Geoconnex/Internet of Water WRRI Report

Kyle Onda

Table of Contents

1 Purpose

The purpose of this report is to document the activities, outputs, and outcomes of the NCSU Water Resources Research Institute award for the "Internet of Water: Research and Development toward a linked data system and foundational knowledge network for the Internet of Water". The linked data system and foundational knowledge network, named Geoconnex, was conceptualized as an operationalization of the Open Geospatial Consortium Environmental Linked Features Interoperability Experiments for the United States in the domains of water science and management. The outcomes include a performant infrastructure leveraging semantic technology and open, modern API standards that allow data providers to independently publish metadata on the web in a manner that results in their data becoming linked to other providers' data where spatially, hydrologically, and topically relevant. If adopted by data providers at wide scale, this would enable much improved discoverability of water-related datasets. Further work is needed to encourage participation in the system and use of the infrastructure. Potential future work could also establish best practices to use the same technologies to enable the automatic translation of observation and model data across data systems, fostering improved interoperability of data in addition tot he improved discoverability already enabled by the currently implemented infrastructure.

2 Introduction

2.1 Objectives

The main objective of this project is to establish a foundational framework for a contributor-based system that facilitates regular harvesting and cross-referencing of metadata through two key components:

1. Facilitating the Internet of Water community to publish detailed, machine-readable, and cross-referenced metadata (linked data)
2. Developing a centralized crawler/harvester to catalog all the linked data into a single knowledge graph, serving as a part of the index for eventual Internet of Water search utilities.

The objectives, therefore, encompassed both data-publisher-oriented research and development, as well as the establishment of a centralized knowledge network and services.

Specifically, the project aims to:

- Generate a demonstrative set of reference web content concerning various environmental features such as watershed boundaries, stream reaches, aquifers, monitoring locations, administrative geographies, and water-related infrastructure (e.g., dams, bridges) for data-publishing organizations to link to.
- Investigate existing hydrologic and web ontologies pertinent to the publication of data relevant to water science and management, and develop guidance for embedding linked-data content into web resources about water data.
- Develop a use case regarding the discovery and use of modeled and observed data for the same real-world feature or system of interest, and to investigate the metadata and knowledge graph infrastructure requirements to realize that use case.
- Develop software tools that enable data providers to generate and publish linked data without imposing burdensome requirements on existing data systems. All documentation and code were made publicly accessible.
- Create an open source web crawler/harvester infrastructure tailored to water data that can navigate web pages about water data, follow embedded links to other water data web pages, and harvest and catalog the metadata and their linkages to construct a single knowledge graph linking all water data.
- Investigate what governance mechanisms might be appropriate to establish and maintain the system to account for the needs of the water data community, incentivise participation by data providers, and facilitate use by data users.

2.2 Background

The Internet of Water is an initiative that aims to create a network of interconnected water data systems, modeled after the organizational structure of the Internet, as recommended by the Aspen Institute Dialog Series on Water Data Internet of Water Report. This aspiration is shared by an emerging community of water data producers and publishers, including but not limited to the USGS, USEPA, CUAHSI, the Western States Water Council, the Water Data Collaborative, and various state water resources and environmental quality agencies. This initiative now has an associated Coalition, with a steering committee composed of Duke University, CUAHSI, the Water Data Collaborative, the Western States Water Council, and the Lincoln Institute of Land Policy. There are several members of the coalition outside the steering committee, including representation from academia, philanthropy, the private sector, government agencies and intergovernmental associations, and professional societies. The Center for Geospatial Solutions at the Lincoln Institute, offers technology development, educational resources, and technical and social coordination for the wider Internet of Water community.

The U.S. Geological Survey Water Mission Area is actively working on the development of the National Hydrologic Geospatial Fabric (NHGF), a significant contributor to the Internet of Water. The NHGF is designed to establish a spatial-temporal framework to support water resources data and modeling across the United States. Currently, the best available water resources data on topics such as water availability, quality, and use for a specific feature of interest are collected, published, and in some instances, aggregated and republished by a diverse range of federal, state, tribal, local government, academic, and community-science organizations. This fragmented approach makes it extremely challenging for the general public, government, and scientific communities to locate all pertinent water data about a specific environmental feature.

To address this challenge, the Internet of Water is seeking to establish a collaborative partnership with the USGS and other researchers and technologists to develop and test metadata web publishing approaches, technologies, and communities. This collaboration aims to enable and incentivize all data producers to make their data discoverable through common internet-based spatial-temporal queries. The objective of this system, named Geoconnex, is to make a maximum amount of water information accessible via user-friendly search applications, without centralizing data governance and storage. Ideally, a user should be able to query a single web interface about a location of interest and receive enough metadata to quickly locate all places on the internet where water data about that location from all relevant organizations can be found.

The technical approach taken in this project leverages prior work, including:

1. The Second Environmental Linked Features Interoperability Experiment (SELFIE), which conceptualized a web architecture consisting of persistent identifiers for real-world water features of interest that direct to landing

pages with structured metadata that includes links to data relevant to the given feature of interest.

2. W3C Web standards and best practices, including for data on the web and spatial data on the web
3. Open Geospatial Consortium (OGC) API Standards that provide specifications for interoperable data sharing and processing services.
4. The OGC WaterML2 family of information models for water data
5. science-on-schema.org guidance for publishing metadata about scientific datasets
6. Several open source software projects. In particular, contributions were made for this project to:

- pygeoapi, a server that implements OGC API Standards
- gleaner, a metadata harvester that implements W3C best practices

2.3 Overview

The rest of the report is organized as follows:

Glossary provides definitions for specific terms and abbreviations used throughout the report.

User Engagement describes how users were identified and engaged throughout the project.

Architecture and Implementation of Geoconnex Linked Data System provides an overview of the implemented infrastructure components and summarizes their current performance and recommended future work.

Use Cases describe the general data discovery and publication use cases, and domain use cases that were developed and how they were addressed by Geoconnex implementation efforts.

Governance describe the general data discovery and publication use cases, and domain use cases that were developed and how they were addressed by Geoconnex implementation efforts.

3 Glossary

API Application Programming Interface, a set of rules for how machines can exchange information

Data Content A document accessible by URL that presents information about an NIR.

GeoSPARQL An OGC standard for representing and querying geospatial data in RDF

HTML HyperText Markup Language, a text format for web content

HY_Features Surface Hydrologic Features Conceptual Model

JSON JavaScript Object Notation, a data format common for web development and data transfer

JSON-LD JSON for Linking Data, a type of JSON designed to map JSON from different sources onto common vocabularies and data models to facilitate interoperability and automated data integration. It is a format of RDF.

Landing Resource A document accessible by URL that presents a default set of metadata –principally, links to Data Resources about a NIR.

