The Water Point Data Exchange Platform

User Guide Annex 1

This working version of the WPdx User Guide was released in March 2022. Regular updates to the WPdx User Guide are anticipated and will be posted to the WPdx website.

The Water Point Data Exchange (WPdx) is an initiative of the Global Water Challenge.

Generous funding and support to the WPdx platform provided by:

Annex 1: The Water Point Data Exchange (WPdx) Data Standard Monitoring, Evaluation, Adapting and Learning (MEAL) Guide

Annex I Objectives:

- Define a set of standard response options for use during data collection efforts and/or organizational level data cleaning exercises.
- Share details about how shared data is cleaned on the WPdx platform.
- Provide a succinct guide for MEAL specialists/water experts who engage in data collection to understand and implement the WPdx standard for data collection and analysis.

Introduction

The Water Point Data Exchange (WPdx) Data Standard was created in 2015 by an expert working group, with substantial input and review from an array of sector actors. The Standard was designed to define a set of basic parameters which should be collected by all organizations when collecting data from water points. The Standard was also intended to act as a funnel to help compile data from different sources that had been collected for a variety of purposes (monitoring, evaluation, research, inventories, census, etc.). Water point data can be uploaded to the WPdx platform in a variety of formats (csv, xls, json, etc.) and/or directly from other data platforms (Akvo Flow, Kobo Toolkit, mWater, etc.). For more information on how to upload data, please see the Section VIII and Appendix C from the WPdx User Guide.

The Standard, in combination with the WPdx ingestion engine, has enabled the compilation of the world's largest open repository of water point data, with almost 700,000 records available on the WPdx platform (as of November 2022). However, because the Standard allows for open text responses, substantial data cleaning is needed to create a dataset which can be easily filtered and analyzed for insights. WPdx automates these cleaning procedures which are completed as part of the data upload process. The resulting dataset has new "clean" versions of key parameters amended (leaving raw data intact). While this cleaning process has been successful, it does have the potential for inadvertent error and misinterpretation.

The purpose of a data standard is to ensure not only that the right parameters are collected, but that data is collected in a way which is consistent and comparable with data from other organizations and collection efforts. While the WPdx Data Standard includes a set of defined standard parameters, response options are open text and can vary substantially between data sharing organizations. This document updates the WPdx Data Standard to define both the standard parameters and the full set of recommended responses. The WPdx platform will continue to clean and categorize data as needed but recommends that this document and its associated parameters and responses is used by entities as the minimum required during data collection efforts.

Background

The WPdx Data Standard was designed to accept data from a variety of formats to help compile a global dataset of water point records. WPdx does not collect data directly and does not independently verify the quality of shared data. Data is checked for completeness and geographical accuracy (i.e., ensuring

that the points fall within the country where the data is reported to be collected). Instead, data is compiled from governments, leading NGOs, researchers, etc. which have completed their own internal validation efforts before upload. One of the challenges in compiling data from a variety of sources is transforming the different datasets to use a uniform set of terms and categories to ensure the final dataset is consistent and analysis-ready. A simple example which demonstrates the needs for this process is a potential set of entries provided under the #water_tech parameter, which describes the system being used to transport water from the source to the point of collection. A common entry for this parameter is a hand pump, and the most common makes and manufacturers of hand pumps include the Afridev, India Mark I, India Mark II, and India Mark III. Depending on the organization collecting data, datasets uploaded to WPdx to describe an Afridev hand pump might include Afridev Handpump, Afridev hand pump, HP – AfriDev, Afri Dev pump, Afri Ev, etc.

To make this information analysis ready, the terms above must be translated into a consistent format. For WPdx, we use "Hand Pump – Afridev". To preserve the raw data provided, the WPdx datasets include both the original entry, under the #water_tech parameter and the cleaned term under an added parameter, #water_tech_clean. Table 1 provides a sample of common entries received and how they appear in the WPdx dataset.

