Water Data Infrastructure for Low- and Middle-Income Countries

Kyle Onda, Associate Director, Internet of Water Center for Geospatial Solutions
Lincoln Institute of Land Polic

Vision

Water is critical to human development and well-being. Its importance is not only captured explicitly in the Sustainable Development Goal 6 (SDG6) regarding universal access to safe, affordable, and adequate water and sanitation, but also is embedded within the goals on energy, food security, poverty, energy, health, disaster risk, and cities. In 2016-2018, the United Nations and World Bank Group convened a High Level Panel on Water to accelerate progress towards SDG6, which identified a number of ways in which such progress was off-track, particularly in low- and middle-income countries (LMICs). This panel recommended an agenda to enable stakeholders to make decisions and take action. Prior to decision-making, it is necessary for stakeholders "to understand the quantity, quality, distribution, use, and risks of the water they have." This understanding in turn depends on investments in institutional and technical infrastructure for "water-related data as well as the systems to share, analyze, and take decisions with this data" (United Nations and World Bank 2018).

This makes sense — decisions require information based on data. But what data investments need to be made? What should water data infrastructure do? "Water-related" data is collected on many topics by many different actors, each in a different format and quality fit for a different, often highly localized primary use, even if secondary use of such data by other parties may be beneficial. Well-informed decisions may require data to be integrated from many of these different sources. Currently, the typical data-to-decisions cycle can be long, winding, difficult, and full of uncertainty. For example, a regional water infrastructure planner may need to answer many interrelated questions in order to even begin costing out alternative scenarios of water supply improvements in an application to a government agency or donor organization:

- How many people live in this region, and how are they spatially distributed among settlements of different sizes?
- Where do existing water points serve each of these settlements? What types of water points are they, and what levels of service in terms of water quantity, quality, travel times, and wait times do they provide? What are people paying for water from these points? Where do people retrieve water if not from one of these points?
- How much water is being used, including for domestic, agricultural, livestock, and commercial purposes?
- What are prevailing rates of waterborne disease in these settlements, and in their vulnerable subpopulations?
- What is the potential to increase safely managed water supply quantity? What are the surface and groundwater availability conditions in the

catchments each settlement is in? Are any of the potential water sources claimed by external agricultural, industrial, or real estate interests? Are there any important ecological habitats, and what flows are required for them?

- Are there any threats of serious chemical or biological contamination to existing or potential water sources?

Multiple organizations collect data relevant to each of these questions. Examples include central and local government agencies, bilateral and multilateral donors, international NGOs, local and international academic institutions, and the private sector, all of whom may or may not make their data available at all, let alone in a format useful to analysis. The planner may miss important data sources, and the data she does manage to gather will probably be in different formats, collected in different units, and at different spatial and temporal scales. It will require tremendous effort to organize this data into usable information. Moreover, the data itself may be quite sparse, and the planner may have no idea if she needs to invest in more data collection or if there are organizations with the required data that she is merely unaware of. All of these issues compound for monitoring and evaluation, which require repeated measurements and recurrent data integration and analysis workflows.

These difficulties should frame the desired interventions into the existing water data ecosystem to improve, create, and connect water data infrastructures relevant to LMICs. While there are diverse, highly localized needs for data improvements, some generalized needs can be inferred to help frame a path forward. Consider three highly generalized, interrelated user stories[1]:

"As a water decision-maker, I need information that communicates key indicators in a clear and timely manner based on best available, up-to-date data, so that I can make informed decisions and justify my decisions to those I am accountable to."

"As a water decision support system creator, I need to automate access, transformation, and analysis workflows for all data necessary to calculate key indicators, so that I can consistently and quickly create and deliver information required by decision-makers, preferably in a dynamic manner based on their input on the fly."

"As a water expert, I need to discover, access, and use data that is relevant socially, hydrologically, spatially, or administratively to a feature[2] I care about from all organizations that hold such data, so I don't need special knowledge to access some data, and so I don't miss potentially relevant data, and so I can direct application developers and decision-makers to the best available data with confidence in a timely manner."

These user stories illustrate some key elements required to improve decision making in the water sector, ranging from the collection, interpretation, and analysis of data, to the development of complex tools and information products based on data, to the taking and justification of decisions based on credible information. These elements can be distilled into three key layers of a social-technical water data infrastructure:

1. For data users to design appropriate information products, the data they use needs to be:
    a. **Discoverable** (i.e., quickly found based on relevant search criteria)
    b. **Accessible** (i.e., available for download from listed sources)
    c. **Interoperable** (i.e., readily interpreted based on good documentation, rich metadata, and predictable open formats)
2. For people to create information products in a timely and cost-effective way, data underlying information products need to be:

1 A user story is an informal description of a system written from the perspective of a user of that system, and is helpful in system design by keeping designers aware of the key problems they are addressing.

2 Any object (digital/virtual or physical) associated with a real-world location relevant to a water-sector activity.

a. **Accessible** (i.e., consistently available in a way that computer programs can immediately access, such as from the internet)

b. **Interoperable** (i.e., conform to a consistent structure in common across datasets that computer programs can interpret and manipulate)

3. For people like decision-makers who just need to know an answer to a question, access to information products is required more so than basic data.

In general then, any water data infrastructure should work to make its constituent data discoverable, accessible, and interoperable. Data must be published online in such a manner that interested data users can find them. Data must be available for download and use to authorized parties, and as much data as possible should be accessible to all. Finally, data about a given topic from multiple sources should have sufficient metadata and documentation to be used, and should be formatted in such a way so as to be able to be integrated automatically with each other. All of these elements are, of course, easier to enumerate than to implement. The remainder of this piece is organized as follows: Section 2 elaborates the technical and institutional requirements of the vision established above, with reference to existing activities, data systems, and technologies in the sector that can be built on to meet these requirements. Much of the material in this section is based on conceptual work undertaken in the United States context by the ***Internet of Water*** project at Duke University (Patterson and Onda, 2020). Section 3 describes some key investments that can be made to progress towards the vision with varying levels of required commitment and associated barriers, risks, and potential payoffs.