NIR Non-Information Resource. A physical (e.g. a river) or conceptual (e.g. institution, jurisdictional area) object

OGC Open Geospatial Consortium, an international consensus standards organization for geospatial and sensor data and data processing and sharing services

OAFeat OGC API-Features, an OGC API Standard designed to provide vector geospatial data in a variety of formats

PID Persistent Identifier. An identifier that never changes for a given resource. In the Geoconnex context, referes to Geoconnex PIDs minted at the Geoconnex Persistent Identifier Registry

pygeoapi An open-source python server that implements OGC API standards

RDF Resource Description Framework, a generalized data model for knowledge graphs and cross-dataset interoperability.

Registry An information system that manages files containing identifiers. In the context of Geoconnex, the Geoconnex Persistent Identifier Registry at https://geoconnex.internetofwater.dev

Resolver A system that redirects URIs to URLs. In the context of Geoconnex, refers to the Geoconnex Resolver that redirects URIs that begin with https://geoconnex.us/

schema.org: A vocabulary for use in structured data embedded into websites for search engine optimization and cross-website data interoperability

SELFIE Second Environmental Linked Features Interoperability Experiment

SPARQL A standard query language for RDF data.

URI Uniform Resource Identifier, a unique set of characters that identifies a resource. Within the geoconnex context, URIs shoud be HTTP URIs, structured like URLs, that identify a real-world resource (NIR), but direct via HTTP code 303 ('See Other') to a Landing Resource about the NIR

URL Uniform Resource Locator, or web address for any kind of web resource. Within Geoconnex, URLs are distinguished from URIs in that URLs point to or perhaps identify information resources, but not NIRs/real-world objects. URIs identify NIRs but direct to web resources that have information about NIRs.

4 User Research and Engagement

4.1 Outreach

Several direct engagements, including rounds of technical assistance in implementing Geoconnex data publication practices, were conducted with data providers (see Section 5.4) who are represented in the Internet of Water Coalition. Insights gleaned from these engagements included the need for web developer-friendly tools to ease the publication of JSON-LD from common data service platforms including CKAN, ArcGIS Enterprise/Online, and TylerTech Data & Insights (formerly Socrata).

In addition presentations and webinars were delivered that included significant Q&A sessions for feedback from the water data community including

- An Internet of Water Webinar (December 2022)
- A presentation at the 2022 National Water Use Data Workshop (August 2022)
- A presentation at FOSS4G in Prizren (July 2023)
- A presentation for NSGIC (August 2023)
- A second Internet of Water Webinar designed for a general public audience (August 2023)

4.2 Interviews

In June-August 2023, funding external to the WRRI grant was leveraged to conduct 10 1-hour interviews with water data experts and federal and state agency water data provider staff to research desired characteristics for a future geoconnex data discovery API, which will influence both the kinds of metadata requested of data providers to publish, and the structure of any interfaces to the knowledge graph.

5 Architecture and Implementation of Geoconnex Linked Data System

All code associated with the architecture is navigable from the GitHub Geoconnex Directory Repository at https://github.com/internetofwater/about.geoconnex.us

5.1 Content Model

The overall content model in this project is derived from the concept proposed in SELFIE (see Figure 1), which applied W3C Data on the Web Best Practices to environmental data use cases.

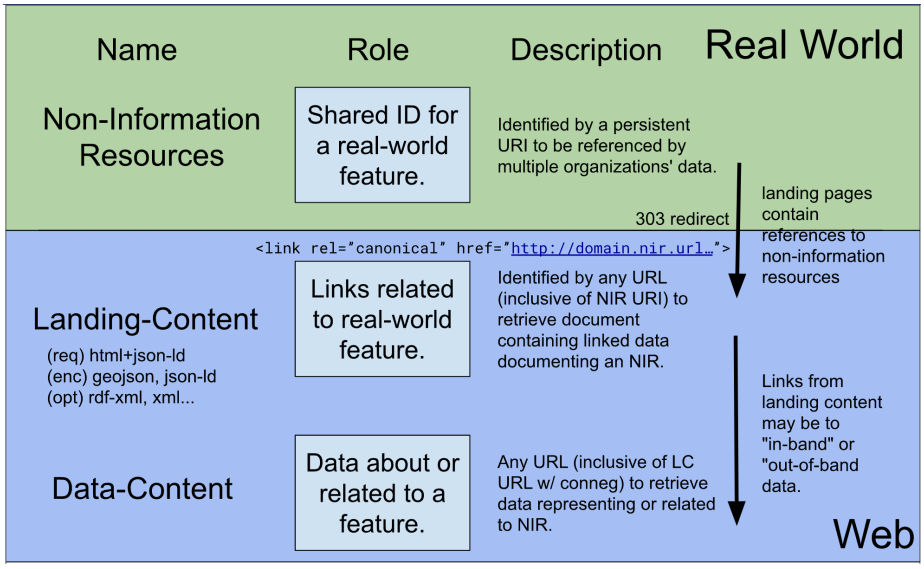

Figure 1: Content Model, derived from SELFIE Engineering Report

Non-information Resources (NIR) are real-world features such as rivers, wells, dams, lakes, public water systems, conduits, etc. NIR are to be identified by HTTP(s) URIs, which are strings of characters that are formatted like URLs (web addresses). For example, the Hoover Dam would have a URI of https://geoconnex.us/ref/dams/1080095. These URIs, when entered into a web browser, are to redirect to a URL where there is **Landing Content** including some structured metadata about the feature the URL identifies in the form of embedded JSON-LD, which makes the metadata machine-readable and interoperable with similarly published metadata. This **Landing Content** then includes links to **Data Content**, which can be any web-accessible structured and unstructured data about a URI. Data content might similarly have a JSON-LD version for maximum interoperability, but could simply be data in any format available at a URL.

For example, https://geoconnex.us/ref/dams/1080095 is a persistent identifier that a resolver responds with a HTTP 303 "See Other" redirect to https://reference.geoconnex.us/collections/dams/items/1080095. This web page has some basic information about the dam (see Figure 2).

