How does the Internet of Water work?

Last Updated March 20, 2020

The Internet of Water (IoW) is designed to address the challenges of fragmented water data over time and space. While countless entities are collecting water data like individual pieces of a puzzle, much of these data are not findable, accessible, or usable to create a coherent strategy for managing water resources across shared aquifers or watersheds. Here, we describe how the IoW is designed to help find, access, and use data collected by different agencies with different purposes, standards, formats, hosted on different platforms, and so on. This framework will allow users to find more pieces of the puzzle when attempting to build a more complete picture for what is happening within their aquifer and/or watershed.

Why do we need better data management for water?

Conventional wisdom says *you can't manage what you don't measure.* There are countless entities measuring water resources for management and supply purposes. A grower is measuring water levels in their wells, a drinking water utility is measuring the quality of the water pulled from the river, a wastewater utility is measuring the quality of water being discharged, a federal agency collects data on snowpack while another collects data on streamflow, and so on. Many entities use the same water flowing through a watershed or stored within an aquifer, and each of these entities holds different pieces of data and makes decisions that impact the water that is accessed and used by many on a daily basis.

We can think of each piece of data as a piece of a larger puzzle that, in aggregate, depicts the present and historic states of our watersheds and aquifers. Currently, each of us holds a couple of puzzle pieces (data) and makes the best decisions we can with the information we have. We might be able to put together small sections of a puzzle, but because we may only be able to find, access, and use a fraction of the data available for a watershed or aquifer, we are not able to clearly or completely see what we are looking at. It is only as we find and put in place more pieces of the puzzle that we are able to build a high resolution, coherent picture of our water bodies and hydrological systems (Figure 1).

Assembling a complete picture of a watershed or aquifer is easier said than done. For a variety of reasons, water data is fragmented. The Internet of Water (IoW) is conducting federal and state-by-state inventories to identify which public agencies are collecting different types of water data. These inventories provide a snapshot of water data that is available or referenced online. Our 2018 federal government inventory alone showed 42 agencies had water within their mission and water data were provided on 56 different platforms. Similarly, states that have been inventoried all have their own agencies collecting different types of water data and, when that data is available online, have made it accessible through different platforms.

The widespread fragmentation of water data raises the following question: what will it take to integrate water data within the United States?