An architecture for water data infrastructure

The vision of seamless data discovery, access, and use across innumerable datasets requires a coherent architecture to realize. At first glance, a tempting architecture would be a unified system, where data on all relevant topics could be collected and standardized in a single database and published in a single data portal, similar to the way that the ***Energy Information Administration*** manages energy-related data in the United States (Josset et al., 2019). However, the EIA handles a limited number of data types about a highly regulated and technically uniform system. Such an effort is unlikely to scale much beyond existing data centralization efforts, such as the Joint Monitoring Program (WHO & UNICEF, 2021). Water data exists across many domains. Water data fragmentation is not merely physical but also social and institutional. Constantly expanding flows of highly heterogeneous data would be enormously costly to control in both a technical and administrative sense, involving countless hours of laborious data extraction, transformation, and loading work conducted by thousands of people. A more realistic architectural approach can be borrowed from the one in use for the internet (Patterson et al.. 2017); that is, a web of interconnected resources with a minimal way to express relationships to each other through links.

The concept of a decentralized, federated data architecture

Such decentralized, federated, linked-data architectures are nascent in the water sector in high-income contexts like North America, Europe, and Australia, let alone in LMICs. However, they show promise, and represent an opportunity for LMICs to leapfrog legacy data architectures that are more heavily reliant on centralization in high-income countries. What are the components of a linked data architecture (Figure 1)?

1. **Data Producers** collect and publish data. They may do so independently, but their data will reach the largest audience when its data is represented in a **Hub**. Data producers can range in complexity from individuals with scanned handwritten documents

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	

to internet-connected real-time water quality sensors.

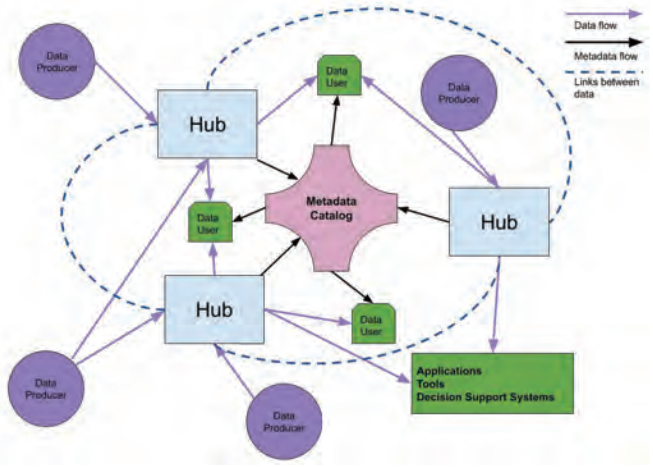
2. **Hubs** are organizations that standardize and publish data and metadata from many data producers about one or more thematic areas in one or more regions. They provide access points for data from many producers.

3. **Links** are encoded relationships between data. They can be published as parts of datasets within **Hubs**, and allow **Data Users** to discover how different datasets might be related and fruitfully integrated.

4. **Metadata catalogs** are repositories of descriptive information about data. They are analogous to library catalogs for bibliographic resources, such as books and journal articles. A catalog may be associated with a single **Hub**, serving essentially as a **Hub's** user interface. Catalogs may also harvest and republish data from multiple **Hubs**, including direct linkages to where data can be accessed directly. They publish metadata in standardized ways that allow search engines, such as Google, Yahoo!, and Bing, to index important information, including **Links**. They thus provide an important entry point for data discovery by **Data Users** who do not know about any particular **Hub** or **Data Producer** ahead of time.

5. **Data Users** might be humans or machines. They may use search engines to naively discover data, use **Metadata Catalogs** to automate data harvesting, and use **Hubs** to access data.

Figure 1.
Federated linked-data architecture conceptual diagram.



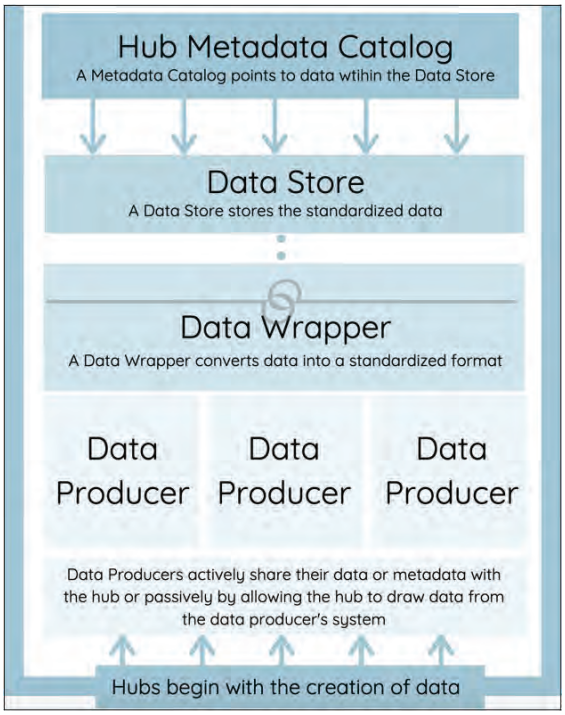
In this architecture, accessibility and interoperability are achieved through the interaction of data producers and hubs in managing data with respect to data standards. Data discoverability is achieved through the interaction of metadata catalogs and links.

Hubs: Institutional and technical organization of accessible and interoperable domain data

Hub Architectures

Within the overall architecture, hubs are the mechanism by which data is organized, standardized, stored, and delivered in interoperable formats. Data hubs are meant to provide "data as a service," moving beyond the traditional publication of data as a static product to be downloaded and used offline by data users. Data "services" allow users to query datasets for the subsets relevant to them in a variety of useful formats through web application programming interfaces (APIs) (Blodgett et al., 2016). APIs allow computers to send and request data to each other. Data services thus allow machine-to-machine communication necessary to create dynamic applications or analysis workflows that integrate data from many sources as needed, on-the-fly. Hubs can be organized in varying ways, depending on the nature of data collection and publication for a given topic and region. However, any hub has a few common functional elements (Figure 2).

Figure 2.
Components of a data hub.



A metadata catalog should provide at minimum a listing of all available datasets within the hub, but can also include search functions and APIs for dynamic data access. A data store can be a database or simple store of independent files, as long as data standards are enforced. A data wrapper can be an automated or manual process of data standardization (reproduced from the Internet of Water Hub Diagram, ***https://internetofwater.org/wp-content/uploads/2020/08/Hub-Diagram.pdf***) .