The web page also has JSON-LD version which can be parsed by web browsers or computer programs with the same information specified using standard vo-

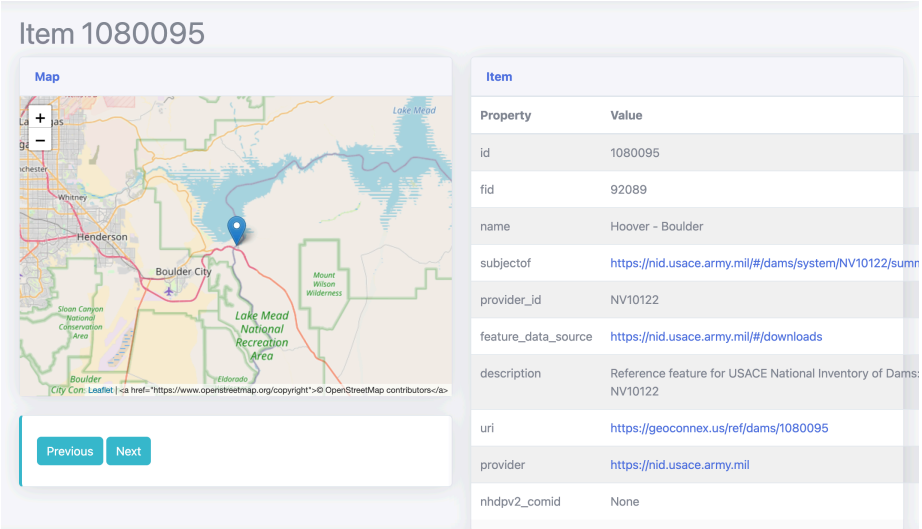

Figure 2: Landing Content

cabularies:


```
{
  "@id": "https://geoconnex.us/ref/dams/1080095",
  "@type": "https://schema.org/Place",
  "http://www.opengis.net/ont/geosparql#hasGeometry": {
    "@type": "http://www.opengis.net/ont/sf#Point",
    "http://www.opengis.net/ont/geosparql#asWKT": {
      "@type": "http://www.opengis.net/ont/geosparql#wktLiteral",
      "@value": "POINT (-114.73740000000001 36.0163)"
    }
  },
  "https://schema.org/description": "Reference feature for USACE National Inventory of Da
  "https://schema.org/geo": {
    "@type": "https://schema.org/GeoCoordinates",
    "https://schema.org/latitude": 36.0163,
    "https://schema.org/longitude": -114.73740000000001
  },
  "https://schema.org/name": "Hoover - Boulder",
  "https://schema.org/provider": {
    "@type": "https://schema.org/url",
    "@value": "https://nid.usace.army.mil"
  },
  "https://schema.org/subjectOf": {
```

```
      "@type": "https://schema.org/url",
      "@value": "https://nid.usace.army.mil/#/dams/system/NV10122/summary"
    }
  }
```

Thus, the identifier https://geoconnex.us/ref/dams/1080095 can unambiguously provide a computer program with the information that this dam has the name (specifically, the https://schema.org/name, itself a URI for the property of "having a name") of "Hoover - Boulder", as well as similarly unambiguous representations of its latitude and longitude. In addition, there are links to **Data Content** via the `schema:subjectOf` property, which in this case is simply a URL for the record of the Hoover Dam published by the National Inventory of Dams of the U.S. Army Corps of Engineers.

In responding to the principal use case of the Geoconnex system, which is to provide a multi-organizational community index for all water-related information, the technical requirement of the Geoconnex system is to facilitate the population of **Landing Content** with well-annotated links to **Data Content** from all organizations who publish data about the same feature represented in the **Landing Content**. For example, the U.S. Bureau of Reclamation has a set of data content about mussel detection on the Hoover Dam here https://data.usbr.gov/catalog/8. The technical aim of the Geoconnex system is to enable the automation of the extension of the landing content to include metadata about this data, including its subject, methods, period of record, and data format, etc. The resulting JSON-LD of the **Landing Content** could thus look something like this:

```
  "https://schema.org/subjectOf": [
  {
    "@type": "https://schema.org/url",
    "name": "National Inventory of Dams",
    "@value": "https://nid.usace.army.mil/#/dams/system/NV10122/summary"
  },
  {
  "@type": "Dataset",
  "@id": "https://data.usbr.gov/catalog/8/item/878"
  "name": "Lake Mead Hoover Dam and Powerplant Intermittent Veliger Density Time Series D
  "description": "Measurements of quagga mussel veliger numbers from Lake Mead Hoover Dam
  "temporalCoverage": "2013-01-01/2018-01-01"
    "distribution": {
    "@type": "DataDownload",
    "name": "USBR RISE API",
    "contentUrl": "https://data.usbr.gov/rise/api/result/download?type=csv&itemId=878&bef
    "encodingFormat": [
      "text/csv"
```

```
      ],
      "dc:conformsTo": "https://data.usbr.gov/rise-api#Result"
    }
    }
  }
    ]
```

The overall architecture of the system as currently implemented is illustrated in Figure 3

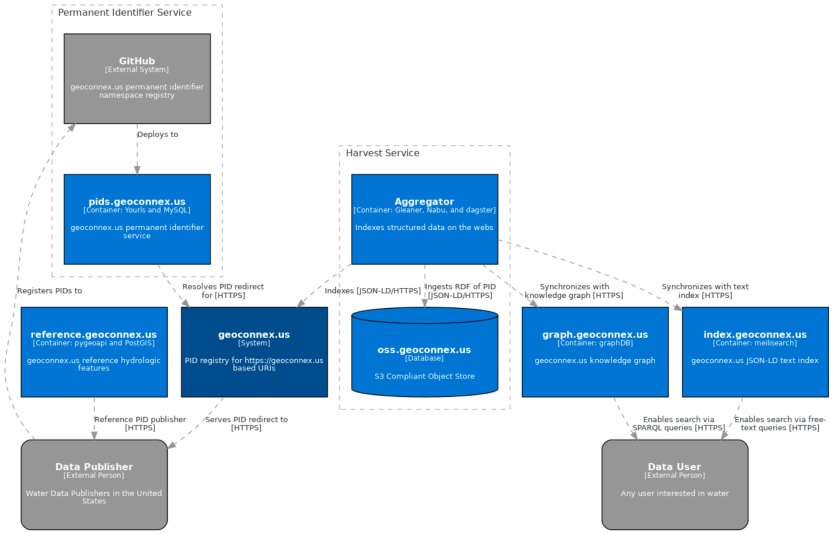

Figure 3: Geoconnex Architecture Diagram

5.2 Persistent Identifier Service

The persistent identifier (PID) service consists of a **registry**[1] and a **resolver**.

Registry (https://geoconnex.us)

The registry mints URIs for real-world resources and specifies which URLs they 303 redirect to. In the Geoconnex context, these URIs are referred to as PIDs. The registry is currently administrated through GitHub https://github.com/internetofwater/geoconnex.us, and allows for moderated contributions of `.csv` files to organizational and reference namespaces. See Figure 4 for an example csv file in the registry, in this case for URIs directing to USGS representations of HUCs.

Organizational namespaces, such as `https://geoconnex.us/usgs/` for USGS, allow organizations to mint identifiers for locations they have data about.

