Data Management Best Practices, Requirements, and Recommendations

Prepared and submitted by the Internet of Water with funding from the Cynthia & George Mitchell Foundation

Internet of Water

CONTENTS

INTRODUCTION

To be effective, water management professionals need quick and accurate information about the water sources they manage. How much water is there? How is it being used? What is its quality? In most cases, the data to answer these questions exists, but is often inaccessible or not formatted in a way that is easy to use. Modernizing our water data infrastructure is the key to unlocking this data and enabling effective water management.

In 2018, water experts and stakeholders in Texas came together to take the first steps toward water data modernization at a workshop at the University of Texas in Austin. The goal of the workshop was to develop a vision for "an ideal water data system for Texas." The system envisioned in that workshop was the jumping-off point for a collaborative project between the Internet of Water (IoW) and the Texas Water Development Board (TWDB) to develop the Texas Water Data Hub. This data hub will give data users easy access to many different types of standardized water data in one place. The hub will support real-time decision-making, enable the identification of opportunities to improve water security and provide decision-makers with a more complete picture of the water cycle in Texas.

As part of this project, the IoW has developed the following best practices and recommendations to support the development of the Texas Water Data Hub. While this document was developed for the Texas Water Data Hub it is broadly applicable and can be used to support the development of other hubs in the future.

BEST PRACTICES & RECOMMENDATIONS

MAKE YOUR DATA DISCOVERABLE

A key step in modernizing water data infrastructure is to improve the findability of water data. Making data easy to locate means it is more widely available for a broad range of purposes. If water management professionals don't know that a particular dataset exists or where to find it, they cannot use it to improve analysis. Because water today is so fragmented, experienced government employees and knowledgeable researchers often spend hours of valuable time, just searching for the data they need before they can even begin to answer their questions.

Increasing data findability is contingent upon three components:

1. A thorough data inventory,
2. The development of associated landing pages,
3. Assigning persistent identifiers to datasets that link them with keywords or to geographic locations.

BUILD A DATA INVENTORY

How do you create a data inventory? There are several resources available detailing how to conduct a data inventory, such as the Data Inventory Guide at John Hopkins. Below is an abbreviated version, tailored for water data.

STEP 1. ESTABLISH AN OVERSIGHT AUTHORITY.

It is very helpful to have a designated person taking responsibility for and leading the effort. The person may develop a committee to help guide and shepherd the process, especially if the inventory will be conducted across multiple organizations. In

STEP 2. DETERMINE THE SCOPE OF THE DATA INVENTORY.

Data inventories can vary dramatically based on scope and depth. For example, an agency might start by inventorying all water-related data, or they might prioritize a specific type of data, such as groundwater data, snow data, spatial data, etc. The inventory may include basic elements (described below) or they might require opening each dataset to understand the frequency of collection and/or the geographic and temporal coverage.

Data are continually collected, and new datasets are developed. Just as a business must conduct continual inventories of their assets, agencies should regularly conduct data inventories. The oversight authority should consider the long-term process for conducting data inventories and how those data inventories and their corresponding datasets will be integrated into a resulting data catalog.

STEP 3. PROVIDE TEMPLATES AND GUIDANCE FOR THE INVENTORY.

Categories are helpful ways to intuitively group and search for data in a catalog. For water data, the categories could be groundwater, surface water, climate, administrative, etc. These are often the "topics" or "categories" present in a data catalog. Categories should resonate with stakeholders to maximize data usage.

For example, in Table 1, the categories reflect the mission of the agency and the types of data they collect. The Water Control Board and Departments of Environmental Quality or Protection have categories for monitoring, regulation, compliance, and remediation. Departments of natural resources oversee more than water and tend to have categories such as air, climate, ecosystems, land, and water (as a whole category). Establishing

Table 1: Examples of water-related categories associated with data catalogs.

