Completing the Puzzle: How Increased Data Sharing and Use Can Fill In the Gaps and Better Enable Evidence-Based Decision-Making

Katy Sill, Program Director, Water Point Data Exchange (WPDx)

Pieces of the Puzzle

Despite progress, the world is not on track to reach Sustainable Development Goal 6 (SDG6), ensuring universal access to water and sanitation services. Basic access for people in rural locations continues to lag urban areas, with over four times more people in rural areas (630 million) lacking access compared to urban (143 million).[1] Rural water provision is especially challenging due to the nature of providing services to decentralized, remote, and low-income populations, particularly while adhering to the principle of "leave no one behind." Simply installing more water points is not necessarily a feasible or effective solution, as less than 15% of countries report having adequate finance to meet national rural drinking water targets[2], and up to 25% of water points fail within the first 4 years.[3] The idea of harnessing data to help prioritize investments and optimize limited resources has (re)gained attention

in recent years alongside the rise in popularity of big data and artificial intelligence. Increased data collection and evidence-based decision-making have been identified by many key entities in the sector, including the United Nations, Sanitation and Water for All, the African Minister's Council on Water, international development banks, international aid agencies, nongovernmental organizations (NGOs), and multi-partner consortiums, as key approaches needed to optimize investments and maximize impact, in terms of people reached and the sustainability of services. However, despite these bold policies and commitments, there is not yet a global, robust, widely accessible data resource for rural water points.

Governments, donors, NGOs, and academic institutions have continued to increase monitoring and evaluation efforts, resulting in an unparalleled amount

---

1 United Nations and World Health Programme Joint Monitoring Program. 2019. Progress on household drinking water, sanitation and hygiene 2000-2017: Special focus on inequalities. Available online: https://washdata.org/sites/default/files/documents/reports/2019-07/jmp-2019-wash-households.pdf

2 UN-Water global analysis and assessment of sanitation and drinking-water (GLAAS). 2019. National systems to support drinking-water, sanitation and hygiene: global status report 2019. Geneva: World Health Organization; 2019. Licence: CC BY-NC-SA 3.0 IGO.

3 Banks, Brian, and Sean Furey. 2016. "What's Working, Where, and for How Long: A 2016 Water Point Update." Poster presented at the 7th RWSN (Rural Water Supply Network) Forum, Abidjan, Côte d'Ivoire, November 29. https://www.researchgate.net/publication/312027512_What's_Working_Where_and_for_How_Long_A_2016_Water_Point_Update.

of data, which could and should be used to better understand the challenges and opportunities within the water sector. However, this data is collected at irregular frequencies, is of varying quality, and for the most part, is not widely shared, meaning that the data cannot easily be used to improve decisions outside of the organization that completed the collection effort. This fragmentation of data is true throughout the WASH sector, but especially prevalent for rural water points, health care facilities and schools.

Government-led national inventories typically offer the most comprehensive perspectives. However, due to high costs and extensive logistical challenges, these intensive surveys are not conducted on an annual basis, meaning that the collected data is quickly out of date, and inventory data is often not shared publicly. NGOs and private operators are key stakeholders and contributors to providing rural water services, and regularly visit water points. However, the average nongovernmental water point inventory contains fewer than 5,000 records[4] representing only a fraction of the conditions in a region. While NGOs are increasingly collaborating with other organizations and finding ways to share data with partners, unless data is shared more widely, each entity or even consortium has only a piece, or at best a few pieces of a much larger puzzle. Additionally, different organizations often define data parameters and metrics in ways that are specific to their data collection goals, which makes integration of data especially challenging. For example, when describing water point functionality, there can be a range of what may be described as functional, partially functional, non-functional, abandoned, etc., although there is no commonly held definition of what these individual terms mean. The varying goals of these collection efforts also lead to a difference in the number and type of parameters that are being collected. Research efforts and evaluations may include many parameters, including those which may be very focused on specific

measures of interest that are unique to the context but may not be appropriate to be collected at scale.

Further complicating this existing challenge, the ongoing global COVID-19 pandemic has greatly limited data collection efforts due to travel and gathering restrictions in place around the world. These limitations underscore the need for expansive collaboration and data sharing to ensure that all stakeholders have access to any and all available data.