Within this basic pattern, hub types can be organized around the questions:

- Who stores and serves data?
- Who standardizes data?

The answer to each of these questions is either the data producers or the hub administrators. The suitability of a hub type for a given scenario thus depends on tradeoffs between technical and administrative capacities of hub administrators and data producers. There are four basic hub configurations:

1. **Centralized** hubs involve hub administrators taking on the burden of both standardizing data from various producers and storing that data. An archetypal example of this in the water sector is the ***Joint Monitoring Program*** (JMP) whose staff aggregates data from various national statistical agencies and surveys, such as the DHS, MICS, and LSMS, aligning the various indicators about water source and sanitation infrastructure types into a common vocabulary, and publishing the combined data with consistent cross-country comparability. The advantage of this approach is the degree of quality control that can be exercised, with a single common data model being enforced by staff of the hub itself. Data producers require little, if any technical capacity, and merely have to direct hub staff to where data can be accessed. It enables extraordinarily complex and heterogeneous data to be integrated and delivered in a common format. The disadvantage of this approach is the concentration of labor in one organization to handle the Data Wrapper component. This cannot

scale to large numbers of data producers and/or high frequency of data updates.

2. **Integrated-push** hubs involve an integrated system of data collection. Data wrapping is generally implemented by requiring all data submissions to come in the form of a data template. The template can vary in complexity from a CSV or Microsoft Excel file with prescribed column headers to a populated file database. An archetypal example of this is the Water Points Data Exchange (WPDx), which requires data to be submitted through upload of a CSV template via web form. Data standardization is thus the responsibility of the data producer, and some degree of data validation can be automated. This allows this kind of hub to scale to large numbers of data producers, even with relatively complex data standards. However, quality control can become difficult, given that data producers may submit data with incorrectly filled or incomplete templates. There becomes a tradeoff between data quality control and the volume of data submitted. In addition, push processes may not be suitable for very high-frequency data.
3. **Integrated-pull** hubs are similar to integrated-push hubs, but instead of data producers submitting data through a central interface, data producers host a dynamic data endpoint that hub administrators ingest data from on creation. Somewhat high-frequency data becomes possible to ingest with this architecture, although it either requires a degree of technical sophistication to data producers that may reduce the scope of participation, or require the maintenance of cloud-based data management software by the hub on behalf of the data producers. An archetypal example of this is mWater, which ingests data from a mobile application that it maintains into a centralized data platform (mWater, 2021).
4. **Federated** hubs do not actually store any data. They function, rather, as metadata catalogs for a community of data producers that maintain independent data systems that can return data in a common format. Data users can query data from any participating data producer from the catalog, but data is delivered directly from the data producer. These systems are the only appropriate type for large networks of high-frequency sensors designed to provide real-time alerts, such as flooding early warning systems, since there is no delay between data production and data standardization and storage. They also require comparatively little storage or technical administrative overhead for the hub relative to the volume of data that can be made available to data users. However, a disadvantage is that if a contributory data producer's system goes offline, its data would not be available from the hub unless the hub caches the data. In general, this type of hub is a poor choice for contexts where many data producers do not have access to a great deal of technical capacity. There are no examples of this type of hub that is focused on LMICs specifically, although individual countries may have hydrometeorological sensor networks that are architected this way. The CUAHSI ***Hydrologic Information System*** provides a metadata catalog to a global community of meteorological and hydrologic sensor networks.

Standards

Standardized data services require data standards and API standards be adopted and enforced by the hubs. Below are data standards and API standards that are relevant to the water sector. Many necessary standards may not yet exist and could be priorities for development within the broader water sector community.

Data Standards

Data standards are rules governing how data are described and recorded. They are typically comprised of four elements:

- **Schemas**, or **data structure standards**, are lists of mandatory and optional data elements. In their simplest form, they are the column headers of a spreadsheet. More complex schemas may involve multiple tables, each with its own lists of elements, or nested structures.

- **Data content standards** provide specific guidance for how to fill out data elements. This includes detailed definitions (e.g., the "identifier" might be the serial number for a water pump) or prescriptions for data type (e.g., text, numeric), format (e.g., YYYY-MM-DD for dates), or units (e.g., "kilogallons" for monthly water consumption). Content standards are often packaged along with schemas. See Table 1 for common schema/content standards in the water sector.
- **Data value standards** or **controlled vocabularies** are lists of specific terms that are allowed to populate certain schema elements. For example, a binary data element may be associated with the controlled vocabulary of the set "Yes, No", or the data element for "country" may be associated with the controlled vocabulary of ISO 2-character country codes. It is often useful for controlled vocabularies to refer to specific terms with unique URLs that point to machine-readable definitions of those terms. For example, the term "Faecal coliforms" in the Food and Agricultural Organization Vocabulary (AGROVOC) can be denoted unambiguously in datasets with the URL *http://aims.fao.org/aos/agrovoc/c_36372*. This can be useful when automating the integration of datasets with different terms for the same concept. See Table 2 for some data value standards that may be useful in the water sector
- **Data format** or **data exchange standards** prescribe the specific file type that the data can be sent to users in. These can include simple CSV for tabular data, specific XML or JSON encodings for complex nested data structures, and particular formats for geospatial data, like ESRI shapefile or OGC GeoPackage. In general, it is preferable for data format standards to prescribe open rather than proprietary formats to allow the maximum number of people to access and use the data. See Table 3 for some data exchange standards useful in the water sector.

Table 1.
Data schema and content standards relevant to the water sector.

Topic	Schema/ Content Standards
General Metadata	*Dublin Core*
Geographic and Jurisdictional areas	FGDC Boundaries
Hydrography	*FGDC Hydrography*
Water Quality Samples	*WQX*, *WaterML2 Water Quality Profile*
Time Series "Sensor" Data	*WaterML2 Part 1*
Surface Hydrology Features (characterizing streams, lakes, etc.)	*WaterML2 Part 3*
Groundwater	*WaterML2 Part 4 (GWML)*
Water Rights and Use	*WaDE Schema*
Water Points of Use	*WPDx Standard*
Donor-funded Projects	*IATI Standards*
Utilities and Infrastructure	*FGDC Utilities*

Table 2.
Data exchange standards relevant to the water sector.