Table 1. Examples of how #water_tech entries are transformed to #water_tech_clean

#water_tech Common entries received for Afridev Hand Pump	**#water_tech_clean** *Amended entry on WPdx*
Afridev	Hand pump – Afridev
AfriDev	
Afridev Handpump	
AfriDev Handpump	
Hand Pump Afridev	
Aferdive pump	
Afridev, Hand pump	
Pump AFRIDEV	
Hand pump Afridev	
Handpump-Afridev	
AFRIDEV	
Afridev pump	

Many datasets shared with WPdx combine entries for #water_source and #water_tech. Table 2 provides a sample of common entries for a hand dug well, some of which are described as having a hand pump installed. Entries provided in the #water_source parameter are divvied up appropriately to both #water_source_clean and #water_tech_clean by the WPdx cleaning algorithms.

Table 2. Examples of how #water_source entries are transformed to #water_source_clean

#water_source *Common entries received for Hand Dug Well*	**#water_source_clean** *Amended entry on WPdx*	**#water_tech_clean** *Amended entry on WPdx*
Protected dug well	Protected Hand Dug Well	
Manual pump on hand-dug well		Hand pump
Improved Protected dug well		
Protected Hand dug well- normal pump		Hand pump
Protected hand dug well with handpump		Hand pump

F. Protected dug well with hand pump		Hand pump
Protected traditional well		
Protected hand dug well with windlass/bucket		Rope and bucket

While some transformations are relatively straightforward, others can be more complex. For example, the #management parameter tracks who manages the water point. The responses received under this parameter range widely from different terms for community management (water user committee, WUC, committee, village comm, elders, etc.) to specific operators or government agencies responsible for the point. WPdx categories responses from the #management parameter into the following categories which are added under the new parameter #management_clean:

Table 3. Examples of how #management are transformed to #management_clean

#management: *Common responses received*	**#management_clean:** *What is added to WPdx*
Community Management	**Community Management**
Small community WSMT (WATSAN)	
Community committee	
VWC	
Direct Government Operations	**Direct Government Operations**
Local Authority	
Government Operations	
Govt	
Private Operator/Delegated Management	**Private Operator/Delegated Management**
Private Person	
Privat owner	
Private entrepreneur	
Institutional Management - Health Care Facility	**Health Care Facility**
PHU	
Health Care Facility	
hospital	
School Management	**School**
Institutional Management – School	
School Authority	
School staff	
Religious group	**Religious Institution**
Mosque	
Catholic mission	
Church management	
Institution	**Other Institutional Management**
Police	
Other Institution	
Military	
No management structure	**No Management**
No management	
No manager	
Aucune gestion	

Implementing the WPdx Data Standard

While WPdx will continue to provide a service to clean and categorize uploaded data into a set of uniform responses, it would be ideal if data collection was more standardized in terms of the responses collected in the field to ensure a complete and accurate representation of the data.

The following tables provide recommended standardized responses for inclusion when developing a survey for both required and optional parameters from the WPdx Data Standard. These responses were developed and approved by the WPdx working group. The recommended standardized responses represent the formats and uniform terms currently used by WPdx to create consistent responses which are added as new columns in the dataset with "_clean" appended (i.e., #water_source_clean, #management_clean). These terms can also be used by organizations interested in cleaning existing data for internal uses and/or for upload to WPdx. The list of standardized responses, like the WPdx standard will be "living" and organizations are welcome to request additions for review by the WPdx working group.

WPdx Standard Parameters:

The WPdx Data Standard includes 25 parameters which describe key features of a water point. The tables below provide summary information for each parameter.

Required Parameters

Table 4 summarizes the parameters which are required for records to be uploaded to the WPdx platform. These parameters provide the basic "who, what, where, when and functionality status" for each water point. This is the minimum amount of information that should be collected when visiting water points.