Bringing the Puzzle Pieces Together: an open platform for data sharing, access, and analysis

In 2014, in recognition of the need for a platform to openly share and access rural public water point data, and in collaboration with sector experts, the Global Water Challenge created the Water Point Data Exchange (WPDx). The first step in the process was to create a common data standard to allow for harmonization of disparate datasets, regardless of the collection methods used or the purpose of the collection effort. The WPDx Data Standard was collaboratively developed by a working group of sector leaders and open for public comment and review, to ensure sector contributions and buy-in. The parameters included in the data standard were identified through an extensive desktop review of more than 40 existing monitoring frameworks, including over 2,500 individual indicators. The indicators most commonly collected across frameworks were selected for the WPDx Data Standard to ensure that exchanging data would be feasible and require minimal effort.[5] The WPDx data standard is a living standard, designed to grow and evolve with the sector. At the core of the standard are six required parameters, which provide basic information about the water point: GPS location (longitude and latitude), functional status (working/non-working), water

---

4 Based on analysis of datasets shared with WPDx

5 WPDx (2014). WASH Data Sharing Update. Available online at:
   https://www.waterpointdata.org/wp-content/uploads/2020/11/wash_datapoint_update_september_2014_compiled_with_appendices.pdf

Table 1.

Summary of WPDx Data Standard Parameters

Required Parameters	Optional Parameters
Latitude	Country
Longitude	Primary Administrative Division
Presence of Water When Assessed	Secondary Administrative Division
Water Source and/or Water Technology	Tertiary Administrative Division
Date of Data Inventory	Water Point ID
Data Source	Scheme ID
	Installation Year
	Installer
	Rehabilitation Year
	Rehabilitator
	Management Structure
	Payment for Water
	Condition
	Public Data Source URL
	Photograph URL
	Fecal Coliform Presence
	Fecal Coliform Value
	Subjective Quality
	Notes

point type (source and/or technology), date of visit, and name of organization providing the data. The remaining parameters provide additional details about the water point's administrative division, management structure, water quality, etc. A summary list can be found in Table 1, and the full standard with parameter definitions is available on the *WPDx website*. (See complete list in Table 1).

The next step was to launch an online platform to enable entities to openly share and access rural water point data aligned to the WPDx Data Standard. The online repository is free and open to all users and does not require a login or organizational membership to access the data. In fall 2020, WPDx launched an overhauled version of its data sharing process, with the goal of reducing the burden on entities seeking to share data. The ingestion engine accepts a variety of common file formats (.csv, .xls, .xlsx, JSON, Google sheets) and then guides organizations to map their

individual datasets to the parameters included in the WPDx Data Standard. Mapped columns are then uploaded and shared to the data repository. As of May 2021, the data repository is home to over 600,000 raw water point records, including data from over 70 countries and almost 100 contributing organizations. By the end of summer 2021, WPDx will host two datasets: WPDx-Basic and WPDx-Plus. The WPDx-Basic dataset will include all records shared by contributors that meet baseline upload requirements, along with the addition of cleaned and categorized versions of the water source, water tech, and management parameters; and versions of the country and administrative division parameters based on GPS locations and the Database of Global Administrative Areas (GADM). The cleaning terms for each parameter can be found on the WPDx website. The WPDx-Plus dataset will offer further enhancements including processes to identify and eliminate duplicate records, match records from different dates with contributors

Table 2.
Summary of features for WPDx-Basic and WPDx-Plus

WPDx-Basic	WPDx-Plus
Full suite of WPDx data standard parameters	Full suite of WPDx-Basic parameters
Cleaned/categorized version of water source entries	Identification and deletion of duplicate records
Cleaned/categorized version of water technology entries	Assignment of WPDx__id to match records for the same water point from different dates and contributors
Cleaned/categorized version of water point management type entries	Addition of parameters, including distance to road, town, city, and land use cover type.
Country name from GADM based on provided GPS location	
Administrative Division 1 (adm1) name from GADM based on provided GPS location	
Administrative Division 2 (adm2) name from GADM based on provided GPS location	
Administrative Division 3 (adm3) name from GADM based on provided GPS location	

for the same water point through the assignment of a unique identifier (known as the WPDx__id), and add relevant external data, such as the proximity to the nearest road and small town for more in-depth analysis (See Table 2).