[1] https://geoconnex.us

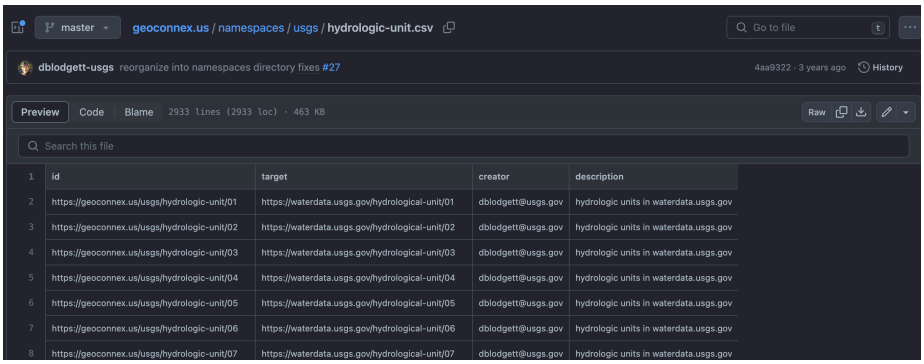

Figure 4: Geoconnex PID registry csv

Further topic hiearchies are possible at each organization's discretion, such as USGS using `https://geoconnex.us/usgs/monitoring-location/` to identify their monitoring locations. Reference namespaces (all beginning with `https://geoconnex.us/ref/`) refer to community-wide, nation-scale locations that many organizations might have data about, such as `https://geoconnex.us/ref/dams/`. Ideally, it is envisioned that data providers are able to tag their datasets with the reference features they are relevant to. Additional reference features are made available for cataloging purposes, such as HUCs, States, and Counties, Aquifers, and Mainstem Rivers. It is possible that datasets may be measuring variables on these features, but in most cases they are cataloging containers to aid in data discovery and filtering for smaller features of interest like specific streamgage and well locations.

PIDs contributed to the registry automatically populate a sitemap index at https://geoconnex.us/sitemap.xml, a standardized format for listing web addresses for web crawlers such as those operated by commercial search engines, as well as the harvester implemented for the Geoconnex system.

For organizations that have more than 300,000 locations, GitHub cannot support the resulting file sizes of csvs, so the Geoconnex project reccomends minting a regex (regular expressions) identifier. This allows for the use of wildcards. For example, https://geoconnex.us/usgs/monitoring-locations/%5B0-9%5D which can redirect to any number in the place of `$1`, such as https://geoconnex.us/usgs/monitoring-locations/04127997 redirecting to https://waterdata.usgs.gov/monitoring-location/04127997.

Resolver (pids.geoconnex.us)

The resolver is a customized fork of the open-source YOURLS URL shortener. The code for the Geoconnex implementation is available from https://github.com/internetofwater/pids.geoconnex.us. YOURLS is written in PHP and interfaces with a MySQL database, and was chosen because of its highly extensible plugin framework. Multiple plugins were developed (**under non-**

WRRI funding) to give YOURLS necessary capabilities for Geoconnex including: a bulk importer, a regex PID resolver, a filetype forwarder, and a connector to serverless cloud databases.

5.2.1 Performance and future work

The registry is now adequate technically, however there are two areas for improvement. First, some data providers are unfamiliar with GitHub or regular expressions, and others find the process of contributing to PIDs via GitHub pull requests to be cumbersome and insufficiently automate-able. Future work should investigate the feasibility of both a more user-friendly dedicated user interface, and the possibility of allowing access to an API to request and verify the minting of PIDs. Second, contributors of regex PIDs nevertheless need a consistent way to provide a comprehensive list of every URL to the Geoconnex system so that they can be crawled. Specific guidance for creating and submitting self-hosted, or alternatively hosted sitemap.xml files will be necessary to develop.

The resolver was initially inadequate to serve redirects to landing resources at scale while the Aggregator was actively retrieving all resources from PIDs in https://geoconnex.us/sitemap.xml, with resolution times in the 1-second range and an error rate of ~5%. To address this, replica persistent identifier databases were set up, each with its own autoscaling, load-balanced redirect server container. This cut resolution times to less than 50ms on average and the error rate to less than 0.1% during aggregator runs. In anticipation of increased usage in the future, the resolver was also put behind Cloudfront, the Amazon Web Services global Content Delivery Network, which caches requests and reduces the load on both the persistent identifier resolver and any servers that provide landing content at the redirected URLs.

5.3 Reference Feature Server (reference.geoconnex.us)

The reference feature server (https://reference.geoconnex.us) provides an OGC-API features endpoint implemented using the open-source python server pygeoapi. A key contribution the Geoconnex project (under **non-WRRI funding**) made to pygeoapi was the capability to add JSON-LD templates to the server's responses for individual items, allowing for completely customized JSON-LD formats for different types of source data. For example, the reference gage URI https://geoconnex.us/ref/gages/1000012 points to https://reference.geoconnex.us/collections/gages/items/1000012, the OGC-API endpoint for that feature. Computer programs can be written to retrieve the HTML (web browser default), geoJSON, or JSON-LD versions of this feature. It includes basic attributes for hydrographic addressing, including the NHDPlusV2 reachcode and measure as well as the relevant mainstem. JSON-LD templating allows the JSON-LD output to conform with the relevant OGC surface hydrology conceptual model HY_Features to allow for intepretation of these attributes in a hydrologically meaningful way:

```
"hyf:referencedPosition":[
        {
            "hyf:HY_IndirectPosition":{
                "hyf:distanceExpression":{
                    "hyf:HY_DistanceFromReferent":{
                        "hyf:interpolative":31.7648
                    }
                },
                "hyf:distanceDescription":{
                    "hyf:HY_DistanceDescription":"upstream"
                },
                "hyf:linearElement":"https://geoconnex.us/nhdplusv2/reachcode/10190004000
            }
        },
        {
            "hyf:HY_IndirectPosition":{
                "hyf:linearElement":{
                    "@id":"https://geoconnex.us/ref/mainstems/461532"
                }
            }
        }
    ]
```


Reference feature collections can searched via the OGC-API Features standard, including the Common Query Language filters, including by bounding box, polygon, datetime, or attribute to aid in discovery of the most relevant features to a data provider's data for tagging and metadata management, or to a data user's use case.

The reference feature server's collections are imagined to be thematic collections of nation-scale common features that many organizations publish data about, and that are managed by a community process that needs to be developed further (see Governance section. The source datasets are available for bulk download with an open license from HydroShare.[2] For the purposes of this project, reference features were curated by USGS and CGS staff. USGS staff curated reference feature sets for HUCs, mainstems, gages, aquifers, hydrogeologic units, and dams. CGS staff curated reference feature sets for Census geographies and public water systems.