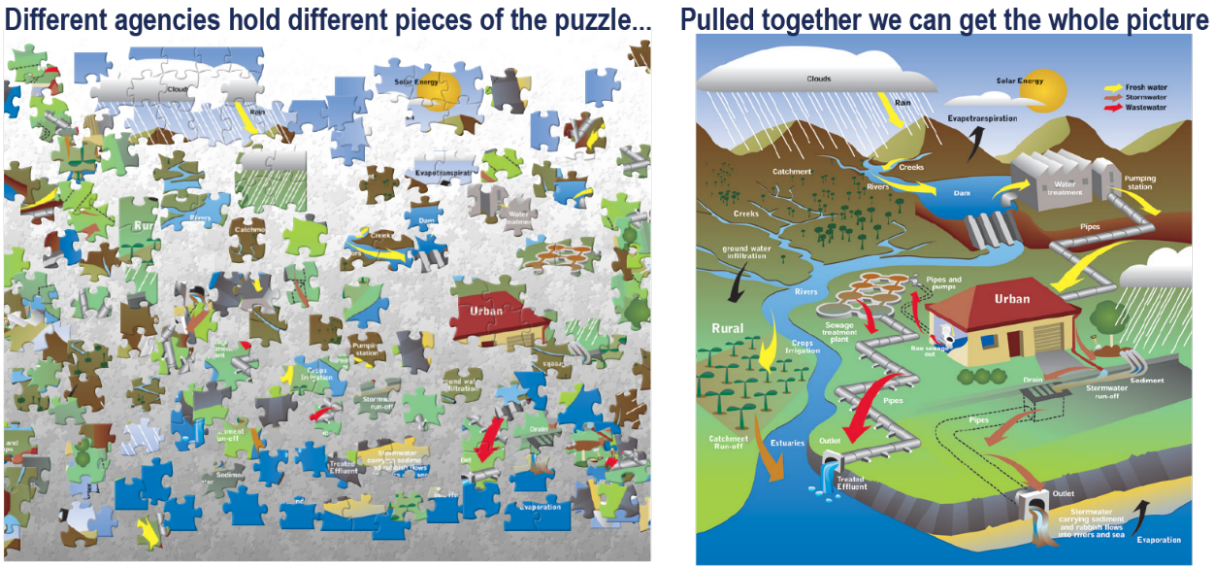


Figure 1: Fragmented data "puzzle" transforms into a complete picture when we integrate puzzle pieces.

How the Internet of Water Works

The 2017 report from the Aspen Institute Dialogue Series on Water Data, *Internet of Water: Sharing and Integrating Water Data for Sustainability,* stated that the architecture for an internet of water, "in which open public water data would be shared through a network of communities," can best address water data fragmentation across space as a federation of data producers, hubs, and users. The data flow more readily between producers, hubs, and users as they become more FAIR: findable, accessible, interoperable, and reusable.

- **Findable** – we know where the data are located (puzzle pieces)
- **Accessible** – data can be obtained when producers choose to share (we can get puzzle pieces)
- **Interoperable** – data have standards or metadata that allow the data to be used and connected to other data correctly (the puzzle pieces can fit together)
- **Reusable** – many users over time can create value from the same data (with the same puzzle pieces we can create the same puzzle)

In order to create a network of communities that allows data producers to maintain ownership of their data, the Internet of Water is promoting and strengthening the connections between these individual data providers.

The Internet of Water is composed of the following components:

- **Data Hub**: a formalized, structured, source of open water data.
- **Data Producer**: an entity that collects data.
- **Data Provider**: an entity that publishes data, either a hub or a data producer.
- **Data User**: an entity, private or public, involved in accessing and investigating data.

Data Producers

Data producers collect data to meet their specific needs and hold the pieces of the puzzle. They may be public agencies, private companies, researchers, non-governmental organizations, or community-based organizations. Producers collect different types of data using different methods and different standards. Some data producers may be willing to share their raw, or aggregated and anonymized data, to others for secondary uses. Shared data can be sent to an IoW data hub for users to then access.

Data Hubs

Data hubs are the key to pulling the pieces of the puzzle together to create FAIR data. An IoW hub is a data hub that contains four essential components to make data FAIR: data producers, wrappers, data store, and metadata catalog.

- **Data Producers** – must share their data or metadata with the hub (shares puzzle pieces)
- **Data Wrappers** – convert raw data into a standardized format (makes sure pieces fit together)
- **Data Store** – persistently stores the standardized data (puts standardized pieces in a box)
- **Metadata Catalog** – points to data within your store (finds which puzzle holds your pieces)

An IoW Hub must have all four components. Most water data hubs today are Non-IoW hubs. Meaning they host water data that adhere to some FAIR principles and have some components of an IoW hub, but not all. Most often these hubs have the data producer and data store but lack the wrapper and/or a metadata catalog. Wrappers are essential for data users to put the data to correct use with other data across a region. The metadata catalog is essential for finding data within and, eventually, across hubs. Non-IoW hubs can still participate in an IoW by adding direct web page indexing or registering their catalog with the IoW.