Data Type	Controlled Vocabularies
Hydrological Observations	*ODM2*, *NEMI*
Quantities, Units, and Dimensions	*QUDT*
Household Surveys	*IPUMS harmonized variables*

Table 3.
Data exchange standards relevant to the water sector.

Data Type	Exchange
Geospatial Vector (point, polygon, etc.)[3]	*GeoJSON*, *GML*, *GeoPackage*
Geospatial Raster (matrix, image, etc.)	*GeoTIFF*, *netCDF*
Tabular Data (general)	CSV, *JSON*, netCDF, *Tabular DataPackage*
Nested and Multidimensional Data	JSON, *XML*, netCDF

API Standards

There are many APIs. While these are incredibly useful to data users, the sheer variety of APIs can limit interoperability between data systems, since any given data integration activity requires custom code to interact with each and every API. Widespread adoption of API standards can reduce this burden. API standards are rules that define the pattern of an API, specifying that an API receiving a request in a given format will deliver a predictable response. API standards often work with data content and structure standards to ensure that the requested data is accurately transferred. So-called RESTful APIs allow data to be queried simply by going to a particular web URL, with specific parameters changed to subset data. These special URLs are referred to as "API calls."

Different data systems can choose to provide data using open API standards, ensuring that similar requests made to each system receive similar responses, even if the underlying databases and API code are fundamentally different. Many open API standards are developed by international communities of practice, such as the *Open Geospatial Consortium*. For example, the EU INSPIRE directive requires that European statistical and environmental agencies publish data using a limited number of OGC standard APIs. *This demonstrates how information on such diverse topics as demography, water quality, and air quality can be delivered using the exact same API*. See Table 4 for some API standards useful in the water sector.

Table 4.
API Standards relevant to the water sector.

Data Type	API Standards	Open Source Implementations	Proprietary Implementations
Geospatial Vector	*WFS* *OGC Features*	*Geoserver*, *QGIS*, *MapServer* *Geoserver*, *PyGeoAPI*, *QGIS*	*ESRI*, CubeWerx *CubeWerx*
Geospatial Raster	*WCS*	*Geoserver*, *QGIS*, *MapServer*	*ESRI*
Map Imagery	*WMS*, *WMTS*	*Geoserver*, *QGIS*, *MapServer*	*ESRI*, *CubeWerx*
Georeferenced Observations/ Time Series/Samples	*SOS* *SensorThings API*	*52North*, *istSOS* *FROST*, *GOST*, *52North*	*Kisters*, *SensorUp* *SensorUp*
Tabular Data	*ODATA*	*CKAN*	*Socrata*

3 A particular omission many may recognize is that of the ESRI Shapefile and geodatabase formats for geospatial vector data, which are very commonly used. Shapefiles have a number of disadvantages that are detailed in at *switchfromshapefile.org*, and for the most part can be completely replaced in practice with GeoPackage for bulk file transfer or GeoJSON or GML for web services/web mapping.

Existing Hub Examples

Having elaborated the purpose, function, and technical requirements of data hubs, the existing landscape of hubs relevant to the water sector in LMICs can be surveyed and evaluated. Many of the use cases listed in Section 1 require data from one of four basic water data domains:

- **Hydrometerological** observations of precipitation, streamflow, and groundwater levels are important to estimate and project surface and groundwater availability.
- **Water quality samples** of environmental waters and water at points of collection and use are important to monitor water safety and sanitation.
- **Household surveys** are important to monitor sociodemographics, economic and health conditions, and water use patterns.
- **Project and system KPIs** are important to monitor service levels, performance, and sustainability of water systems.

Below, a few examples of hubs addressing one or more of these data domains are profiled, including hubs in high-income contexts that nevertheless have important lessons relevant to LMICs.

Hydrometeorological Hubs

1. The Global Runoff Data Centre (***GRDC***) of the ***Global Terrestrial Network*** — Hydrology

The GRDC is an "integrated-push" hub for streamflow data, collecting, and standardizing data from the national hydrologic or meteorological agencies of 159 countries. Coverage tends to be greatest in North America, the EU, Australia, and New Zealand for up-to-date records, although there is also substantial coverage in Brazil, Chile, South Africa, and Namibia. The GRDC does not truly operate as a "service," as data is only available using a human-oriented user interface, although it is delivered in standard formats including WaterML2. The GRDC has a fairly restrictive data license which prohibits commercial use or redistribution of data to third parties. The GRDC is financed by the German federal government.

2. The Trans-African Hydro-Meteorological Observatory (***TAHMO***)

TAHMO is a nonprofit integrated-pull network of 20,000 weather stations installed across many eastern and western African countries, typically in schools. It uses a standard suite of sensors , data loggers, and cellular telemetry (from ***METER Group***) to stream data into a central cloud database. It operates with official MOUs with relevant national meteorological agencies, and makes data available to these agencies for free. Data use requires a data use agreement, although data for academic or public use is free, and there is a charge for commercial use. TAHMO is largely financed by public and corporate donors, but is attempting to develop information products for commercial sale.

Water Quality Sample Hubs

1. Global Environmental Monitoring System for Freshwater (***GEMStat***)

GEMStat is a global water quality "integrated-push" hub providing water quality sample data from ~13,000 river, lake, reservoir, wetland, and groundwater locations globally, including substantial coverage throughout South America, and to a lesser extent Sub-Saharan Africa and South Asia. It is operated by the International Centre for Water Resources and Global Change, Koblenz, Germany, in coordination with the UN Environment Program. Data providers manually submit Excel templates via email. Most data providers are national environmental regulatory or science agencies. Data providers can choose whether data available in the hub can be provided with open or more restrictive licenses. It operates a metadata catalog and a human-oriented map interface to visualize or download data in a nonstandard CSV format via email delivery; it offers no API. This may be due to the general data use agreement of the host organization, which does not allow for redistribution of data without written permission.