5.3.1 Performance and future work

Technically, the reference feature server performs well at scale. For example, USGS Web Communications Branch staff crawled the entire mainstems feature collection to add mainstem identifiers to the USGS SensorThings API instance,

[2] https://www.hydroshare.org/resource/3295a17b4cc24d34bd6a5c5aaf753c50/

and the Geoconnex crawler can similarly crawl every individual item across all collections without sacrificing performance for other users. Future work should focus on fostering a community of practice that can take responsibility for curating additional reference feature collections (e.g. for wells, lakes, wetlands, conduits, aquatic barriers) and maintaining all collections over time as the water data community is made aware of new data sources.

5.4 Organizational Web Content

Organizational Landing and Data content was fostered using a number of strategies. Direct engagement (see Section 4) with data providers including USGS Water Mission Area Web Communications Branch, the US Bureau of Reclamation RISE system, the New Mexico Water Data Initiative (NMWDI), the California State Water Resources Control Board (SWRCB), the Texas Water Development Board's Texas Water Data Hub project, the Western States Water Council Water Data Exchange (WaDE) system, and the Oak Ridge National Laboratory HydroSource data catalog were pursued. The following activities resulted:

- PIDs were minted for relevant USGS resources that were already published, such as the monitoring location pages.
- USBR RISE implemented location-based HTML landing content (e.g. https://data.usbr.gov/location/314) but is still in the process of implementing embedded JSON-LD as of the writing of this report.
- The New Mexico Water Data Initiative's primary data access mode is a number of OGC SensorThings API (STA) endpoints. To support their publication of landing content, this project implemented an STA provider for pygeoapi, and assisted NMWDI in deploying it. Thus, NMWDI PIDs (e.g. https://geoconnex.us/nmwdi/st/locations/4108) direct to a pygeoapi-implemented OGC-API Features endpoint which includes JSON-LD, which includes semantic links to data content from the STA endpoint.
- The SWRCB and HydroSource platforms rely on ESRI ArcGIS REST services to publish geospatial data and had no other easy framework to implement location-based landing content. An ArcGIS REST service provider for pygeoapi OGC-API Features was thus implemented. HydroSource is in the process of implementing semantic landing content at https://hydrosource-features.ornl.gov as of the writing of this report. The SWRCB minted PIDs for a streamgage catalog that was also inorporated into the reference gages layer, and is currently represented on CGS infrastructure at https://sb19.linked-data.internetofwater.dev.
- The Texas Water Data Hub is implementing the open-source CKAN as its primary data catalog and the CKAN datastore and API as its primary data service, and is in the process of implementing dataset-oriented

content on the data resources pages, and investigating how to leverage the platform to serve location-oriented landing content. In support of potential organizations that use CKAN and the similar proprietary Socrata platform as data services, providers for both APIs for pygeoapi OGC-API Features were implemented.

- The WaDE platform uses Mapbox and a custom database to deliver data services regarding more than 1 million points of water diversion and use across 17 western states. While the WaDE team has implemented template-based dynamic HTML/JSON-lD landing content, and minted Geoconnex persistent identifiers for each of WaDE's locations (e.g. https://geoconnex.us/wade/sites/MTwr_SPOD437023), crawling the pages incurred undesirable data egress charges for WaDE. As an alternative approach, the creation of a bulk JSON-LD graph of all WaDE's metadata content was piloted. It was able to be ingested into the triple store manually and via Gleaner with a custom sitemap entry for the entire graph file.

5.5 Author Guidance

Throughout engagement regarding organizational web content, guidance for how to format and publish JSON-LD was iterated on an ad-hoc basis and synthesized over time. Documentation specific for Geoconnex data providers-as-web content authors is ongoing at https://docs.geoconnex.us. A coherent, minimum set of guidance for JSON-LD content which emphasizes https://schema.org and iterates on science-on-schema.org is available at https://geoconnex.us/iow/guidance.

5.5.1 Performance and future work

It was found that uptake is far more likely when JSON-LD requirements are as parsimonious as possible and reduce reliance on ontologies and vocabularies other than https://schema.org, which web developers are familiar with and which has robust documentation. For example, attempts to rigorously implement patterns from the Semantic Sensor Network ontology found that the concepts therein, especially around Sensors, Procedures, Platforms, Deployments, and Samplings were difficult to consistently apply to the metadata that was easily available for publication from participating source data systems. Nevertheless, it was found that properties from HY_Features, GeoSPARQL, and QUDT were absolutely necessary, and properties from the Observations Data Model 2 (ODM2) Vocabularies were useful in most cases. It is plausible that more user-friendly publication and documentation of domain models would enable greater degrees of standardization. Domain models and authoritative or community vocabularies that need to be published with resolvable URIs and accessible documentation and search interfaces include:

- groundwater features

- water quality and quantity regulatory concepts (e.g. action and enforcement thresholds, beneficial use categories, water rights allocations and priorities)
- quantity kinds and parameters
- environmental observation and model methods
- infrastructure properties (e.g. dams, reservoirs) and anthropogenic hydrology (e.g. conduits, culverts, levees)

5.6 Aggregator service

The aggregator services deploys the open source softwares Gleaner and Nabu which were developed initially for NSF's EarthCube. The activity flow is shown in Figure 5.

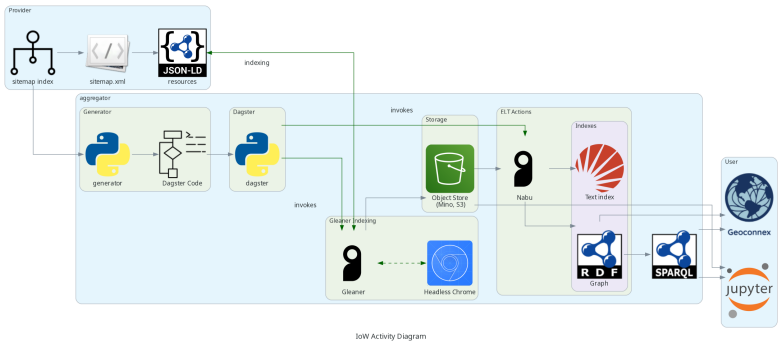

Figure 5: Geoconnex Architecture Diagram

Gleaner is configured against one or more web-accessible `sitemap.xml` files, and deploys a "headless" Google Chrome browser to retrieve JSON-LD from the URIs in the sitemaps, and load them individually into an S3-compliant object store. It is scheduled to run via python scripts that are scheduled and orchestrated by a deployment of dagster, an open-source data pipeline scheduler. The object store serves as the "source of truth" for the knowledge graph/triple store. Resources that are crawled multiple times with changes in between results in the JSON-LD documents in the object store being overwritten.

Nabu performs two functions:

1. loads the JSON-LD documents into a triple store, which in this case is an instance of GraphDB with a free license (see Section 5.7 for further elaboration on the function and performance of GraphDB).

2. loads the JSON-LD documents into a free text index of the JSON-LD documents, which provides a way to quickly retrieve relevant URIs based on free text search, for referral to a knowledge graph for more semantic queries.