Amassing water point information is a key step, but a repository of data alone is not enough to improve decisions in the rural water sector. As Fiorina once said, "The goal is to turn data into information and information into insight." The transformation

from data to insights requires well-defined use cases combined with expertise in the fields of data management and data science. In recognition of existing sectoral capacity gaps in data expertise, the WPDx platform developed a suite of decision-support tools to enable government and NGO decision-makers to utilize shared data to improve rural water services by answering four key questions (see Table 3). The tools employ cutting- edge geospatial analysis and advanced machine learning algorithms but are freely available and simple to use. The tools include:

Table 3.
WPDx Decision Support Tool Overview

WPDx Decision Support Tool	Key Question	Outputs from Tool*
Measure Water Access by District	What is the rural drinking water coverage at the district scale?	Calculates the percentage of people currently unserved at the district scale. Allows for comparison and prioritization of investments across districts (Figure 1).
Prioritization of Locations for Rehabilitation	Given limited resources, which non-functional water points should be given priority for rehabilitation?	Identifies the most impactful water points for repair and rehabilitation based on the potential number of people currently unserved by an existing, functional water point (Figure 2).
Prioritization of Locations for New Construction	Which populations are not yet served by an existing water point?	Determines the most impactful locations for new water point installation based on the potential number of people currently unserved by an existing, functional water point. (Figure 3)
Prediction of Water Point Status	Which water points are most likely to break down? Why?	Utilizes machine learning to predict water point status as of today, one year from today, and three years from today. Detects factors that drive water point status. (Figure 4)

*The output from each tool includes a map visualization including a satellite view and a downloadable CSV of the results.

Figure 1.
Sample output from Measure Access by District tool. For Sierra Leone, the results are shown at the sub-district or chiefdom level.
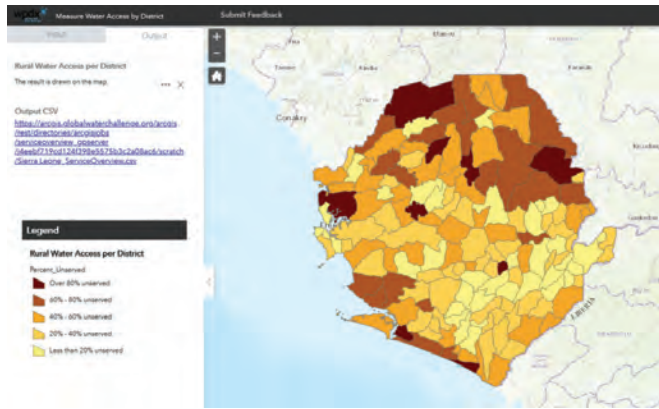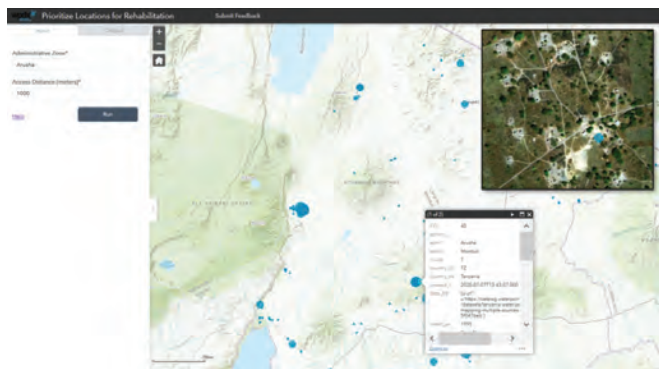

Figure 2.
Sample output from Rehabilitation Priority tool. Larger circles indicate greater number of unserved people who could be reached by water point repair.

Figure 3.
Sample result from Prioritization of Locations for New Construction tool. An "X" marks the spot for a potential location of a new installation based on unserved population data.
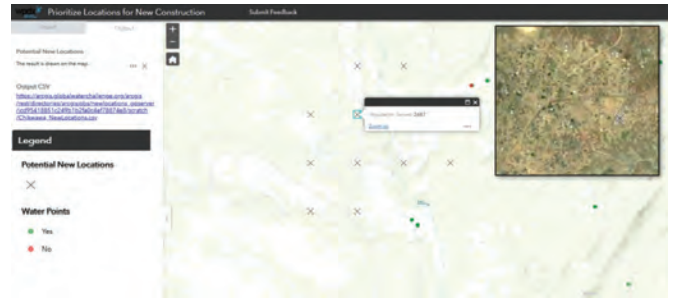

Figure 4.
Output from Water Point Status Tool. Results show map at national, regional, and/or district scale. Pop-ups provide details on water point, including photograph if available. Tabular version of output available for download as a CSV.