Table 4. Required parameters from WPdx Data Standard

Parameter	Hashtag	Format	Description	Standardized responses
Latitude	*#lat_deg*	Float	Provide the decimal value of the latitude in WGS 1984. At least four decimals (but more are encouraged) should be included. North and East should be noted as positive numbers.	Decimal value from WGS 1984
Longitude	*#lon_deg*	Float	Provide the decimal value of the longitude in WGS 1984. At least four decimals (but more are encouraged) should be included. North and East should be noted as positive numbers.	Decimal value from WGS 1984
Data Source	*#source*	Text	Provide the name of the organization collecting the data record.	Formal name of organization sharing data
Date of Data Inventory	*#report_date*	Date; see suggested format	Provide the date that the data was collected on using ISO 8601. Time and time zone designator are optional.	YYYY-MM-DD
Presence of Water when Assessed	*#status_id*	Binary: yes/no	Identify if any water is available on the day of the visit, recognizing that it may be a limited flow.	Yes, No

At least one of the following (ideally both if available):				
Water Source	*#water_source*	open text; see suggested standardized responses	Describe the water source	- Piped Water - Borehole - Protected hand dug well - Unprotected hand dug well - Protected Shallow Well - Unprotected Shallow Well - Protected Well - Protected Spring - Unprotected Spring - Rainwater Harvesting - Sand or Sub-surface Dam - Delivered Water - Packaged Water
Water Point Technology	*#water_tech*	open text; see suggested standardized responses	Describe the system being used to transport the water from the source to the point of collection (e.g. Handpump (include manufacturer, i.e. Afridev, India Mark II, Malda, etc.), Kiosk, Tapstand, etc.).	- Hand Pump - Hand Pump – Make - Kiosk - Mechanized Pump - Mechanized Pump – Diesel - Mechanized Pump -Hydram - Mechanized Pump – Solar - Mechanized Pump – Wind - Rope and Bucket - Tapstand

Optional Parameters

WPdx includes 19 optional parameters. These parameters are designated as optional in recognition that not all entities are able to collect all this information. However, based on experience in developing the analytics for the WPdx decision support tools, some of the optional parameters have been found to be quite informative and are highly recommended for inclusion in data collection efforts. The optional parameters are divided below into two groups: Group A, as shown in Table 3 are highly recommended for inclusion in data collection. Group B, as shown in Table 5 are recommended parameters, but not vital.

Table 5. Optional Group A Parameters from WPdx Data Standard

Parameter	Hashtag	format	Description	standardized responses
Management	#management	Open text	Select the classification of the entity that directly manages the water point.	- Community Management - Direct Government Operations - Private Operator/Delegated Management - Health Care Facility - School - Religious Institution - Other Institutional Management - No Management - Other - Unknown
Installation Year	#install_year	date	Provide the 4-digit installation year (e.g. 1994).	YYYY
Rehabilitation Year	#rehab_year	date	Provide the 4-digit year when the most recent major rehabilitation (not just regular maintenance) occurred (e.g. 1994).	YYYY
Condition	#status	open text	Provide a status of the physical/mechanical condition of the water point.	Open text
Fecal Coliform Presence	#fecal_coliform_presence	Binary	Results of e. coli or thermotolerant coliform water quality test from a 100ml water sample collected directly from water point. Total coliform should not be included. If thermotolerant, must be noted in the metadata. Value should represent presence or absence.	Absent Present
Fecal Coliform Value	#fecal_coliform_value	Float	Results of e. coli or thermotolerant coliform water quality test from a 100ml water sample collected directly from water point. Total coliform should not be included. If thermotolerant, must be noted in the metadata.	Value should represent the most probable number or colony forming units in 100ml. (e.g. "20" to represent 20 colonies per 100 ml)
Subjective Quality	#subjective_quality	open text	Information regarding the perceived quality of the	Suggested responses include: Acceptable

			water including taste, appearance, and/or odor.	quality, unacceptable quality, bad taste, bad odor, bad smell
Photograph	#photo_lnk	Link	Provide the URL of a photograph of the water system.	Link to a hosted photograph. Multiple URLs can be included, with each URL separated by a semi-colon (;).
Payment for Water	#pay	open text	Provide the payment amount and basis. If no amount is provided, the basis can be provided alone. An amount without a payment basis cannot be included.	Suggested entries include: No payment, fees collected – at point of collection, fees collected - metered, fees collected – monthly, fees collected – upon breakdown, fees collected – basis unknown