Category	AZ DWR	CA Water Boards	NM Water Data	NC DEQ	PA DEP
Agriculture	▨				
Boundaries	▨			▨	
Climate			▨		
Contaminants		▨			
Drinking Water; Public Water Supply		▨			▨
Ecosystems & Wildlife			▨		
Education/Outreach				▨	
Energy or Oil & Gas			▨	▨	▨
Facilities		▨			
Funding		▨			
Groundwater; Wells	▨	▨			
Infrastructure			▨	▨	
Land				▨	
Monitoring				▨	
Natural Hazards			▨		
Open Data and Maps		▨			
Regulatory; Compliance & Enforcement		▨		▨	
Recovery, Restoration, Remediation				▨	
Surface Water or Streams & Lakes	▨	▨			▨
Waste Management				▨	▨
Water; Water Resources				▨	▨
Water Conservation		▨			
Water Planning			▨		
Water Quality			▨		
Water Quantity			▨		
Number of Categories	4	9	8	10	5

broad categories is important in tagging and classifying data for easy discovery.

In Texas, a stakeholder engagement process generated a two-tiered system of data categories and tags (additional data topic, format, or quality descriptors) to allow for search indexing at multiple levels of detail.

STEP 4. CATALOG DATA ASSETS

Inventoried data should be assigned metadata, which will allow the data to be cataloged. Just like any library, a catalog allows users to search for data using keywords or other categories (such as geographic location or data type) to locate all data in the catalog related to their area of interest. With an ever-growing number of datasets available, it is more important than ever to organize datasets in a catalog so that users can quickly find all data relevant to their project.

Once data is inventoried and cataloged, the next step in improving data findability is creating landing pages for each dataset that describe the dataset and expose it to search indexes so that the data can be found in search engines.

DEVELOP LANDING CONTENT

An important way to make data findable on the internet is to publish human-readable “landing content” about it on a “landing page”. People find resources on the internet that are in the form of web pages that are linked to other web pages. A landing page is a stand-alone web page, typically focused on a single topic. This is unlike a web page that has many goals and encourages exploration throughout a website that contains multiple pages. Landing pages have a single focus, making them useful for improving data discoverability by connecting the data to a single URL via the landing page.

By linking data sets to a landing page, when a user searches for data using keywords or geographic locations, the metadata and identifiers embedded in the landing page will result in a successful search. Additional information can also be included on landing pages, such as dataset metadata and links to other related data. It is best practice to assign landing pages persistent identifiers.

These identifiers are important because often organizations and agencies make changes to their websites and landing pages, resulting in a broken link when a user attempts to locate the page. Have you ever searched for a link and received an Error 404 code? This means the original location, referred to in the link, no longer exists. The owner of that information may have simply moved it to a different page. Persistent identifiers, linked to landing pages, ensure that users are automatically redirected to the current address/URL for the landing page, regardless of the changes made at the organizational or agency level.

MINT PERSISTENT IDENTIFIERS

To connect datasets and increase their findability, Persistent Identifiers must be associated with each dataset. Persistent Identifiers (PIDs) are unique values attached to an individual object, person, institution, or dataset, and are a critical component for improving findability, fostering transparency, and ensuring collaborative parties are referring to the same datasets. Persistent identifiers are frequently used in everyday life. Every person in the U.S. has a persistent identifier: the federal government assigns each person a social security number (SSN), unique to that person, when he or she is born. Social security numbers allow the government and other entities (such as banks, schools, and license branches) to ensure they create services for the correct person and are the main way the federal government identifies individuals. Creating and attaching persistent identifiers to datasets ensures they can be located and ensures that the user locates the intended dataset.

The most common way to mint persistent identifiers for datasets is to use the Digital Object Identifier system through becoming a member of a DOI registration agency like a DataCite or CrossRef. If the expense associated with membership in these agencies is impractical, there are free options, including:

PURL: Persistent URLs (PURL) provide permanent addresses for resources. The client searches for the PURL and the PURL redirects the client to the current URL holding the data object to the client using a standard HTTP redirect. PURL allows a persistent identifier for the URL while the data provider may change their web address. If the provider does not update the PURL, the redirection will fail. PURLs are provided by the Internet Archive, a 501(c)3 nonprofit.