In locations with national or district level coverage, the results from WPDx tools offer key insights, which can be used by decision-makers to inform where to prioritize investments for maximum impact. The results from WPDx analyses have been well received by government and NGO partners.

In Sierra Leone, the Ministry of Water Resources (MOWR) launched the National Digital Monitoring Approach, which included the development of national survey protocols and a requirement that all data be shared to the Water Point Data Exchange's global data repository. This harmonized fragmented data and allowed many different sources to be stitched together, creating the most complete picture of water services in Sierra Leone's history. Building on this, the Ministry engaged a local university to host a training for all district officials on how to use the WPDx decision support tools. Finally, MOWR issued a policy directive to the Ministry of Finance instructing that "All selection of location for construction and rehabilitation of water points should be derived from the Water Point Data Exchange." Today, district officials are using WPDx tools to identify priority

locations for rehabilitation and new construction to reach more of the unserved population, harnessing evidence to improve the way water decisions are made for millions of people in Sierra Leone.

A retrospective analysis at the district scale in Sierra Leone showed that at the national scale, almost four times as many people could have been reached with water services had the WPDx Prioritization of Locations for Rehabilitation tool been utilized (Figure 5), and costs could have been reduced by a third.[6]

In addition to Sierra Leone, WPDx trainings have been held in eSwatini, Ethiopia, Liberia, and Uganda, and active project support is ongoing in Ethiopia, Ghana, and Uganda. Similar work to develop a new data exchange platform for WASH in Health Care Facilities is in the early stages of development.

Figure 5.
SEQ Figure * ARABIC. Sample output from retrospective analysis for Tonkolili, Sierra Leone
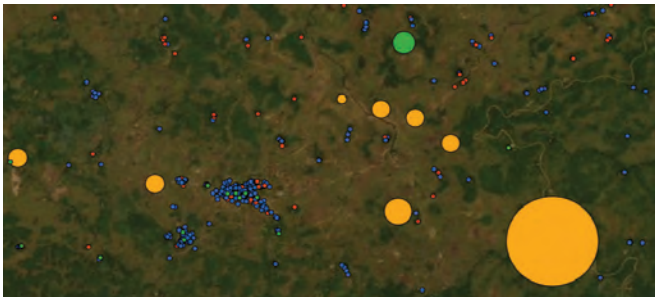
The majority of water investments made in 2012 were located in areas already served by existing functional water points, resulting in fewer unserved people being reached and higher costs-per-person served compared to the recommendations from WPDx decision-support tools.

What makes WPDx unique?

WPDx is a platform for sharing, accessing, and analyzing rural water point data. WPDx is agnostic to the collection platform utilized and can accept data from spreadsheet programs, including Excel and Google, as well as commonly used data collection platforms. WPDx is focused on demonstrating the power of open data and stands apart from other data repositories in the following ways:

- Targeted focus on rural water points

- Governed by a collaboratively designed data standard, allowing for the harmonization of disparate datasets

- Agnostic to the approach and platform used to collect data

- Open, accessible, and downloadable by all, regardless of organizational affiliation

- Easily updated data repository through a simple, user-friendly interface

- Matches records from the same water point, regardless of the organization that provided the data, making it more feasible to collect time-series data for individual water points

- Provides decision-support tools utilizing the WPDx-Plus dataset to give insights to improve four specific key decisions

Filling in the Missing Pieces: overcoming the hurdles to widespread data sharing

As true for all models and tools, WPDx decision-support tools require high-quality, up-to-date inputs to provide the most useful results. While the WPDx database has grown considerably over the years, it still represents only a fraction of the rural water points relied on by millions for their drinking water (see call-out box). There are a number of reasons cited by

---

6   WPDx white paper. Working to publish via blog or other medium ahead of this paper being shared publicly.

organizations regarding why they do not share data publicly, including:

- **Data ownership.** Unclear determinations on whether an NGO or a host-country government owns data on water points, and if data sharing would be acceptable by key stakeholders.

- **Reputational risks.** Concerns about the consequences associated with data transparency related to partnerships with governments and funding organizations, especially if data shows high amounts of non-functional points.

- **Security matters.** Worries about whether data should be made public, or if this information should be protected and kept private in part due to potential threats related to domestic and international terrorism.

