Geoconnex

A New Way to Connect Data

Cassidy White • November 2021

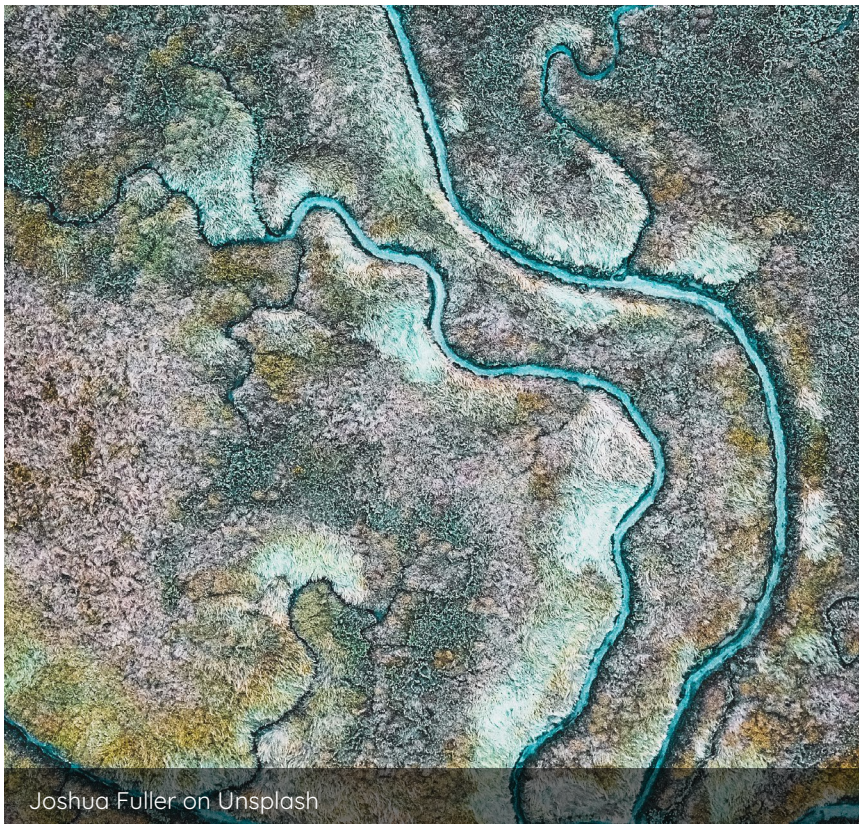
The fragmentation of water data creates an inherent problem for water managers, decision-makers, and researchers alike. Without consistent, consolidated, standardized, and easily accessible water data, the true state of water resources (e.g., quantity, quality, location) is not accurately represented. As a result, data users may inadvertently make faulty assumptions or predictions which ultimately lead to ill-informed decisions. Poor decisions can be catastrophic for the water sector, especially as water resources become both increasingly scarce and in high demand. Staff at the Internet of Water (IoW) are developing a solution.

In March 2020, IoW together with partners at the USGS and the SELFIE project, created the idea of developing structured metadata for water data. Geoconnex, a web tool for tagging and identifying water data, has since been in development by the IoW team with the intention of being fully operational by December 2022. A conversation with Kyle Onda, the IoW's Data Architect, sheds light on the need for structured water metadata and explains how Geoconnex will connect searchers to the data they seek.

"Poor decisions can be catastrophic for the water sector, especially as water resources become both increasingly scarce and in high demand. Staff at the Internet of Water (IoW) are developing a solution"

Geoconnex, Kyle explains, is a family of software and practices designed to create a water data specific search index. The tool involves linking water data with descriptive metadata, enabling the data to be organized (and discoverable) by theme and location with site-level granularity. Think of it like searching online for movie times. When you type "movie times" in the Google search bar, a neat, well-organized list of all movie showtimes in your area will be returned. You don't need to have any pre-existing knowledge about what movies are playing or the name of the theater.

Geoconnex will enable the same to be true for water data. Currently, a Google search for data about a particular water body often turns up results that provide general information about that water body, but rarely lead users to the data itself. Finding data about water can be difficult, but it doesn't need to be that way.


Joshua Fuller on Unsplash

An individual – be they an experienced researcher or a curious member of the public – should not need to know anything about where water data is housed to find it. With Geoconnex, the time-consuming process of searching the web for water data can be eliminated. But first, there are three key steps water data providers will have to follow for Geoconnex to be successful.

1. Every organization publishing water data needs to create a different webpage for each location for which they have water data. For example, the USGS has a different webpage for each of their stream gauges. These pages can be called "landing pages", as they are entrypoints to more detailed information that may be available elsewhere

2. Data providers need to embed structured metadata (following the JSON-LD format used by major search engines like Google, Bing, and Yahoo) inside each landing page housing water data. Metadata is essentially data about data – or a bit of coded information describing what kind of information can be found on a webpage. The metadata should include reference features, providing information on *where* the data came from (e.g., watershed, city, state, agency, etc.) and *what* kind of water data the site houses (e.g., groundwater vs surface water, depth, temperature, volume, etc.). For example, if a data provider has a landing page about a well in a county (County X), their metadata should be structured to say, "The data on this page is within County X." Ultimately, the metadata improves the relevance of search results and enables data discovery by the Geoconnex web crawler (described below).

3. All water data landing pages must have persistent identifiers (PIDs). Similar to the DOI system for academic articles, water data PIDs take users directly to a linked landing page. At geoconnex.us, data providers can mint identifiers for all their landing pages (See http://geoconnex.us for more information). If their landing page changes, they can then return to Geoconnex.us and remap their PID to the new URL, preserving the integrity of the search index and preventing the recurrence of broken URLs.

With the geoconnex.us PIDs and metadata in place, linked landing pages are organized and easily accessible through the https://reference.geoconnex.us landing page about a given reference feature that any data provider may have published data about. However, the ultimate data discovery tool will come in the form of a web crawler and the web applications it enables. A web crawler is essentially a robot that goes to landing pages and harvests data about that site. In this case, the IoW will have a list of all the water data landing pages with geoconnex.us identifiers. The web crawler can then go to all such landing pages, interpret and compact their JSON-LD formatted metadata, then copy and store it in a new, centralized database. From there, the IoW will build a Google-like interface where one can map out or search the data by organization, parameter, site name, location, watershed, upstream/downstream, and more.


Steven Sepulveda on Unsplash

Once the web crawler and database are complete, all linked water data will be available to the public in a single database for use through the IoW's associated web application. In addition, the public will be free to develop their own applications to manipulate and search the database. With these tools, a simple search will return the data from all organizations and landing pages with geoconnex.us PIDs whose metadata are embedded with relevant reference features (e.g., all data for a given location).

At the end of the day, the Geoconnex data identification process and tools can save water managers/researchers countless hours of navigating websites to locate the data they need. Nonetheless, it will not function without the participation of data providers. Agencies and other organizations must be willing to publish their landing pages with JSON-LD metadata, links to the Geoconnex reference features, and https://geoconnex.us/ identifiers. Geoconnex will be a free and open-source tool for use across the public domain and, although still under development, offers a promising pathway for overcoming water data fragmentation.

For a Geoconnex demo please visit https://geoconnex.internetofwater.dev/demo/index.html.

To assign Geoconnex PIDs to your organization's water data please visit https://geoconnex.internetofwater.dev/.

For more information feel free to contact Kyle Onda at kyle.onda@duke.edu.

Photo by Jason Schuller on Unsplash
Header Photo by USGS on Unsplash

Internet of Water