5.6.1 Performance and future work

Gleaner and Nabu were demonstrated to successfully harvest JSON-LD and load them into the triple store. The initial deployment deployed all components of the aggregator service in a single docker-compose framework on a single virtuam machine. This deployment experienced upper limit on the number of resources crawled due to the memory allocatable to the S3 data store within the framework. This was resolved by reconfiguring Gleaner and Nabu to work with a serverless S3 infrastructure (in this case, Google Cloud Storage). To fulfill two key functions of public data access and data versioning, a third capability should be finalized with Nabu: to concatenate the JSON-LD documents from a given sitemap or namespace into a single, versioned and timestamped RDF release graph for archiving and download at a publicly accessible S3-compliant object store bucket.

5.7 Knowledge Graph/ Triple Store

The knowledge graph is primarily instantiated as an RDF triple store, in this case using GraphDB. GraphDB, though proprietary software, implements all data interfaces with open standards, including SPARQL and RDF4J. GraphDB was chosen over available open source triple stores due to it support for GeoSPARQL inference and query operations. GraphDB's RDF4J REST API includes a SPARQL query endpoint that can be used in a read-only mode for public access. https://graph.geoconnex.us/repositories/iow is the service endpoint for both GET and POST SPARQL queries. In addition, this endpoint was demonstrated to be a practical back end to OGC-API Features endpoints delivering GeoJSON and HTML output for simple SPARQL queries, as implemented in a custom SPARQL provider for pygeoapi.

5.7.1 Performance and future work

GraphDB is a capable enterprise-grade triple store and can process SPARQL and GeoSPARQL transactions to update and query data. However, at the imagined scale of Geoconnex if widely used and adopted, including many simultaneous, asynchronous queries integrated into major high-traffic Federal and State data systems, a more robust architecture including database replicas with failover protection will need to be investigated and costed.

In addition, the vast majority of even highly technically skilled water data analysts are not familiar with SPARQL. More traditional RESTful API entrypoints to the knowledge graph should be investigated. One approach would be to investigate several common SPARQL property paths that address key use cases,

and implement them at the reference feature OGC-API Features endpoints. Another approach would be to implement a number of OGC-API Processes which implement common graph queries using a constrained set of query parameters.

5.8 Usage statistic tracking investigation

The following data sources were investigated to track and provide usage statistics for community linked data resources

- Google Analytics
- PID server logs
- CDN distribution logs

5.8.1 Google Analytics

The PID server was unable to incorporate Google Analytics.

5.8.2 PID server logs

PID server logs provided reasonably complete metrics for hits, referrers, and data transfer volume information for each PID. However, once the PID server was put behind the Cloudfront CDN, they become less reliable as the CDN began serving from its cache for popular resources.

5.8.3 CDN distribution logs

The Cloudfront logs provide complete information on which PIDs are being accessed, whether or not the response is forwarded from the source or the cache, but raw logs are difficult to interpret and should be visualized and/or tabulated.

5.8.3.1 AWS CloudWatch

Overall hit and error rates and data transfer volume are easily configurable metrics for AWS CloudWatch dashboards, which can provided via publicly accessible link or only to specific users. It was also possible to configure more specific metrics, such as for hits for PIDs in specific namespaces. However, each namespace needed to be configured individually. Thus, if the use case involves providing custom metrics for data contributors for their specific namespace only, AWS CloudWatch would be arduous to configure and maintain.

5.8.3.2 Data export

It was relatively straightforward to export and concatenate the CloudFront logs as they were posted to an AWS S3 bucket. The complete log file could thus be concatenated on at least a daily basis using a simple shell script. The resulting file can be easily queried and analyzed with spreadsheet programs, databases, or R, Python, or other programming languages.

5.8.4 Performance and Future work

For the simple use case of tracking access to specific Geoconnex knowledge network resources, either on an individual or aggregated by namespace basis, maintaining a data export for visualization in an R Shiny App or Python Dash App is most likely the next step to creating a publicly viewable and resource tracking system. However, this solution would not be able to track second-order links (that is, non-Geoconnex PID URLs in the linked data). Moreover, data providers wishing to track access hits to their systems that are referred via Geoconnex resources would need to use their own server logs. Moreover, in many cases they would only see that hits were referred via landing resource URLs, not the upstream PIDs, and so only reference feature referrers (from https://reference.geoconnex.us) could be reliably tracked.

Future work could investigate the feasibility of populating the Geoconnex CDN logs with hits to linked data. One approach would be to periodically mint special geoconnex URLs that redirect to each dataset URL detected in the knowledge graph. For example, a dataset about a specific dam might have a DOI-URL like https://doi.org/10.1111/12234, and landing content that includes `"schema:about":"https://geoconnex.us/ref/dams/1000001"`. The Geoconnex system could in principle add a redirect from `https://geoconnex.us/data/doi.org/10.1111/12234` to https://doi.org/10.1111/12234. Then, these minted URLs could serve as the primary entrypoint for linked datasets in user interfaces and information products in the Geoconnex system, such that data providers' data that is discovered via Geoconnex is likely to be accessed via a redirect from `https://geoconnex.us/data/doi.org/10.1111/12234`, hits to which could be tracked via the Geoconnex PID server.

6 Use Cases

Below both a highly general use case is described, as well as a specific use case involving the discovery and integration of hydrologically related observed and modeled data.

6.1 General use case

The general geoconnex use case is essentially the same as the "Internet of Water" use case articulated in Internet of Water Aspen Institute Report[3] and summarized in the SELFIE Engineering Report[4].

6.1.1 User Story

As a user of water data, I need to discover and access water information relevant to the environmental feature I care about from all the organizations that

[3] https://www.aspeninstitute.org/publications/internet-of-water/

[4] https://docs.ogc.org/per/20-067.html#_us_internet_of_water_distributed_data_and_observations

hold data about it, so I don't have to have special knowledge to access some information and so I don't miss some potentially relevant information.[5]

6.1.2 Datasets and sources

- USGS Reference Hydrography
- State and local data and observations
- University consortia aggregated data services
- Federal aggregated data and services
- Nongovernmental aggregated data and services

6.1.3 Conceptual Demonstration

The Geoconnex system is ultimately designed to meet this highly general use case. The overall architecture envisions a federated system of data producers that all participate by publishing landing content that references Community Reference Features, and registering URIs and/or PIDS for that content with the persistent identifier registry. The combined landing content can then be organized into a knowledge graph that is made publicly accessible. Diverse data discovery workflows can be accommodated via SPARQL query to the knowledge graph or instantiated via SQL-enabled data stores or tabular and geospatial data files created from the knowledge graph.