The responsibility for providing and managing the four components of an IoW hub varies. Currently four primary types of IoW hubs exist, as well as some hybrids. At one end of the spectrum, data producers are primarily responsible for all four components with the exception of the metadata catalog (Type A: Distributed). At the other end of the spectrum, a centralized hub organization is responsible for all

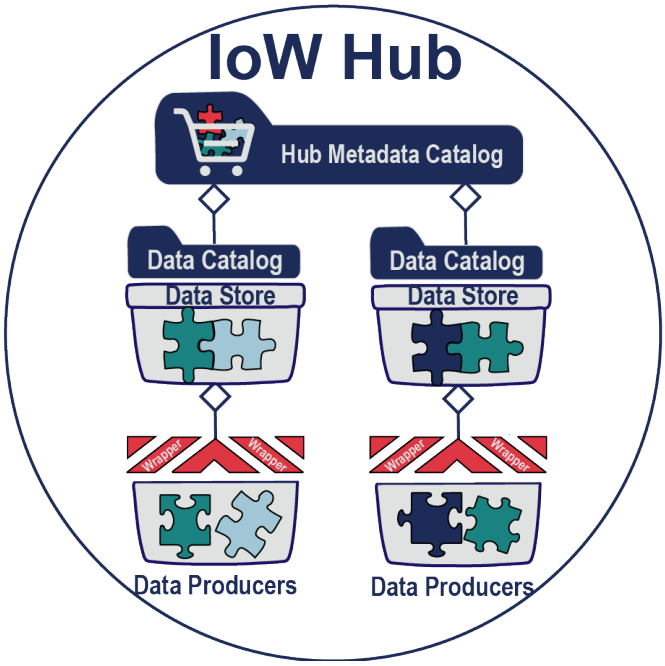
components of the IoW hub, with the exception of collecting the raw data (Type D: Centralized). Type B and Type C hubs are some combination of these extremes.

- **Type A: Distributed** – producers are responsible for all components except the metadata catalog.
- **Type B: Blended, Producers Push to Hubs** – the components are shared between producers and hubs, but producers are responsible to push data to hubs
- **Type C: Blended, Hubs Pull from Producers** – the components are shared between producers and hubs, but hubs are responsible to pull data from producers.
- **Type D: Centralized** – hubs are responsible for all components except data collection.

Type A: Distributed Hub

In a distributed hub, data are collected and converted into a standardized format that are *pulled* into a local data store with a catalog that points to the hub's main metadata catalog. The hub metadata catalog can search through local catalogs to pull queried data in real time.

Type A: Distributed



Ideal For: managing high volumes of similar data types in real time

Advantages: low computation and storage requirements for the hub

Limitations: user query is limited

Barriers: requires significant capacity for all data producers

Examples: EPA Interoperable Watersheds Network; CUAHSI Hydroclient; WaDE 1.0

Type B: Blended, Producers Push to Hubs

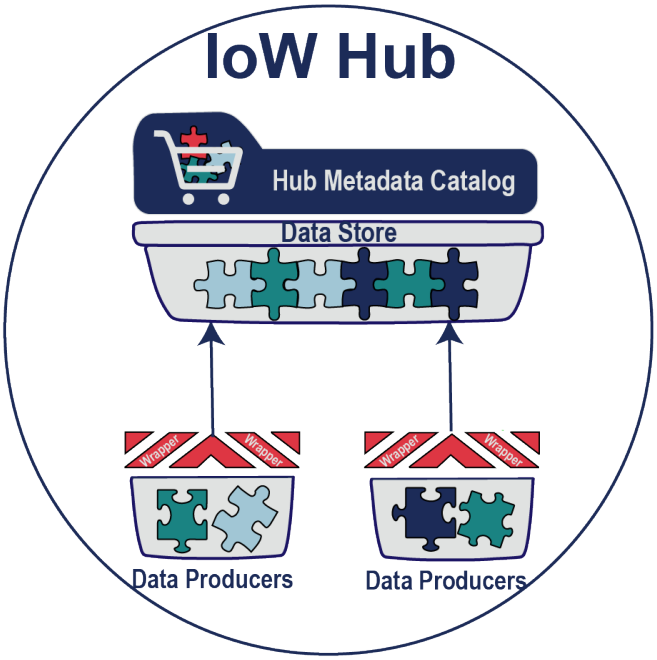
Data are collected and converted into a standardized format that are *pushed* by the producers into a centralized data store managed by the hub.