2. The National Water Quality Monitoring Council Water Quality Portal (*WQP*)

The WQP is a U.S.-focused but globally open "integrated-push" hub providing water quality sample data from hundreds of thousands of sites in the United States. It is operated by the United States Geological Survey (USGS) in coordination with the U.S. Environmental Protection Agency (USEPA). A variety of international sites also have data, mostly due to cooperative activities with U.S. Federal agencies or academic institutions. However, data can be submitted by any organization anywhere to the portal via upload of Excel templates to the *USEPA WQX system*, using the WQX data standard, which is quite comprehensive for water quality metadata. Data is freely available and can be redistributed, and is published via both web interface and an *API (WFS standard)*, and is thus incorporated into many third- party tools.

Household Survey Hubs

1. The Joint Monitoring Program (*JMP*) for Water Supply and Sanitation by WHO and UNICEF

The JMP is the official UN body responsible for monitoring progress towards SDG6. It is a "centralized" hub regarding water and sanitation service levels according to a minimally general data model. It aggregates and standardizes data from nationally representative censuses and surveys, as well as donor agency surveys, such as the USAID Demographic and Health Surveys (DHS) and the UNICEF Multiple Indicator Cluster Surveys (MICS). Data is available for bulk download but not via API. Also, the spatial granularity of the data is limited, usually only at a national or regional scope.

2. *IPUMS*

IPUMS is a "centralized" hub that provides access to standardized census and survey tabulations and microdata from most countries. It is funded primarily by U.S. Federal agency grants and operated by the University of Minnesota. Spatial data granularity depends on the source data and can range from 1st-level administrative divisions, like states, to 3rd-level, like sub-districts. Both raw representations of source data and "harmonized" versions that standardize variables across countries and time periods are possible to access. Data is available for bulk download via the web interface, and *custom API development* is underway.

Project/System Hubs

1. The International Aid Transparency Initiative (*IATI*)

The IATI is an initiative to provide open access to records regarding international aid spending. It is an "integrated-push" hub that relies on donors submitting data in the *IATI Standard*, which has an XML-formatted schema and content standard, including *controlled vocabularies* for items such as countries, regions, and sectors, and specific activity types. Location data is expressed as a point location as well as administrative geography. IATI data is fully open source and available for bulk download and through an API, as well as through third-party APIs and value-adding analytical tools. It uses the open source *CKAN* platform as its data catalog and management system.

2. The Water Point Data Exchange *WPDx*

The WPDx is a global "integrated-push" hub focused on organizing data about rural water point locations, water sources, and operational status, including faecal coliform test results. It has a *simple schema* with rather flexible content standards (with many free text possibilities), and data providers upload data in CSV templates through the web interface or from a remote file provided by an API. Location data is expressed as a point location as well as administrative geography. It is operated by the *Global Water Challenge* coalition. *Data access* is provided through an instance of *Socrata*, a proprietary data management platform similar to the CKAN platform used by IATI. Data can be sorted or visualized within the platform, downloaded in bulk, or accessed via the Socrata API, as well as a standard OData API suitable for connection to third-party data analysis tools. WPDx also provides some *analytical tools* that incorporate data from its own system, as well as from ESRI population estimates,

demonstrating the power of data integration using APIs.

3. The International Benchmarking Network (***IBNET***)

IBNET is an "integrated-push" hub focused on key performance indicators for water and sanitation utilities. It has a large and complex schema, expressed in Microsoft Excel templates for data providers distributed as a "***toolkit***." Aggregated data by country can be downloaded and utility-specific data reports can be dynamically visualized using the web interface, which can also facilitate manual downloads. The degree of spatial granularity is limited, with "country" being the only spatial filter. There is no API for automated data querying. IBNET is funded by the World Bank.

4. The Rural Water and Sanitation Information System (***SIASAR***)

SIASAR is an "integrated-pull" hub that is similar in data content to IBNET, but with a simplified set of indicators and a particular focus on rural water systems. It began as a collaboration among the governments of Honduras, Nicaragua, and Panama, but has since expanded to other Latin American countries as well as Uganda and Kyrgyzstan. Data entry is mediated through a ***mobile app*** that enforces the data standard. Location data is expressed as a point location as well as administrative geography. Data is provided under an open license in the form of bulk Excel files on a country basis. There is no API for automated data querying.

Thus, there are a variety of hubs that are active and could be extended, providing access to data relevant to the water sector in LMICs. The degree to which they serve accessible and interoperable data is variable. WQP, WPDx, and IATI in particular provide admirable levels of accessibility and data standardization, while IBNET, SIASAR, and IPUMS could be improved with more modern data practices such as APIs. TAHMO, GRDC, and GEMStat provide valuable data, but accessibility is somewhat limited by valid data-licensing concerns. There are notable gaps in sanitation point operational data, as well as streamflow and well-level monitoring. All of these hubs, however, are currently difficult to integrate data from, due to a lack of proper data discoverability infrastructure or practices. What would it take to make data from the above hubs discoverable and integratable on demand?

Linked data: A framework for data discovery and integration within and across hubs

Currently, data integration in the water sector (and most sectors) is done manually. Data users interested in a particular area need to create ad-hoc procedures to determine which subsets of each of many datasets match up with each other before creating a unified data product for analysis. On the internet, similarly complex data integration tasks have been automated for use cases, such as discovering restaurants and their menus along transportation routes. This is accomplished not by curated, large combined datasets but by *indexing linked data*.