A simple example of discovering streamgages from all participating organizations that monitor a specific river is elaborated below:

The Reference Gages OGC-API Features endpoint could be configured to include an attribute which on the back end is a response to SPARQL query that returns URIs, names, and geometries for all reference gages that are indexed to the same mainstem river:

```
PREFIX hyf: <https://www.opengis.net/def/schema/hy_features/hyf/>
PREFIX schema: <https://schema.org/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>

select DISTINCT ?mainstem ?gage ?gagename ?streamname ?provider ?gwkt ?mswkt where {
    <https://geoconnex.us/ref/gages/1118104> hyf:referencedPosition ?rp .
    ?rp hyf:HY_IndirectPosition ?ip .
    ?ip hyf:linearElement ?mainstem .
    BIND (<https://geoconnex.us/ref/mainstems/29559> as ?target)
    ?gage hyf:referencedPosition ?rp2 .
    ?rp2 hyf:HY_IndirectPosition ?ip2.
    ?ip2 hyf:linearElement <https://geoconnex.us/ref/mainstems/1> .
```

[5] Reproduced from the SELFIE engineering report

```
    <https://geoconnex.us/ref/mainstems/1> schema:name ?streamname .
    ?gage geo:hasGeometry ?ggeom .
    ?gage schema:name ?gagename .
    ?gage schema:provider ?provider .
    ?ggeom geo:asWKT ?gwkt
        }
```

The resulting GeoJSON response for that gage from https://reference.geoconnex.us/collections/gages/items/1118104 could then be:


```
{
    "type":"Feature",
    "properties":{
        "uri":"https://geoconnex.us/ref/gages/1118104",
        "description":"USGS NWIS Stream/River/Lake Site 05451910: Iowa River at Chelsea,
        "provider":"https://waterdata.usgs.gov",
        "provider_id":"05451910",
        "nhdpv2_reach_measure":30.648,
        "mainstem_uri":"https://geoconnex.us/ref/mainstems/324976",
        "fid":120134,
        "name":"Iowa River at Chelsea, Iowa",
        "subjectof":"https://waterdata.usgs.gov/monitoring-location/05451910",
        "nhdpv2_reachcode":"07080208000193",
        "nhdpv2_comid":17541869.0,
        "gages_same_mainstem":[
        {
            "uri":"https://geoconnex.us/ref/gages/1017824",
            "name":"Mill Race at Amana, IA",
            "latitude":41.79611839,
            "longitude": -91.86517779
        },
        {
            "uri":"https://geoconnex.us/ref/gages/1017821",
            "name":"Iowa River at Columbus Junction, IA",
            "latitude": 41.27835889,
            "longitude": -91.3468216
        }
    ]
    },
    "id":"1100229",
    "geometry":{
        "type":"Point",
        "coordinates":[
            -106.7075374,
```

```
                39.8894315
            ]
        }
    }
```

The GeoJSON representation of the `mainstem_uri` can be retrieved easily and a second GeoJSON would be simple to construct from the array at the node `gages_same_mainstem`.

A visual representation of this example is available in a web application[6]. The workflow is as follows:

First, an area of interest can be navigated to on a web map that displays mainstem rivers and all known surface water monitoring locations:

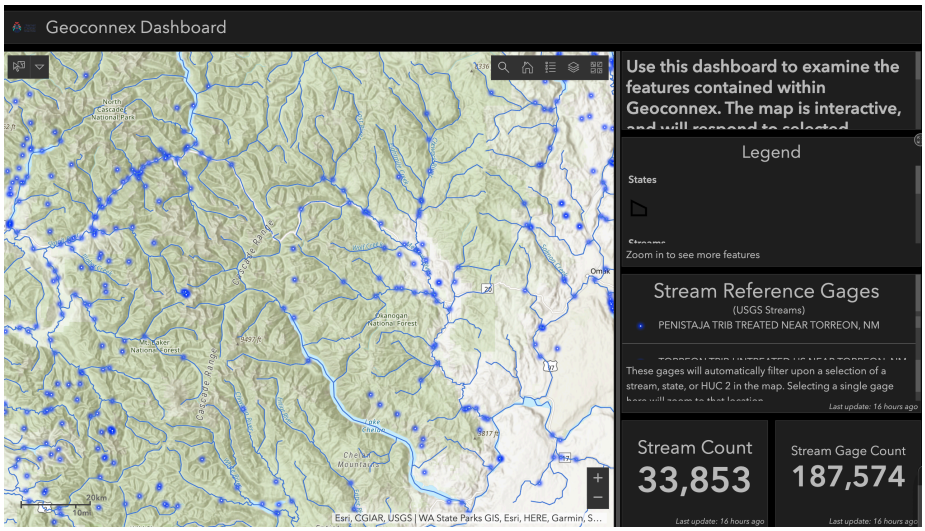

Figure 6: General Use Case: Finding monitoring locations by river, step 1

Second, any monitoring location can be selected. This results in the view being subsetted to only the mainstem river that the selected monitoring location is linked to in the knowledge graph, and all other monitoring locations linked to the same mainstem river.

6.1.4 Next steps

At this point, the technical baseline has been established to meet a wide variety of data discovery and access use cases. Reasonable next steps might include:

- incorporating next-generation USGS hydrography products (e.g. 3DHP mainstems) into reference feature collections

[6] https://gis.cgs.earth/portal/apps/opsdashboard/index.html#/0e113bea7c0542c18adef00d910c330e

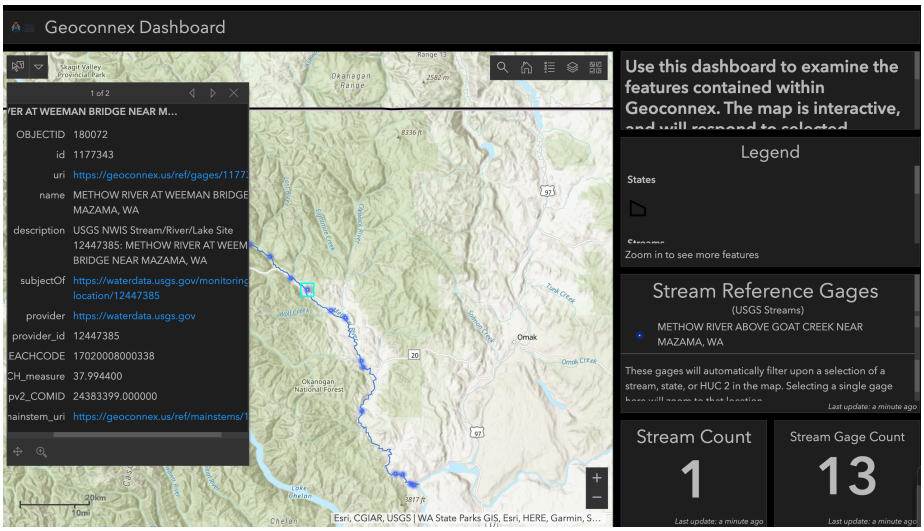

Figure 7: General Use Case: Finding monitoring locations by river, step 2

- creating and publishing domain models for specific feature types (e.g. anthropogenic water features)
- scaling participation in the system from data providers through targeted implementation of domain use cases that actively use the Geoconnex infrastructure
- creating institutional mechanisms to steward reference feature collections

6.2 Discovering, Distinguishing, and Integrating Hydrologically Relevant Observed and Modeled Datasets

6.2.1 Description

This use case involves a hydrologist a who aims to integrate historical observed and modeled streamflow data for a specific hydrologic feature or region in order to validate a novel model under development and compare it to other models and observations relevant to the same time and place.