Ideal For: regulatory data collected at a daily or higher (sub-daily) frequency

Advantages: users can make complex queries; accessible when data producers are offline; hubs can ensure data producers meets certain standards

Limitations: large computation and storage requirements for the hub

Type B: Blended, Producers Push to Hubs



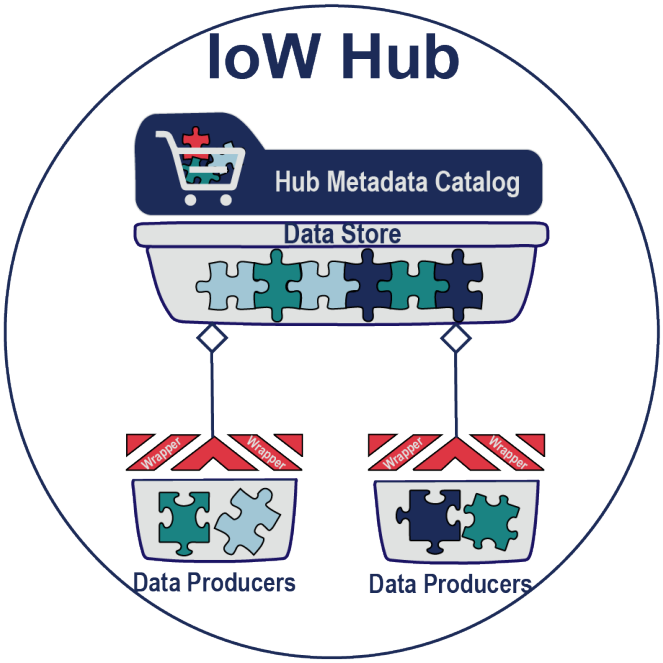
Barriers: requires producer capacity to wrap data and agreement to share data

Examples: Water Quality Portal; USGS NWIS; Reclamation Water and Information System

Type C: Blended, Hubs Pull from Producers

In Type C, data is then *pulled* by the hub into its centralized data store, rather than *pushed* by local entities as in Type B. This may address some of the limitations and barriers to Type B in terms of local capacity and burden.

Type C: Blended, Hubs Pull from Producers



Ideal For: non-regulatory data collected by multiple producers with varying capacity

Advantages: users can make complex queries; accessible when data producers are offline

Limitations: large computation and storage requirements for the hub; hubs have less control on ensuring data standards

Barriers: requires producer capacity to wrap data and agreement to share data

Examples: NOAA Integrated Ocean Observing System

Type D: Centralized

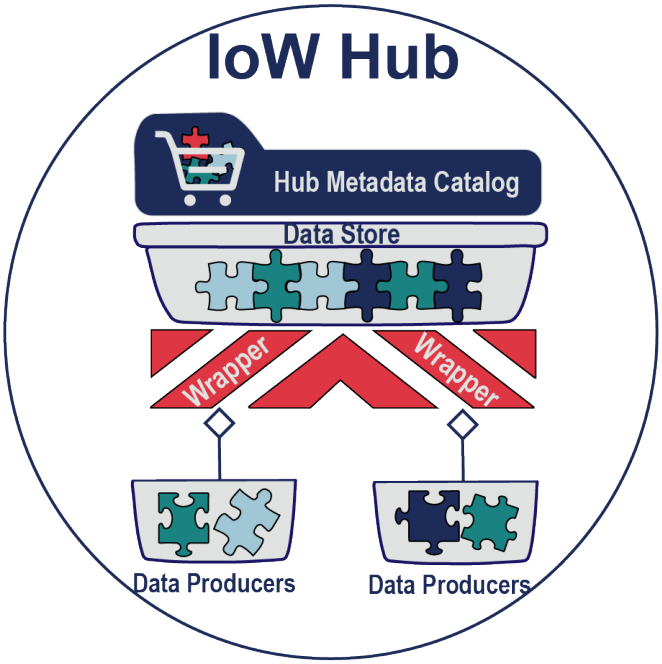
Hubs *pull* raw data from producers and convert the data into a standardized format that is saved in their data store with a metadata catalog.