WPdx uses a data cleaning algorithm to add a new parameter #status_clean to the dataset. The #status_clean parameter is created based on cleaning a concatenated version of #status_id and #status. This new parameter is used to identify points which are abandoned to be removed from the Rehabilitation Priority analysis as well as to provide a more detailed picture of the functionality of the water point. Please see Table 4. for an example.

Table 6. Example of how #status_clean is created from #status_id and #status parameters

#status_id	#status	#status_clean
Yes	No problems	Functional
Yes	But functioning poorly	Functional, but needs repair
Yes	Dry condition	Functional, but low yield
Yes	Not in use	Functional, but not in use
Yes	Abandoned	Abandoned
No	Pump breakdown	Non-functional, technical issue
No	Pump stolen	Non-functional, stolen parts
No	Dry	Non-functional, dry
No	No funds	Non-functional, financial issue
No	Abandoned	Abandoned

The parameters below in Table 7 are still ideal to include when available to provide additional details about the water point. The country and administrative divisions are added automatically by using the provided GPS coordinates, under the parameters #country_name_clean, #adm1_clean, #adm2_clean, #adm3_clean, though user entries can be useful to have included in areas where there are disputes regarding boundaries and to verify for good quality data.

Water point ID and scheme IDs would be incredibly helpful to have but appear to be rarely used in a consistent fashion in most datasets received by WPdx. If an organization is using a consistent water point ID and/or scheme ID, it is highly recommended to include these in the dataset and to mention either in the metadata section and/or through direct communication with WPdx via info@waterpointdata.org.

Table 7. Optional Group B Parameters from WPdx Data Standard

Parameter	Hashtag	format	Description	standardized response
Primary Administrative Division	#adm1	open text	Provide the name of the primary administrative division.	The correct unit can be found at http://www.statoids.com. This corresponds to "First Order" and "First Level" administrative units at http://Geonames.org and http://www.gadm.org respectively.
Secondary Administrative Division	#adm2	open text	Provide the name of the secondary administrative division.	The correct unit can be found at http://www.statoids.com. This corresponds to "Second Order" and "Second Level" administrative units at http://Geonames.org and http://www.gadm.org respectively.
Tertiary Administrative Division	#adm3	open text	Provide the name of the tertiary administrative division.	The correct unit can be found at http://www.statoids.com. This corresponds to "Third Order" and "Third Level" administrative units at http://Geonames.org and http://www.gadm.org respectively.
Water Point ID	#activity_id	Alphanumeric/numeric depending on format used by organization	Provide the Unique ID for the specific water point infrastructure, as reported by data collector.	Record the physical ID on the water point or an internal system ID.
Scheme Identification	#scheme_id	Alphanumeric/numeric depending on format used by organization	The identifier for a small-piped scheme that connects multiple individual water points.	This could be a physical ID on the scheme or an internal system ID.
Installer	#installer	open text	Provide the name of the entity or entities that installed the water system. This should be the entities that completed or were directly responsible for the construction, rather than a donor or other involved stakeholder.	Open text. Formal name of installing organization.

Rehabilitator	#rehabilitator	open text	Provide the name of the entity or entities that completed the most recent rehabilitation of the water system. This should be the entities that complete or were directly responsible for the construction, rather than a donor or other involved stakeholder.	Open text. Formal name of the rehabilitating organization.
Notes	#notes	open text	This field can be used to incorporate any additional information not already part of the WPdx standard that is useful to the data provider.	Open text
Public Data Source URL	#orig_lnk	link	Provide the public link to the data record for a specific water point or full data set, including any non-standard compliant data.	Link to the organization's full data set