- **Issues with data quality.** Recognition that inconsistent methods may have been utilized during data collection leading to issues with completeness and accuracy of collected data.

- **Unclear value proposition.** Questions about why organizations should make the effort to share data without assurance of a specific direct benefit for contributors.

- **Academic publications.** Requirement/preference to keep data private when conducting analysis and writing papers for publication in academic journals.

- **Non-digital data.** While more and more organizations are using mobile applications to collect water point monitoring data, many still rely on analog methods, meaning that records are in notebooks and not easily digitized for sharing without considerable effort.

How many pieces of the rural water point puzzle are there?

Based on JMP statistics from 2017, 2.2 billion people do not have access to safely managed drinking water services.

- People are instead reliant on water points and/or unimproved sources:

  - 1.4 billion people with basic services, meaning an improved water source located within a round trip of 30 minutes

  - 206 million people with limited services, or an improved water source requiring more than 30 minutes to collect water

  - 435 million people taking water from unprotected wells and springs

  - 144 million people collecting untreated surface water from lakes, ponds, rivers, and streams.

- These figures give rise to estimates that between 3.2 million and 6.4 million water points are needed to serve people with existing basic and limited services (1.6 billion people, tapstands, and boreholes with handpumps that serve a maximum of 250–500 people each).

The concerns raised are valid and represent important issues for the sector to discuss and resolve. However, to actualize the policies and fulfill the commitments around improving evidence-based decision-making described in the introduction, the sector at-large must first find a way to overcome the barriers to widespread data sharing. There are several potential approaches that could help open the data flood gates:

- **Governments encourage and/or require data sharing with chosen platform for NGOs working in their countries.** In countries where governments have provided letters of support, WPDx has seen considerable increases in data sharing both from government and NGO sources. For example, in the past six months, the amount of data from Ethiopia increased 10-fold following the issuance of a letter

of support from the Water Development Commission (WDC) in the Ministry of Water, Irrigation and Energy (MoWIE) and included contributions from woreda (district) water offices, NGOs, and consortiums.

- **Donors/funders require data sharing with chosen platform from all WASH grantees**. An increasing number of international donors and private funders have increased their focus on data collection and analysis to track and improve the programs they support. Requiring grantees to openly share data would be a clear demonstration of donor/funder commitment to transparency and accountability. Since 2017, the Coca-Cola RAIN program has required all grantees to upload data on their new water points, which has resulted in uploads from over 15 organizations.

- **Sector consortiums require data sharing through chosen platform for members.** The membership lists of highly respected consortiums, such as the Millennium Water Alliance (MWA) and Agenda for Change, include most leading NGOs. These types of organizations have the power to help institutionalize data sharing as a norm within the rural water sector. In just nine months, a partnership between MWA and WPDx has resulted in over 10,000 new data records being uploaded by six members.

- **Increased demonstrations of the power of open data creates organic incentive for data sharing.** In a limited number of countries where WPDx has national level coverage, decision-support tools can be used to showcase how shared data resources can improve decisions. A retrospective analysis in Sierra Leone showed that on average, using the WPDx repair priority tool would have resulted in at least a 3-fold increase in the number of unserved populations reached. New analyses on the WPDx platform and beyond could provide additional examples of what can be done with a robust open dataset for governments, NGOs, and other stakeholders.

- **Incentives and rewards for data sharing.** Establish sector, funder, national and/or local awards for sharing data and demonstrating commitments to open and transparent information provision. Awards could include opportunities for funding as well as recognition for contributions, and should be widely publicized, for example with a ceremony at high profile events such as World Water Week.

- **Private sector use cases.** In the process of working towards normalizing data sharing, the sector should convene to discuss and resolve the barriers to decision-making mentioned previously. There is no one size fits all solution, but many of these issues can be addressed if the greater sector comes to agreement that there is value in the open sharing of data.

Data Sharing for the Public Good

The WPDx platform was conceived as a public good, a platform for providing open access to monitoring data that was otherwise only available within different organizations. Through the creation of a data standard, data from almost 100 entities has been harmonized and enhanced through data processing and cleaning to create an analysis-ready data set at global, national, and local scales. However, WPDx relies on data sharing from governments, NGOs, academics and others to share data they have collected for a variety of purposes. For WPDx to best serve the public good, there must be broad agreement from stakeholders that open data sharing should be normalized within the sector to help provide transparency, track accountability, and prioritize resources.