Ideal For: storing a few data types across a few producers who have low capacity

Advantages: users can make complex queries; accessible when data producers are offline

Limitations: large computation and storage requirements for the hub; hubs must have high capacity to create and maintain wrappers

Type D: Centralized



Barriers: potential reservations by data producers to allow hub to standardize data

Examples: National Groundwater Monitoring Network

Which types of hubs for which circumstances?

Different types of hubs may be best suited for different circumstances depending on the capacity of data producers and hubs (Table 1). For instance, a sensor network requires a distributed (Type A) hub, whereas state administrative water rights might be better suited for Type D (a configuration to which WaDE is transitioning) (Figure 2).

Table 1: Summary of hub types and their attributes: *Note: in the table below, "Hub Set-Up" and "Hub Maintain" refer to the organizations that maintain a hub.*

Type	Producer Set-up	Producer Maintain	Hub Set-up	Hub Maintain	Real Time Data	Accessible if Producer Offline	Complex Queries Possible
Type A	Hard	Medium	Hard	Easy	Very Good	No	No
Type B	Medium	Medium	Hard	Easy	Poor	Yes	Yes
Type C	Medium	Easy	Hard	Medium	Poor	Yes	Yes
Type D	Very Easy	Very Easy	Hard	Very Hard	Very Poor	Yes	Yes

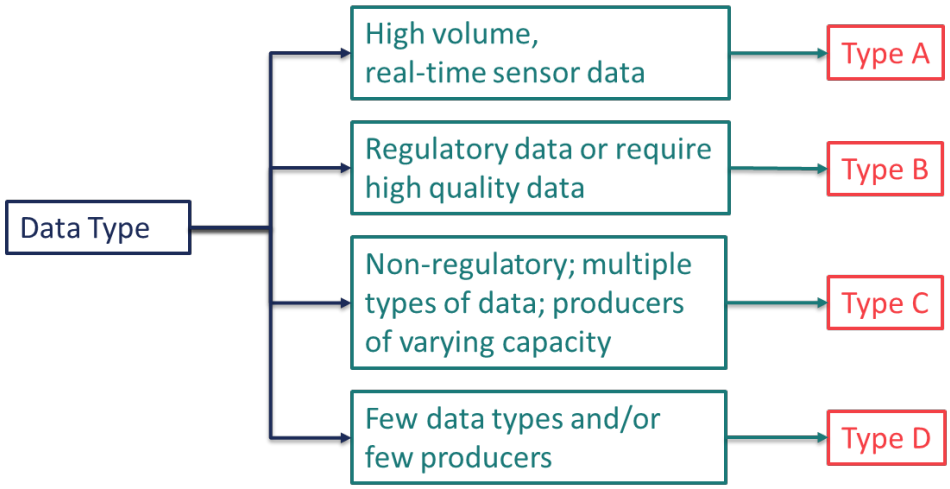


Figure 2: Different type(s) of data may be best suited for different types of hubs

Similarly, the producer's capacity and drivers for sharing data may also lend themselves to certain types of hub configurations (Figure 3). Note that these are descriptive tendencies and not prescriptive configurations for hubs.