Geoconnex: Geoconnex is an Internet of Water linked data project that includes a free persistent identifier service similar to PURL. The project also provides facilities to link resources with geoconnex identifiers to other ones. Geoconnex is based on YOURLS, an open-source URL redirector software that is another option for organizations wishing to implement their own persistent identification infrastructure.

MAKE YOUR DATA ACCESSIBLE AND INTEROPERABLE

Improvements in data collection technologies have greatly increased water data coverage but have also introduced new challenges in data acquisition and processing. Today greater volumes of water data, are being collected at higher frequencies, by a greater number of organizations than ever before. However, because this data is often stored in hard-to-access formats, the amount of time it would take to process the data prevents many would-be users from accessing it. Improved data accessibility will enable us to realize the benefits of improved data collection technologies.

BUILD PROCESSES TO INGEST, STANDARDIZE AND SERVE DATA ON THE WEB.

The Internet of Water recommends that organizations begin stewarding their data as a service. This means offering datasets that are standardized by topic, continuously maintained, and available for download on-demand. Data services make data available in ways that users and software developers can access it easily in the context of a national, integrated water framework. Building a data service involves implementing:

1. ingestion mechanisms from data producers;
2. wrapping mechanisms that standardize disparate data from many providers;
3. data stores of standardized data; and
4. catalogs providing information about the data contained within the data stores and access points to the data.

WHAT ARE DATA WRAPPERS?

A data wrapper is a social and technical process that performs ETL functions. ETL stands for extract, transform, and load. A data wrapper extracts data from data producers, transforms that data by cleaning it and placing it in the proper formats, and loads it into the target datastore. Just as data are being produced at record levels, data are being produced by larger numbers of entities, each with its own formatting practices.

Data must be transformed into standard formats so that data can be integrated with other datasets without users needing to process the data for each different use.

There are essentially three types of data wrappers that data producers might interact with: (1) data submission templates, (2) direct data entry forms, and (3) fully automated processes.

1. Templates are intermediate data files. A familiar form of a template is an Excel file with prescribed column headers. Data producers would be expected to fill out the template, essentially copy-and-pasting data manually or with an automated script from their existing datasheets. Producers may then send their completed template files to target email addresses or upload them manually to a website, share drive, or similar location. Automated scripts designed to work on correctly completed templates then copy or ETL this data into the data store. An example of this type of data wrapper is the WQX Web template used to submit data to the Water Quality Portal. Templates are common for data hubs that lack the expertise

or capacity to build and maintain more complex infrastructure.

2. They are also common for data hubs that involve very disparate data producers with different practices and strict IT protocols that prevent direct system integrations from submitting data to a commonplace. This is also somewhat common in contexts involving sensors with data loggers, where data logs must be manually copied from sensor internal storage. In this case, the template(s) could largely be based on the given data logger native format(s).
3. Direct data entry forms involve the data producer staff manually entering data as it is being produced directly into the data store. Familiar examples of this kind of data wrapper are mobile/ web surveys used by the U.S. Census or Electronic Medical Records systems. Data entry forms are common when the data hub can exert direct control over the data collection process, as is usually the case for data hubs that are confined to only one organization, data hubs purpose-built around a singular data collection project such as an NGO environmental monitoring program, or regulatory reporting contexts.
4. Fully automated processes involve data from producers being automatically ingested by infrastructure operated by the data hub. This is common in internet of things (IoT) contexts, where sensors can send data directly to databases over radio or wireless internet services. This is also common when data producers can serve their data out with their own API. As long as the API is well-documented, it is possible for the data hub to write computer programs that automatically request new data from the API and copy it to its own data store. If all data producers serve their data with a limited number of common, standard APIs, the data hub can be scaled to encompass a very large number of data producers.

WHAT ARE DATA STORES?

A datastore is where data physically live. Datastores have many definitions in the IT sector. In the context of the Internet of Water data hubs, data stores refer to the physical location where standardized data are stored and from which data is accessed by users. There are many kinds of data stores and many ways to categorize them. Table 2 provides a categorization, a description of the appropriate use of each category, implementing technologies for both on-premises and cloud-based solutions, and links to real-world examples.