Linked Data Best Practices

The Open Geospatial Consortium has been developing methods to emulate the organization of the internet for water-related data through the environmental-linked features interoperability experiments (Schleidt et al., 2020; Blodgett et al., 2020). The experiments are now being operationalized in Australia with the ***Location Index*** project, and in the United States with the ***geoconnex.us*** project. This involves a simple web architecture with three best practices:

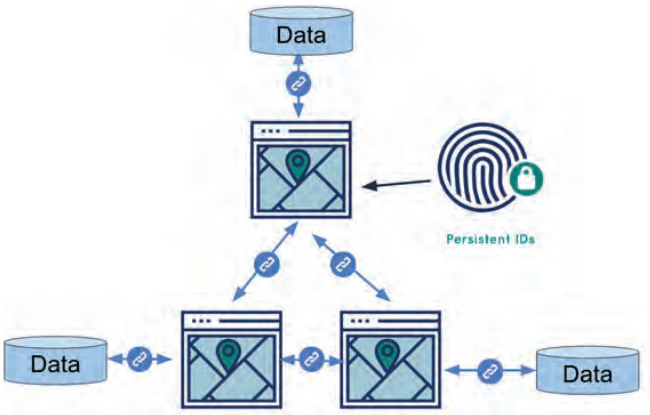
1. **Landing pages** for features of interest. This means that every real-world feature that an organization publishes data about should have a web page providing basic metadata about that feature. For example, ***https://waterdata.usgs.gov/monitoring-location/11164500/*** is the USGS web page about a particular USGS stream gage near Stanford University. This practice allows highly specific data subsets to be indexed by search engines (which index web pages), allowing a search for "USGS stream gage Stanford" to return the desired links. Where data privacy issues preclude publication of such granular data, the feature of interest can be an area for which aggregated or otherwise de-identified data is reported.

2. **Persistent (HTTP) identifiers** for features of interest. This means that every real-world feature that an organization publishes data about should have an identifier in the form of a URL that redirects to the corresponding **landing page**. This is similar to the Digital Object Identifier (DOI) System used for academic publications and datasets, but much more granular. For example, *https://geoconnex.us/usgs/monitoring-location/11164500/* is the persistent identifier that redirects to *https://waterdata.usgs.gov/monitoring-location/11164500/*. Persistent identifiers allow datasets to reference each other in a persistent way that is robust to changes in data provider websites. In this case, the persistent identifier directed users to a previous generation USGS website that will be retired (*https://waterdata.usgs.gov/nwis/uv?site_no=11164500*), but the USGS updated the identifier to point to a new web page.

3. **Embedded links** to data and related features. This means that landing pages include links to various datasets about the feature of interest. The links can be both in human and machine-readable form. In the stream gage example, an embedded link to *machine-readable data* (in the form of an API call to a related USGS data hub) can be processed and *visualized automatically*. Another link might be embedded to *https://geoconnex.us/ref/hu10/1805000304*, the identifier for the watershed in which the gage is located, or to *https://geoconnex.us/ref/places/0673906*, the U.S. Census Place boundary it is within. Embedded links need to be structured in highly specific ways to be machine readable; the preferred way to do this for web development is to embed *JSON-LD* metadata, with special controlled vocabularies to denote different types of links (Sporny, Kellogg, and Lanthaler, 2014). One important controlled vocabulary is *schema.org*, which was developed by Google, Yahoo, and Microsoft to standardize how web pages describe themselves to their search indexes (Guha, Brickley, and Macbeth, 2016). Other controlled vocabularies may need to be developed specifically for the water sector.

Many of these practices are actually precluded by the complex legacy data systems in use by environmental data agencies in high-income countries. However, in LMICs, where data hubs may be newer and simpler, these practices may represent more of an add-on or plug-in. From a community of hubs publishing data in this way would emerge an implicit web of interconnected data, often called a "knowledge graph" (Figure 3).

Figure 3.
A knowledge graph formed by persistent identifiers redirecting to landing pages that link to each other and to data via API calls.



If data providers register their lists of persistent identifiers with metadata catalogs, the links within the landing pages that the persistent identifiers direct to can be ingested and inferred automatically by computer programs. Thus, it is the combination of data hubs and these web publication practices that together construct the complete linked data architecture (Figure 1) that can help realize the vision of easily discovered, accessed, and used water sector data.

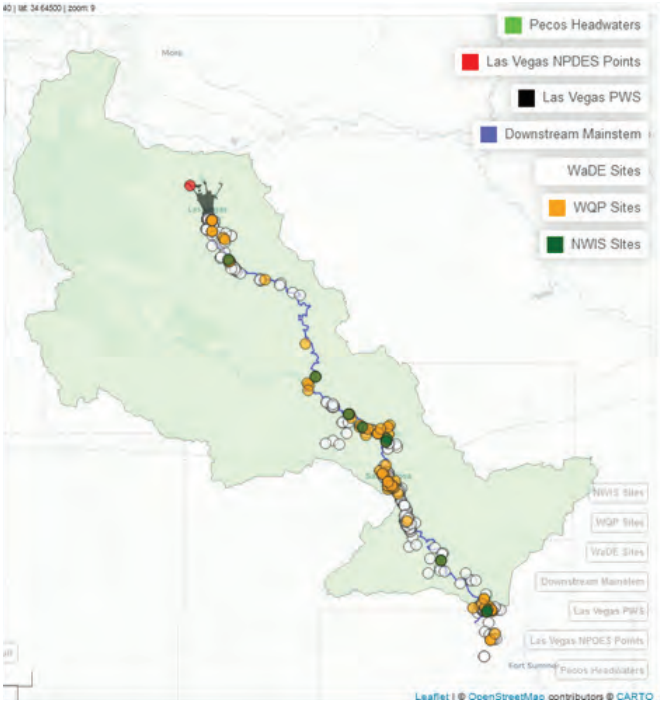
Foundational Linked Data Indexes

In order to ensure that the knowledge graph and data discovery process can be useful and orderly, a certain set of persistent identifiers and landing pages for common features is necessary for organizations to link their data to. In the water sector, two important foundational feature sets are administrative/statistical geographies and hydrography.

Administrative and statistical geographies provide important spatial context, linking together water infrastructure, other infrastructure, political control, demography, and so on. Thus, having persistent identifiers and landing pages for all of the various countries, states, territories, districts, sub-districts, municipalities, villages, etc. would enable data providers to link to these features unambiguously. Metadata catalogs could then automatically create lists of features from all data providers that are located within these geographies without any geoprocessing. The current most advanced system using this method is the Australian Location Index (Bastrakova and Crossman, 2020). The ***demonstration application*** allows a user to click any point within Australia and retrieve all data linked to that point spatially, including via census statistical geographies, local and state government boundaries, electoral districts, and delineated urbanized areas.