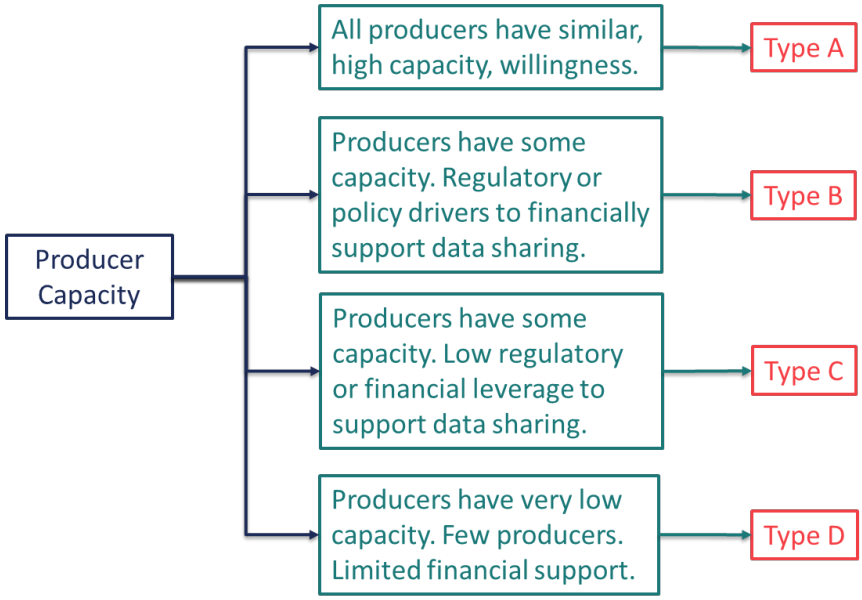


Figure 3: Producer capacity and/or the drivers for data sharing may lend themselves to certain types of hub configurations (descriptive, not prescriptive).

Data Users

The data users access the Internet of Water via simple web searches through a browser or the IoW metadata catalog (Figure 4). The IoW metadata catalog is being developed by the IoW community. This catalog will allow users to find data hubs and discover the data held within those hubs through search queries. This search is primarily designed based on key words and metadata. The IoW is currently partnering with the USGS to develop a water data knowledge graph, called Geoconnex, that will enable water data to be found based on location through commercial search index. Ideally, in a few years you will be able to google "water quality data Durham, NC" and quickly find the relevant data.

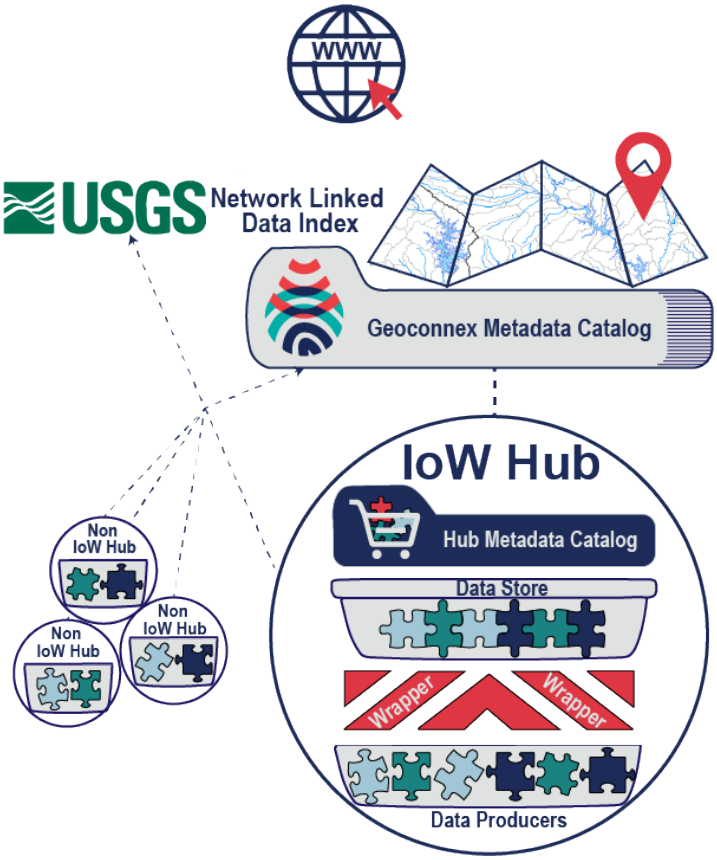


Figure 4: Users can discover data stored in IoW and Non-IoW hubs through search engines and the IoW Geoconnex Metadata Catalog.