“Datastore” may refer to the software holding data (such as Microsoft Excel or Access, Oracle Database, SQL server, etc.) in a centralized database or on a cloud-based platform. The same data may be stored in one format by the data provider and in another format by the agency sharing data with the public. Most agencies are familiar with centralized data storage and infrastructure. In this model, an agency purchases on-premise storage and determines the number of servers needed, cooling and energy costs, backup, and physical security of the data.

In contrast, in a cloud model, the infrastructure and server costs are outsourced to a third party (i.e., you rent the physical infrastructure for data storage). Agencies may rent physical infrastructure to store data only (infrastructure as a service), or they may rent operating systems (platform as a service), and software (software as a service) as well.

Table 2: Data Store types and examples

Type	Appropriate Use	On-premises/ Intranet	Cloud	Example
Singular file	Managing a single small data collection project and sharing the resulting dataset	An excel file	Google sheets	NM Energy, Minerals & Natural Resources Department Coal Water Quality Dataset
Object Storage/ Data Lake	Managing and sharing multiple types of data, including unstructured data.	Intranet Share Drives (e.g. Z:// network drive)	Amazon S3 Google Cloud Storage, Azure Blob Storage, Dropbox, Box, Google Drive	California Natural Resources Agency Open Data
Database	Managing a single complex, structured, standardized dataset that undergoes real-time updates	Local instance of Oracle, MS SQL Server, MySQL, PostgreSQL	Amazon RDS Google Cloud SQL Microsoft Azure SQL	NM Office of the State Engineer Water Rights Reporting System
Data Warehouse	Integrating multiple datasets on different topics for real-time processing and analytics	Multiple local database instances	Snowflake Amazon Redshift Google BigQuery Databricks	EPA Envirofacts

WHAT ARE DATA CATALOGS?

Many water agencies do not currently provide a way for the public to discover or access data on their own, even though much of the data is legally public information. Agency staff, therefore, spend a lot of time addressing Freedom of Information Act (FOIA) requests, manually updating html tables online or ftp datasets, or writing and debugging queries from internal databases. Data catalogs are searchable, structured collections of metadata that serve direct users the data they are interested in. They can considerably improve public access to data, provided they also incorporate appropriate data governance to enforce and maintain the quality of the included data and metadata. Internet of Water data hubs consist of catalogs that provide metadata about their datastore and include only standardized data for the type of data in that hub (e.g., time-series streamflow measurement data, groundwater level data, etc.). Data catalogs are the primary way that data users interact with data hubs. Some examples of data catalog software are detailed in Table 3.

Data catalog administrators, at a minimum, should enforce a metadata standard on included datasets and publish the metadata in machine-readable formats. The process described below employs the widely used Data Catalog Vocabulary (DCAT). DCAT was developed in the context of government data to provide a standardized vocabulary for publishing data catalogs.

Many local and state agencies use Esri's ArcGIS catalogs, which are compatible with DCAT. DCAT includes metadata standards for catalogs, datasets, and data services. The use of standardized metadata (and vocabulary) increases the findability of datasets and data services across hubs (other catalogs).

It takes time and effort to adapt to the format of the DCAT. DCAT is both the name of the metadata standard and is the "namespace" that tells machines we are using DCAT metadata to structure the catalog. This is why you see "dcat:xxx". DCAT has standards to help structure the catalog, locate datasets based on a set of details, as well as information

Table 3: Data catalog software

Software	Cost structure	Include water-sector specific data management & quality control	Integrated data visualizations	Public API for data access	Standard APIs supported	Example
CKAN	Free/ Open Source	No	No	Yes		https://newmexicowaterdata.org/
Vendor-Managed CKAN Datopian	Subscription	No	No	Yes		https://en.energinet.dk/Electricity/Energy-data
Socrata	Subscription	No	Yes	Yes		https://data.colorado.gov/
ESRI ArcGIS Hub	Subscription	No	No	Yes	WFS, WMS	https://gis.data.ca.gov/
Geonode	Free/ Open Source	No	Yes	Yes	WCS, WFS, WMS	https://www.geonode-gfdrrlab.org/
Clowder	Free/ Open Source	No	No	Yes		https://terraref.org/
Kisters	Per-user License or Subscription	Yes	Yes	add-on	SOS	http://portal.gemstat.org/
Aquatic Informatics	Per-user License or Subscription	Yes	Yes	add-on		https://idastream.idahopower.com/

