators from 51 databases, handling API communication, pagination, caching, and metadata management within a single stable interface.

This longevity reflects deliberate design choices: maintaining backward compatibility across API changes, encapsulating complexity behind domain-specific parameters, and treating data acquisition as code rather than manual preprocessing. Version 17.7.1 extends this tradition with publication-quality metadata line-wrapping, default country-context attributes, and fixes for multi-indicator handling in the `latest` option.

The command's pure-Stata implementation ensures compatibility with institutional computing environments, while the bundled live test suite provides confidence in cross-platform reliability. Yet `wbopendata`'s most important contribution may be methodological rather than technical: it demonstrates that programmatic data access—with explicit parameters, reproducible workflows, and verifiable provenance—can become routine practice without requiring researchers to become software engineers.

Acknowledgments

The author thanks the World Bank Open Data Initiative for making development data freely accessible, and the many users who have contributed bug reports and feature suggestions through GitHub.

8 References

Azevedo, J. P. 2011. WBOPENDATA: Stata module to access World Bank databases. Statistical Software Components S457234, Boston College Department of Economics. Accessed: 2026-01-05. https://ideas.repec.org/c/boc/bocode/s457234.html.

———. 2024. unicefData: Programmatic access to UNICEF SDMX data warehouse. GitHub repository. Interfaces for R, Python, and Stata. https://github.com/unicef-drp/unicefData.

Baker, M. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533(7604): 452–454.

Clarke, D. 2012. WORLDSTAT: Stata module to produce World Bank visualizations. *Statistical Software Components* (S457565). https://ideas.repec.org/c/boc/bocode/s457540.html.

Elliott, D. C. 2002. TKNZ: Stata module to tokenize string into named macros. *Statistical Software Components* (S426302). Revised 17 Oct 2006. https://ideas.repec.org/c/boc/bocode/s426302.html.

Long, J. S., and J. Freese. 2001. _PECATS: Utility to determine names and values of categories of dependent variable. *Stata Technical Bulletin* 58(sg155). Part of MLOGTEST package. https://www.stata.com/stb/stb58/sg155/_pecats.hlp.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251): aac4716.

Over, M., and J. a. P. Azevedo. 2000. LINEWRAP: Split a long string into shorter strings and, optionally, display them (Version 2.1 4Jun2023) Version 2.1. http://digital.cgdev.org/doc/stata/MO/Misc/linewrap/linewrap.html.

Pisati, M. 2007. SPMAP: Stata module to visualize spatial data. *Statistical Software Components* (S456812). https://ideas.repec.org/c/boc/bocode/s456812.html.

PUC-Rio Department of Economics. 2020. datazoom_social: Stata package for Brazilian household surveys. GitHub repository. Accessed: 2026-01-05. https://github.com/datazoompuc/datazoom_social_Stata.

World Bank. 2010. World Bank Open Data Initiative. Press release. Launched April 2010. https://www.worldbank.org/en/news/press-release/2010/04/20/world-bank-group-opens-data-to-all.

———. 2018. datalibweb: Stata package for accessing harmonized microdata. Internal World Bank infrastructure. Provides version-controlled retrieval of household survey datasets. https://github.com/worldbank/datalibweb.

———. 2019. Learning Poverty: Reproducible Stata codebase for global learning poverty indicators. GitHub repository. MIT License, versions 1.0–4.0. https://github.com/worldbank/LearningPoverty.

———. 2024. World Bank API Documentation. Developer documentation. https://datahelpdesk.worldbank.org/knowledgebase/topics/125589-developer-information.

About the authors

João Pedro Azevedo is Deputy Director and Chief Statistician in UNICEF's Division of Data, Analytics, Planning and Monitoring. Previously, he worked for 16 years at the World Bank as Lead Economist. His research focuses on poverty measurement, education statistics, and reproducible and scalable analytical pipelines for global, regional, and national monitoring systems to inform policy.

The software repository (https://github.com/jpazvd/wbopendata) includes example do-files, test scripts, and complete documentation.