The full Internet of Water framework example

A data producer collects several types of data: automated streamflow sensors, discrete water quality samples for regulatory purposes, and water rights (Figure 5). The data producer wants to make those data as findable as possible, so shares the data through multiple hubs corresponding to the appropriate data type. The data producer:

- shares real-time streamflow data with CUAHSI (Type A);
- converts the water quality data into WQX standards to comply with EPA, making the data findable through the Water Quality Portal (Type B); and
- tells WaDE 2.0 (Type D) where the water rights data are located so WaDE can pull those data into its own hub and convert the data into the hub's standardized format.

A user can now find the data producers data in multiple ways; they can do a web search, browse the IoW metadata catalog, or search the catalogs of the different hubs.

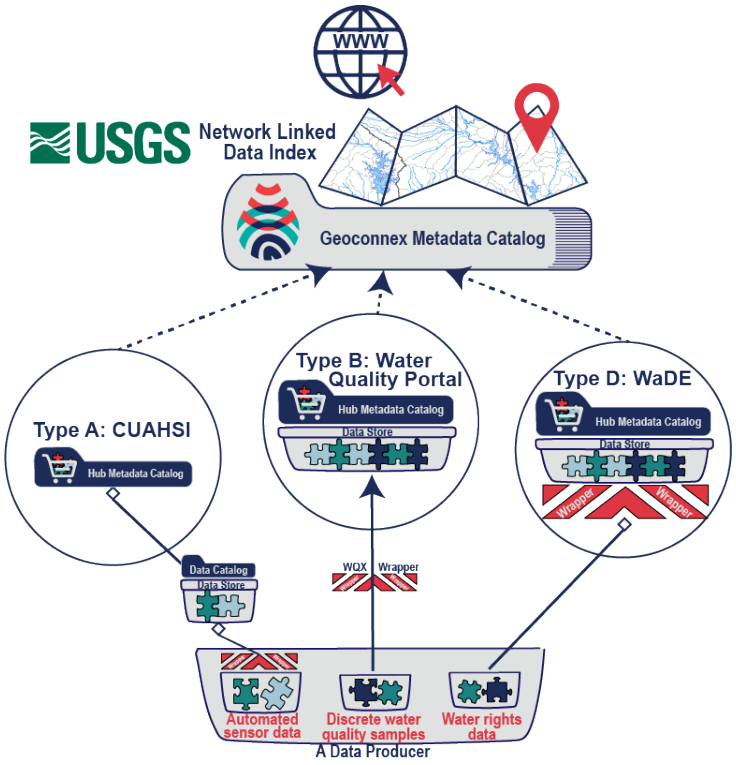


Figure 5: An example of a data producer sharing different types of data with different hubs. The data stored in the hubs are findable through web searches, the IoW metadata catalog, and the individual hub metadata catalogs.

Appendix: Terms

A technical working group convened by the IoW start-up team on October 9, 2019, developed a number of additional concepts and definitions necessary to further elaborate the IoW architecture, specifically:

- **(Meta)data Catalog:** A list of datasets, including structured metadata, that points to data sources.
- **Data Source**: A collection of data in a native, possibly non-IoW specification compliant format, as produced by a documented data collection and analysis process.
- **Data Standards**: documented agreements on the representation, format, definition, structuring, tagging, transmission, manipulation, use, or management of data.
- **Data Store**: Any object that persistently stores data. This includes relational databases as well as other types of data storage such as collections of documents and flat files. A data store may be considered FAIR when it meets a specific set of criteria (in development).
- **Data Wrapper**: An automated process that translates data from a native format into an IoW specification standard compliant format for storage in a data store.
- **IoW Hub**: A structured source of FAIR water data formally included in the "IoW Community" which conforms to best practices and specifications of the IoW and is interconnected with other IoW hubs.
- **Metadata:** Data describing who collected data, about what parameters, for what purposes, over what time period(s), at what location, and with what collection and analytical methods. This information should be sufficient to enable a determination about reuse of the data. (i.e., the Who, What, Where, When, Why, and How of the data).
- **Search Index**: A list of summarized versions of dataset contents produced by search engine crawlers that allow fast processing of search queries and data