Hydrography refers to stream flowlines, watersheds, and aquifers. Features can be hydrologically related by being within the same watershed, withdrawing from the same aquifer, or being up/downstream of each other. Persistent identification of hydrographic features allows data to link to these features, although this first requires accurate lists of hydrographic features to be created. They are usually derived from elevation datasets that get simplified into stream line segments and watershed polygon vector geometries. The most advanced hydrographic indexing systems exist in the United States. The USEPA links many of its datasets (which can be queried with APIs) to the U.S. National Hydrography Dataset. The "***How's My Waterway***" system demonstrates how selecting a location can retrieve enormous amounts of information relevant to the watershed the location is within AND can be quickly organized. However, the USEPA only provides this facility for its own data. The USGS operates the ***Hydro-Network-Linked Data Index*** (NLDI), which allows any data provider to index their data to the national stream network and provides an API for the general public to search up or downstream of a point of interest, and retrieve a list of persistent identifiers linked to the relevant stream segments (Figure 4).

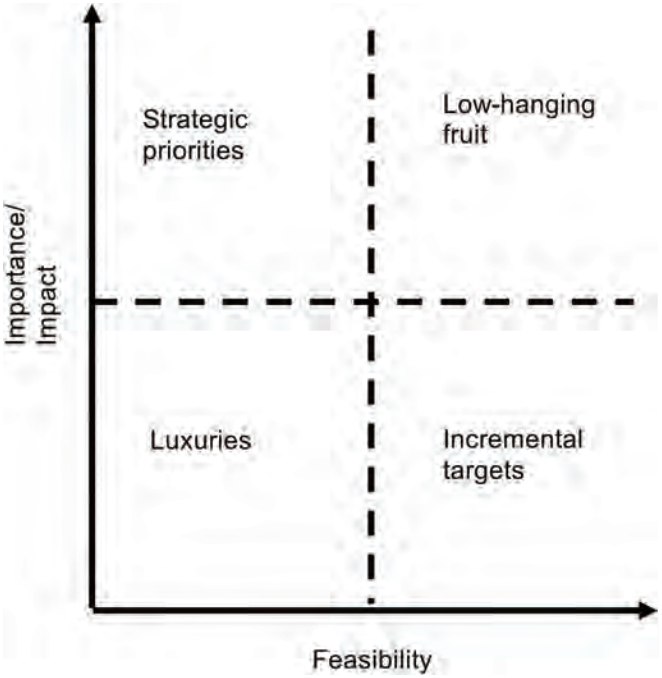
Figure 4.
Demonstration of the USGS Network Linked Data Index. Reproduced from *https://geoconnex.us/demo*



Priority activities and Investments

A useful way to think about priorities for investing in water data infrastructure for LMICs is to consider dual axes of relative importance and feasibility that together define four basic categories of activities (Figure 5).

Figure 5.
Investment prioritization framework.



Low-hanging fruit

Create and publish controlled vocabularies for the sector. There is a general lack of vocabulary standardization for data in the sector concerning the description of water and sanitation facilities. The sector should consider creating an open, community-driven vocabulary service similar to the AGROVOC of the FAO. The service could provide a forum for community members to suggest standard vocabulary terms that can then populate code lists for data hubs, reducing ambiguity in the interpretation of free-text type data entry fields. It is likely that many members will have different words for identical concepts. If the service adopts linked data best practices, then such terms could be published in ways that indicate that they are synonyms. Several enterprise and open-source platforms exist to manage vocabulary development and publication, so this is mostly a social and institution-building activity.

Encourage existing mature hubs with data services to adopt linked data best practices. The IATI, WPDx, and potentially TAHMO all provide mature data services. It would be only a marginal technical step forward for these hubs to adopt linked data practices, publishing feature-level landing pages for individual aid activity records, water points, and weather stations respectively. This would increase the web visibility of the data within these hubs and lay the groundwork for cross-linking data between hubs. In addition, the experience of these hubs can be used to help less mature hubs adopt these practices and avoid steps in their technical evolution that might make these practices more difficult to implement or maintain.

Encourage water quality sample data to be submitted to WQX. The USEPA WQX is a mature platform with a rigorous data standard to characterize water quality samples in the environment, facilities, and points of use. It is also institutionally supported for the long term by U.S. federal agencies, and requires no funding from the international donor community or LMIC governments. Any actors in the sector can be encouraged and trained to submit data to WQX. This data would then be available in standard formats to all without a significant investment in new IT infrastructure.

Establish administrative geography linked data indexing. Linked-data versions (including landing pages) of globally harmonized administrative boundary data would not be difficult to publish and maintain by such projects as ***GADM*** or ***geoBoundaries***. More fine-grained location data is already published in this manner by ***geonames***. It is even possible to publish linked-data administrative boundaries in such a way that other data hubs can automatically create links in their data to the relevant boundaries. Funding may be required to support these additional capabilities for hubs, however. WPDx and IATI, which already collect such information, could easily pilot this activity.

Strategic priorities

Establish persistent identification service and institutions. Persistent identification of data features is crucial for the linked data architecture to function sustainably. Creating a sector-wide registration agency within the *DOI* network is possible but could be very costly at the scale required. Technically, persistent identification is not difficult, with many open-source options available to implement such services (e.g., "PID Service – an Advanced Persistent Identifier Management Service for the Semantic Web," 2015). It is more complex to ensure that persistent identifier redirection can truly be maintained long term by a trusted entity with a secure funding model.

Upgrade existing less mature hubs to mature data services. IBNET and SIASAR in particular provide very important data in ways unsuitable for linked data. Significant investments would need to be made to change their data delivery systems.

Identify and address data gaps. The larger water sector community should be engaged to identify important gaps in thematic or regional data, and create new or expand existing hubs to fill these gaps. Some potential examples include sanitation facilities, water and sanitation provider service area boundaries, source water bodies/watersheds, and regional surface and groundwater quality monitoring.

Incremental targets

Establish a global hydro network-linked data index based on the USGS implementation. The USGS NLDI could be replicated using global hydrography datasets, such as MERIT-Hydro (Yamazaki et al., 2019). The USGS *code for the NLDI* is open source and could be applied to the global hydrography data if technical expertise could be retained to implement the port. Data hubs with fine spatial granularity like WPDx, WQX, and IATI could pilot linking to such an index, enabling useful upstream/downstream relationships to be identified and recorded between and among locations (e.g., water points with sources downstream of water quality sampling locations).