Making Sense of the Picture

A robust, harmonized, and regularly updated water point dataset with strong sub-national representation for the majority of countries still working towards reaching SDG6 would provide governments, donors, and NGOs with an unprecedented "near real-time"

understanding of the sector. This consolidated dataset would provide a powerful tool for analyses that could support improved legislation, focused budgeting, optimized work planning, increased accountability, and more holistic evaluation. Instead of basing decisions on outdated national surveys or small-scale non-representative district or sub-district data, all stakeholders could access a common resource to actively discuss challenges and opportunities, develop plans to partner on reaching unserved populations and maintaining sustainable services, and be able to continually contribute to and access the most up-to-date record of what is going on. The possibilities are endless, but an immediate list of potential analyses includes:

- Tracking sub-national progress towards SDGs for rural water and/or WASH at health care facilities and/or schools

- Identifying districts and sub-districts for priority investment to reach the most vulnerable populations at a regional and/or international scale

- Prioritizing locations for rehabilitation and new construction to democratize investments and reach unserved people

- Serving as market research for private operators to identify potential areas for focus

- Enabling improvements to the supply chain

- Determining which technologies and/or management approaches are best suited for a particular location/use case

- Charting trends in service level evolution from water points to small-scale piped systems

- Holding stakeholders accountable for improving and maintaining sustainable services

- Allowing deeper explorations of which factors lead to water point and WASH service sustainability and/or failure

- Combining with groundwater data for a more holistic understanding of resource sustainability

- Defining protocols for minimum standards in data collection

It is key to note that even the best and most cutting-edge analysis is meaningless if the findings are not utilized by decision-makers to improve operations and increase impact. Data analysis and decision-support tools must be developed and refined in collaboration with decision-makers from all stakeholder groups. A number of organizations are working to develop and test approaches for the iteration and integration of data-driven analysis into decision-making processes to improve decisions around budgeting and work planning.[7,8] Challenges abound in this key step towards evidence-based decisions, as decision-makers must overcome substantial political, bureaucratic, and financial hurdles to change existing traditional decision-making processes. A potential pathway to success is to "Align, Harness, Analyze, Integrate, and Accelerate."

- **Align** on Approach

  - Collaborate with decision-makers and stakeholders to agree on common goals

  - What are the key questions?

  - What data is needed to answer these questions?

  - How and by whom will the results be utilized to generate change?

  - Define the data use cases

  - Governments (ministry of water, district water officers, regulatory agencies, etc.)

  - NGOs

  - Donors

  - Communities

7   IRC. 2020. The journey so far: Working together for safe water in Uganda. Available online:
    https://www.ircwash.org/sites/default/files/working_together_for_safe_water_in_uganda_the_journey_so_far_0.pdf

8   WaterAid. 2020. From data to decisions: data use planning guide. Available online:
    https://washmatters.wateraid.org/sites/g/files/jkxoof256/files/from-data-to-decisions-data-use-planning-guide.pdf

- **Harness** Data
  - Conduct data landscape assessments and identify key parameters of interest
  - Harmonize existing government data sources
  - Work with governments to request NGOs to share data
  - Provide training on data sharing
  - Create robust national and/or district level datasets with contributions from all stakeholders
- **Analyze** for Action
  - Support stakeholders to carry out and interpret results from decision-support tools
  - Collaborate in workshops to brainstorm how results can best improve ongoing or future rural water programs
- **Integrate** evidence into decision-making process
  - Use results to inform budgets and work plans
  - Evaluate how data use improved the impact of rural water programs
- **Accelerate** rural water service provision

While it can at times feel like a chicken and egg situation (which comes first, good data or a good decision-making process?), all evidence-based decisions require high-quality data at the foundation. This is a pivotal moment in the sector. More water point data is being collected than ever, and there is an enormous opportunity to promote data sharing to help provide a clearer picture of the landscape than previously available to all stakeholders. Increasing data sharing and/or conducting new high-powered analyses alone are not solutions in their own right but are key ingredients to the feasibility of improving decisions to optimize investments and increase the impact of programs in the rural water sector. If a common, up-to-date, robust dataset is openly available and widely used, the sector at-large will be able to better explore factors that support or hinder sustainability, benchmark across different countries and contexts, and collaborate with stakeholders and decision-makers to take action based on the resulting evidence. If the sector can first commit to sharing data in a routine, open ,and transparent way, the possibilities are truly limitless.