on how to download the data. Within DCAT, other structured metadata standards exist. “dct” is the namespace signifying a dataset. Within a dataset, key information is provided such as the title, description, keywords, publisher, date last modified, license information, spatial coverage, and temporal coverage. There are standards around each of the details (such as date format). These details should be collected during the data inventory process.

Google has also been actively developing schemas to make data easily searchable via their search engine; however, schema. org does not fully adhere to international metadata standards for describing geospatial datasets. We recommend adopting standards for geospatial data since all water data are attached to a location. ISO is an international organization that develops standards that describe the best way to make a product, manage a process, deliver services, and even share data.

These standards are communally developed by subject matter experts and reflect the collective wisdom on the best way to do something. This standard-setting organization developed standard ISO 19115 in 2003 as an internationally adopted schema for describing geographic information and services regarding the extent, quality, spatial details, temporal details, and distribution of geographic data. DCAT general-purpose standard for project metadata includes ISO 19115 standards for describing the geospatial datasets. For this reason, we recommend using DCAT, however, there is an ISO 19115 crosswalk between DCAT and schema.org metadata templates should you choose to adopt schema.org templates. The metadata collected in this inventory will be used to populate the metadata fields for whichever catalog you later choose to make your data discoverable and accessible.

In the preview template below we only provide columns for describing the dataset (not including distribution information), and we only include some of the attributes provided by DCAT. The full range of lists, and details about each list, can be found here. Additionally, the Federal Geographic Data Committee provides a suite of tools for creating metadata, as does the USGS.

Each agency should fill out details about the dataset, while the oversight authority or the agency responsible for setting up the data catalog may later provide details regarding persistent identifiers (unique ID; for the importance of unique ids see persistent identifiers section), distribution

information, APIs, and other attributes relative to cataloging and distributing the dataset. This would be akin to a seller providing all the details about their product to Amazon and Amazon adding a rating system and product code to each item provided by the producer.

You may include additional columns in your inventory that are not present in the metadata template. For example, you may want to know if a data dictionary exists and where it is located, what data standards are used, and a sense of data quality. We recommend providing a finalized metadata template for organizations to fill out. This will lower the amount of work required by the coordinating agency to reconcile any discrepancies.

Table 4: Preview of template for datasets

DCAT Title	dct:title	dct:description	dcat:theme	dcat: keyword	dct: identifier
Column Header	Title	Description	Category	Keyword	Unique ID
Description	name of dataset	describe dataset	the main category of the resource. A resource can have multiple themes separated by commas.	in dcat, this is a singular word	a unique identifier for the dataset
Data Type	text	text	text	text	text
Recommended Standard	none	none	none	none	http URI
Priority	essential	essential	optional	essential	optional
Example	streamflow	Daily records of streamflow from 12 gages in xxxx.	Surface water	Streamflow, discharge	https://doi.123

MAKE YOUR DATA USABLE AND INTEROPERABLE

Data fragmentation makes data difficult to find and access. However, even after a data user finds and downloads data, there is still much work required to properly process the data to make it usable. People often refer to the “big data” needed to solve our stickiest questions. However, “big data” is dependent upon structured, quality “small data.” In other words, building big datasets to answer our most challenging questions begins with integrating smaller data sets from a variety of different sources, often referencing different locations, collected using different methods, across different timescales, and often in different formats.

The degree to which data can be integrated is often called interoperability. The successful integration of these data depends upon data and metadata standards, agreed upon by each of the agencies and organizations providing data. Data that follow different standards or lack proper metadata are difficult to integrate. Without data standards and metadata, data users are left with endless columns and rows of data without a clear understanding of how data from one source is connected or related to data from another source.

