FS 2.1: 2023 Expansion of the OLAS Household Survey Data Set

The Expansion of the OLAS Household Survey Data Set was a project implemented in collaboration between the OLAS and the Social Division (SCL) of the Inter-American Development Bank (IADB) from December 2022-July 2023. The project's objectives were multi-fold, with a focus on improving the information available on water and sanitation access while simultaneously streamlining the data generation process.

Antecedents

Before this project, there were two regional data sets produced from national household surveys: SLC's Household Socio-Economic Surveys Data Set and the OLAS Household Survey Data Set.

The SCL's Household Socio-Economic Surveys Data Set includes information on topics such as education, labor, income, education, and living conditions; however, prior to this project the data set incorporated few indicators on water and sanitation, most of which used outdated definitions and terms. Conversely, the OLAS Household Survey Data Set included many valuable indicators with sector-relevant definitions but lacked many of the benefits of the SLC Household Survey Data Set such as its large temporal coverage, variety of breakdown dimensions, and established resources and data processes for updating the data set. Merging the data sets offered clear benefits and eliminated duplicity.

Table 1: Characteristics of both data sets, showing the advantage of merging the projects. (Author's own elaboration)

OLAS Household Survey Data Set:		SCL's Household Socio-Economic Surveys Data Set	
✅	Based on microdata produced by countries	✅	Based on microdata produced by countries
⬛	Updates were slow due to lack of human and technological resources	✅	Update process already established
⬛	Did not allow for temporal analysis (only had 1 or 2 years of data per country, with years spanning from 2014-2020)	✅	Allowed for temporal analysis (data from 2003-2022 with an average of 13 years per country)
⬛	2 breakdown dimensiones (zone and income quintile)	✅	8 breakdown dimensiones (zone, quintile, sex, age, ethnicity, migratory status, disability status, and education level)
✅	32 water and sanitation indicators	⬛	4 water and sanitation indicators
✅	Water and sanitation indicators based on Joint Monitoring Programme's access framework	⬛	Water and sanitation indicators used outdated definitions (MDG definitions)

Execution

The project was carried out from January 2023 – July 2023 by a small team comprised of members of the SCL Data team and the OLAS. Over three hundred surveys and their associated data sets and questionnaires were evaluated and harmonized according to the desired water and sanitation indicator definitions, covering 23 countries from 2003-2022. These harmonized datasets were then used to generate indicators at the country-level and broken down by several socioeconomic dimensions. This effort was divided into two sprints, the first covering surveys from 2013-2022 and the second harmonizing information from 2003-2012, after which five rounds of validation were performed, allowing for iterative adjustments, and corrections. Results of the project were presented to the OLAS country focal points on August 22, 2023.

Methodology

The process of creating the data set involves three steps:

- Harmonization the microdata of each survey,
- Generation of indicators from the harmonized data, and
- Unification, cleaning, and validation of the final data set.

Of these three steps, the harmonization of the microdata is the most vulnerable because it determines what values and fields in the

Figure 1: Rules for harmonizing information on drinking water sources. The values of this intermediate variable were then used to generate the final indicators.

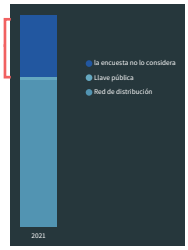
Example of the harmonization of variables

Definition of an intermediate variable

aguafconsumo_ch	**Description:** Water source used by the household for human consumption **Rules:** If the survey does not differentiate between drinking water and water for general use, assign a value of 0. If the categorization is not clear (e.g. "well" instead of "protected well") assign a value of 10.	0	Survey does not ask about drinking water or specify the sources' potability.
		1	Water distribution network, with at-home tap.
		2	Standpipe or other public tap outside of home
		3	Bottled water
		4	Protected well
		5	Rainwater
		6	Delivered or trucked water
		7	Other improved source
		8	Surface waterbody
		9	Other unimproved source
		10	Well, spring, or other source without clear classification

Final indicators

aguafconnulo_ch



aguafredcon_ch

OLAS WATER AND SANITATION OBSERVATORY FOR LATIN AMERICA AND THE CARIBBEAN

IDB

Contact: Inter-American Development Bank - Water and Sanitation Division
Water and Sanitation Observatory for Latin America and the Caribbean

Authors: Jesse Madden Libra and María Alejandra Baquero

olasdata@iadb.org

original surveys will contribute to each indicator. The process consists of creating new fields in each survey's micro data set that represent the same information across all countries and years. Harmonizing large numbers of different survey questions and responses presents a significant challenge due to the heterogeneity of the surveys across the region. So the harmonization variables were designed to allow for the categorization of ambiguous data and data gaps.

Figure 1 provides an example of the harmonization categories for water sources used for human consumption. The creation of a value (0) representing individuals who were not asked about the topic as well as categories 7, 9 and 10 for sources that do not fall clearly into the more specific categories, allows for any response option available on a national survey to be categorized. Each of these classifications is in turn used to create a final indicator in the indicator generation process.

The harmonization process involved reviewing more than 300 surveys, questionnaires, and micro data sets. Once all data was harmonized, the indicator generation process, which is common to all datasets, was run and the indicator data files were generated. These files were then compiled, cleaned, and analyzed to identify potential issues.

Detailed information on the three stages of data generation can be found in the supporting documents and associated code repositories:

Harmonization guide for intermediate variables

Harmonization process GitHub

Indicator creation GitHub

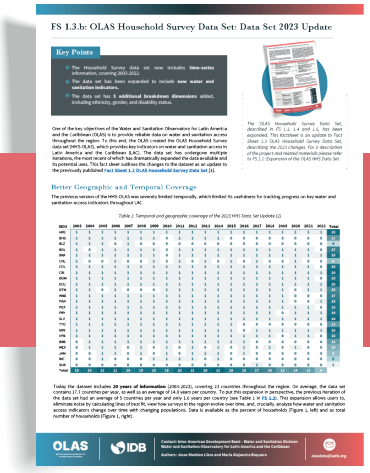
Data dictionary for the SCL Indicators

Cleaning process GitHub

Data and methodology document of the OLAS Household Survey data set

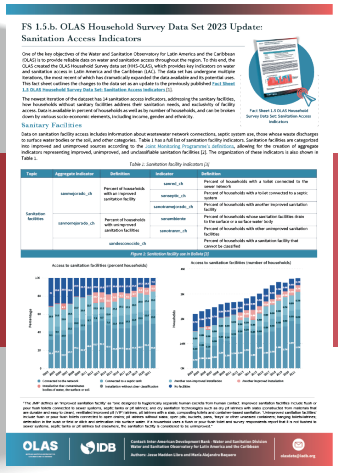
Results

The resulting data set contains 47 indicators available over 20 years for 23 countries in LAC, which can in turn be analyzed through several lenses such as age, income, zone (urban/rural), ethnicity, and more. This expansive data set allows for in-depth analysis of water and sanitation access with respect to socioeconomic inequality, which in turn can aid policy makers in addressing accessibility gaps. The data set as well as dashboards that allow for easy analysis of the data are available on the OLAS. A summary of the information in the new dataset can be found in:



FS 1.3.b: OLAS Household Survey Data Set: Data Set 2023 Update

FS 1.4.b. OLAS Household Survey Data Set 2023 Update: Water Access Indicators



FS 1.5.b. OLAS Household Survey Data Set 2023 Update: Sanitation Access Indicators