Create cross-links between existing hubs. WPDx, IATI, SIASAR, and IBNET all conceivably hold data that have important links. Aid projects finance water systems and water points. Water points are components of water systems. Particularly as these hubs adopt linked data practices, actual linked data should begin to be included. It is worth attempting to identify any IATI projects associated with WPDx points, for example, even before any linked data practices are implemented.

Luxuries

Expand hydrometeorological networks in LMICs. This can be quite expensive and is also already the subject of substantial World Meteorological Organization activity. However, persistent gaps exist and may be quite important to address in water-scarce and/or flood-prone areas.

Encourage hubs to adopt uniform API standards. In the EU INSPIRE program, a variety of social and environmental data can be visualized with very simple programs, since these data are all being delivered with the same structure of the *OGC SensorThings API* standards. However, this required copying wholesale these data sources into new standardized databases. Given other priorities, this additional functionality may not be worth the trouble. However, any new hubs created should adopt some of the standard APIs recommended in this document. This would allow newly created hubs to leapfrog the technical debt of more established ones and allow them to immediately interoperate with each other.

Conclusion

In this Think Piece, a vision for water data integration relevant to LMICs as easy, fast, and seamless as the data integration provided by large technology companies for their commercial offerings was offered. The technical requirements to achieve this vision were elaborated, and some specific potential investments to make progress were recommended, with reference to insights gathered from similar efforts in the water sector in high-income countries. Broadly, the think

piece advocates for a strategy of federated data hubs that can send and receive data to each other, and publish metadata in a way that commercial search engines as well as purpose-built tools can quickly harvest and organize. Many of the recommendations can be piloted incrementally and informally. With some key targeted investments, a positive spirit of experimentation, and an iterative development approach, substantial progress in improving data for decision making is certainly possible.

References

Bastrakova, Irina, and Shane Crossman. 2020. "Enabling Communities to Integrate Earth, Space and Environmental Data — Australian Location Index." Other. Informatics. *https://doi.org/10.1002/essoar.10501689.1*.

Blodgett, David, Emily Read, Jessica Lucido, Tad Slawecki, and Dwane Young. 2016. "An Analysis of Water Data Systems to Inform the Open Water Data Initiative." *JAWRA Journal of the American Water Resources Association* 52 (4): 845–58. *https://doi.org/10.1111/1752-1688.12417*.

Blodgett, David, Alistair Ritchie, Bruce Simons, Eric Boisvert, Abdelfettah Feliachi, and Sylvain Grellet. 2020. "Second Environmental Linked Features Experiment." OGC Public Engineering Report 20–067. Open Geost. *https://docs.ogc.org/per/20-067.html*.

Guha, R. V., Dan Brickley, and Steve Macbeth. 2016. "Schema.Org: Evolution of Structured Data on the Web." *Communications of the ACM* 59 (2): 44–51. *https://doi.org/10.1145/2844544*.

Josset, Laureline, Maura Allaire, Carolyn Hayek, James Rising, Chacko Thomas, and Upmanu Lall. 2019. "The U.S. Water Data Gap—A Survey of State-Level Water Data Platforms to Inform the Development of a National Water Portal." *Earth's Future* 7 (4): 433–49. *https://doi.org/10.1029/2018EF001063*.

mWater. 2021. "MWater." 2021. *https://www.mwater.co/*.

Patterson, Lauren, Martin Doyle, Kathy King, and David Monsma. 2017. "INTERNET OF WATER: Sharing and Integrating Water Data for Sustainability. A Report from the Aspen Institute Dialogue Series on Water Data." The Aspen Institute.

Patterson, Lauren, and Kyle Onda. 2020. "How Does the Internet of Water Work?" Internet of Water. *https://internetofwater.org/wp-content/uploads/2020/08/Architecture-Storyboard_FINAL.pdf*.

"PID Service – an Advanced Persistent Identifier Management Service for the Semantic Web." 2015. In *Weber, T., McPhee, M.J. and Anderssen, R.S. (Eds) MODSIM2015, 21st International Congress on Modelling and Simulation.* Modelling and Simulation Society of Australia and New Zealand. *https://doi.org/10.36334/MODSIM.2015.C8.golodoniuc*.

Schleidt, Kathi, Michael O'Grady, Sylvain Grellet, Abdelfettah Feliachi, and Hylke van der Schaaf. 2020. "ELFIE — The OGC Environmental Linked Features Interoperability Experiment." In *Environmental Software Systems.* Data Science in Action, edited by Ioannis N. Athanasiadis, Steven P. Frysinger, Gerald Schimak, and Willem Jan Knibbe, 554:188–93. IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing. *https://doi.org/10.1007/978-3-030-39815-6_18*.

Sporny, Manu, Gregg Kellogg, and Markus Lanthaler. 2014. "JSON-LD 1.0 A JSON-Based Serialization for Linked Data." W3C. *http://www.w3.org/TR/json-ld/*.

United Nations and World Bank. 2018. "Making Every Drop Count. An Agenda for Water Action." High-Level Panel on Water Outcome Report. United Nations. *https://sustainabledevelopment.un.org/content/documents/17825HLPW_Outcome.pdf*.

WHO & UNICEF. 2021. "Joint Monitoring Program." Washdata.Org. 2021. *https://washdata.org/*.

Yamazaki, Dai, Daiki Ikeshima, Jeison Sosa, Paul D. Bates, George H. Allen, and Tamlin M. Pavelsky. 2019. "MERIT Hydro: A High-Resolution Global Hydrography Map Based on Latest Topography Dataset." *Water Resources Research* 55 (6): 5053–73. *https://doi.org/10.1029/2019WR024873*.

Stanford's Program on Water, Health and Development is working to improve the health and well-being of communities by creating the knowledge, skills and solutions needed to support effective management of water and wastes, and to ensure sustained, equitable access to water supply and sanitation services.

This document was created thanks to the generous support of the Conrad N. Hilton Foundation.

Learn more at *water.stanford.edu*

Stanford Woods Institute for the Environment
Stanford University
Jerry Yang & Akiko Yamazaki Environment & Energy Building
473 Via Ortega, MC 4205, Stanford, CA 94305

environment@stanford.edu
woods.stanford.edu