INCLUDE DETAILED METADATA

Data usability is heavily dependent upon detailed metadata. Metadata are data that describe data. Metadata allow users to search for data using key terms or categories, such as type, title, subject, geographic area, author, etc. However, metadata is also necessary to be able to send and receive data. It defines the file formats, the data transfer interfaces, and the download links. Metadata defines how data is administered. Who has access to the data? How should the data be cited? Who can a user contact if he or she has questions about the data? These answers are all defined in the metadata.

Metadata gives the user important information about how to use the data as well. What was the original purpose of the data? How was the data collected? What is the quality of the data? And very importantly, what do the column headings mean? All of this information is necessary for the user to better understand the data and determine if a dataset is useful for a specific purpose.

Dataset metadata describes the data in a specific dataset. Adopting a metadata standard means all datasets are described in similar details across an agency or organization. If agencies across a state or region use the same metadata standards for describing their datasets, then users can quickly and consistently locate and access those data, and easily integrate them with other datasets.

Once data are properly tagged with metadata, the metadata can be stored and published in a data catalog. In the data catalog, metadata allows the sharing of data or information among or between entities.

Metadata Standards include:

- Schema: A metadata schema is a list of elements that should be included in the metadata and the rules that govern how these elements should be used. For example, the Dublin Core Metadata Initiative defines a schema with 15 core elements for metadata, including categories such as creator, data format, language, and many others.
- Standards: There are three components to forming metadata standards: content, value, and structure. Content standards define the metadata elements and how they should be generated. Value standards restrict the elements to allowable words in a controlled vocabulary (e.g. ISO 19115 topics) or data types (e.g. "integers", or "character string"). Structure standards define the format required for the metadata to ensure it is machine-readable (Ex: XML, JSON, JSON-LD).
- Data Dictionary: A data dictionary is a human-readable description of the data

attributes and their allowed values (i.e. the content and value standards of the data being described by the metadata). For example, a dictionary may describe the “decade” attribute, as a “4-digit number representing the first calendar year of a decade.” This means that a data entry of “2000” in the column “decade” would indicate that this is data collected from January 1, 2000, to December 31, 2000.

Some examples of data of metadata standards can be found at:

- Dublin Core (for general dataset description)
- ISO 19115 (for geospatial data description)
- FGDC Standards (Federal Geographic Data Committee)
- Schema.org (for describing the content of websites – including dataset landing pages – for search engines)

Some examples of tools and guides to create metadata include:

- ESRI Geoportal (open-source metadata catalog with metadata editor)
- Geometa (R package)
- Dublin Core Guide

Some examples of controlled vocabularies and data value lists relevant to water data can be found at:

- ODM2 Controlled Vocabularies for Hydrologic Observations
- NEMI Catalog for Water Quality Analytes and Methods
- QUDT Vocabularies for Quantities, Units, Dimensions, and Data Types

FOLLOW ESTABLISHED DATA STANDARDS

Data standards are essential for data sharing. While individual datasets are valuable, the most value is realized when datasets can be integrated with other datasets to create a complete representation of complex features or systems. Integrating data from a river system allows us to examine the entire river system (upstream and downstream effects) rather than a single component of that system.

Like metadata standards, data standards include:

- Schema (lists of required elements)
- Content standards (definitions and descriptions of schema elements)
- Value standards
- Restrictions on what should populate schema elements (e.g. numeric, text, dates/ times in particular formats
- Controlled vocabularies (lists of allowed terms)
- Structure standards (definitions for how data should be encoded for machine-readability. General examples include Shapefile, XML, and JSON)

Table 5 provides links to data standards that have been created and vetted in collaborative processes to represent different themes of water-relevant data in the United States.

Table 5: Topical Data Schema and Content Standards

Topic	Schema/Content Standards
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
Utilities and Infrastructure	FGDC Utilities

Table 6 provides links to data structure standards for general data types that are open source (non-proprietary), widely adopted, and have been shown to be compatible with most common data transfer, analysis, and visualization frameworks. Other standards exist that meet more niche use cases or may be required for certain proprietary software packages.

Table 6: General data structure standards for data sharing

Data Type	Structure Standard
Geospatial Vector (point, polygon, etc.)	GeoJSON, GML, GeoPackage
Geospatial Raster (matrix, image, etc.)	GeoTIFF, netCDF
Tabular data (general)	CSV, JSON, netCDF, Tabular DataPackage
Nested data	JSON, XML

Table 7 provides links to open API standards for data types and topics, as well as open-source and proprietary software that include options to share data using those APIs.

Table 7: Open API Standards by data type

Data Type	API Standards	Open-Source Implementations	Proprietary Implementations
Geospatial vector	WFS	Geoserver, QGIS, MapServer	ESRI ArcGIS Server, CubeWerx
	OGC Features	Geoserver, PyGeoAPI, QGIS	CubeWerx
Geospatial raster	WCS	Geoserver, QGIS, MapServer	ESRI
Map imagery	WMS, WMTS	Geoserver, QGIS, MapServer	ESRI, CubeWerx
Georeferenced observations/ time series/ samples	SOS	52North, istSOS	Kisters, SensorUp
	SensorThings API	FROST, GOST, 52North	SensorUp
Tabular data	Data Package	OKI, CKAN	NA

WHAT IS AN API?

An Application Programming Interface, or API, is a set of functions and procedures allowing the creation of applications that access the features or data of an operating system, application, or other service. There are many APIs. Many data management systems offer custom APIs that allow for data sharing. 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 would need to write custom code to access data from each API.

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. 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. For example, the OGC WMS is an API standard for delivering web map images from geospatial databases. Products as diverse as ESRI ArcGIS, the open-source Geoserver, and Oracle databases include WMS as an optional interface to query and receive web map imagery, but hundreds of different software clients can import data made available through this interface without difficulty.

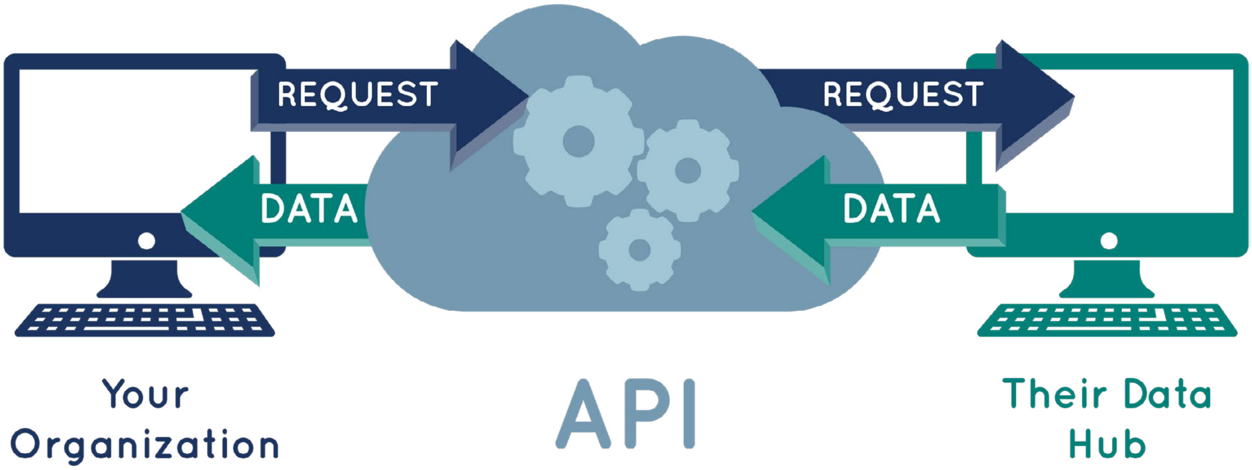

Figure 1: What is an API

The Internet of Water recommends wherever possible, to make data available using APIs compliant with API standards relevant to a particular data type. In particular, we recommend open, community standards that are not associated with any particular vendor data system, and that have been created and are maintained in open collaborative processes. The API standards published by the Open Geospatial Consortium are good examples of such standards.