6.2.2 User Story

As a Hydrologist, I want to integrate all observed and modeled data for a specific hydrologic feature or region, so that I can analyze all available models together and compare them to observed data to identify patterns, trends, and discrepancies and make recommendations for improving hydrologic models or data collection processes.

6.2.3 Conceptual Demonstration

The User first identifies the hydrologic feature or region of interest, perhaps a polygon for a custom-delineated area for which they developed their novel model.

The User could then query the geoconnex reference feature server for all mainstem rivers within that polygon using a CQL filter:

https://reference.geoconnex.us/collections/mainstems/items?filter=INTERSECTS(geom, POLYGON((-79.0228331610021%2035.81119501416316,%20-79.0195124291939%2035.88387125491053,%20-79.00622950196069%2035.916150422114285,%20-79.00456913605635%2035.97529469562092,%20-79.0643423086061%2035.97529469562092,%20-79.14570023791025%2035.960512777684514,%20-79.14902096971844%2035.88387125491053,%20-79.12577584706017%2035.84081189259446,%20-79.0228331610021%2035.81119501416316)))

This results in two relevant rivers, Little Creek and Morgan Creek:

Conceptually, there could be multiple streamflow monitoring locations and modeled streamflow datasets with features of interest relevant to the two rivers. For example, USGS has a gage on Morgan Creek, for which it could publish the following JSON-LD in its landing content:

```
{
  "@context": {
    "@vocab": "https://schema.org/",
    "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
    "dc": "http://purl.org/dc/terms/",
    "qudt": "http://qudt.org/schema/qudt/",
    "qudt-units": "http://qudt.org/vocab/unit/",
    "qudt-quantkinds": "http://qudt.org/vocab/quantitykind/",
    "gsp": "http://www.opengis.net/ont/geosparql#",
    "locType": "http://vocabulary.odm2.org/sitetype",
    "odm2var":"http://vocabulary.odm2.org/variablename/",
    "odm2varType": "http://vocabulary.odm2.org/variabletype/",
    "hyf": "https://www.opengis.net/def/schema/hy_features/hyf/",
    "skos": "https://www.opengis.net/def/schema/hy_features/hyf/HY_HydroLocationType",
    "ssn": "http://www.w3.org/ns/ssn/",
    "ssn-system":  "http://www.w3.org/ns/ssn/systems/"
  },
  "@id": "https://geoconnex.us/usgs/monitoring-location/08282300",
  "@type": [
    "hyf:HY_HydrometricFeature",
    "hyf:HY_HydroLocation",
    "locType:stream"
  ],
```

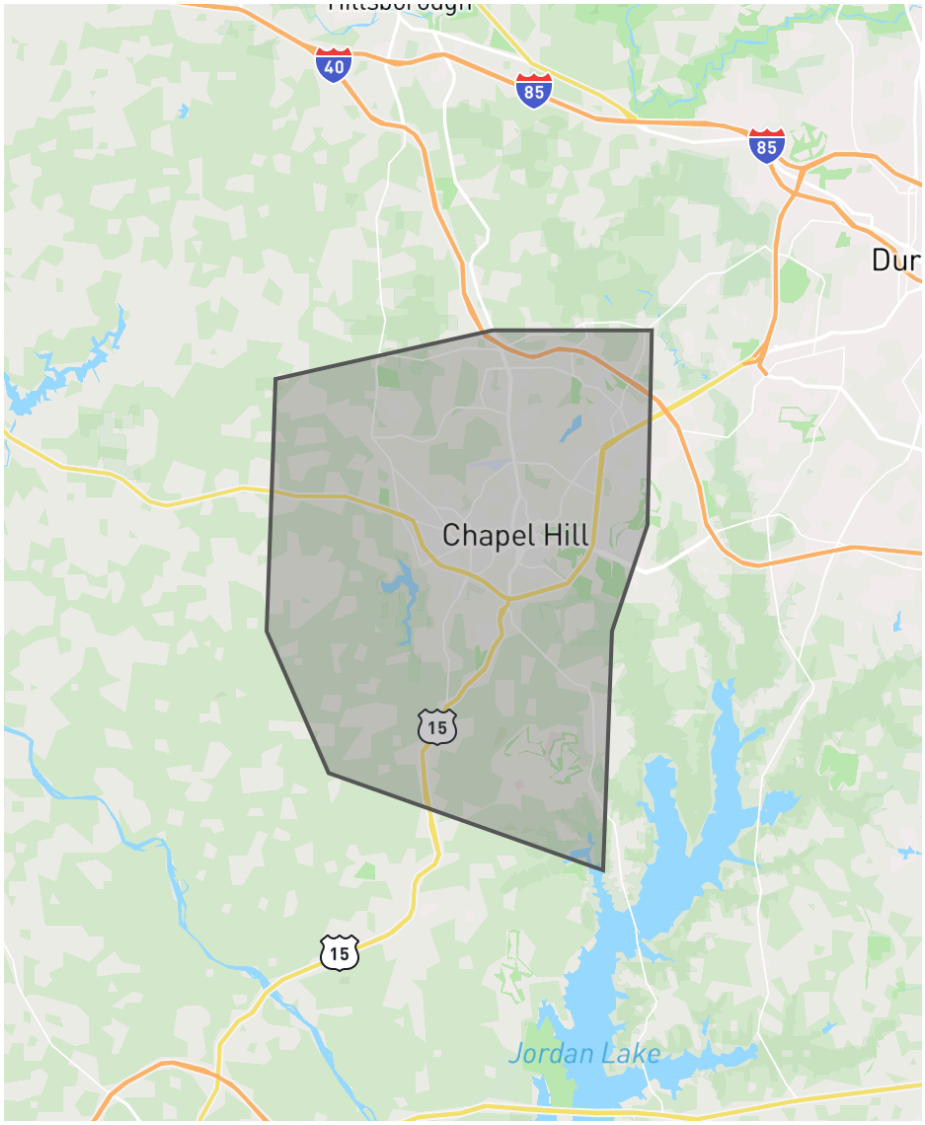

Figure 8: Observed/Modeled Use Case example area of interest

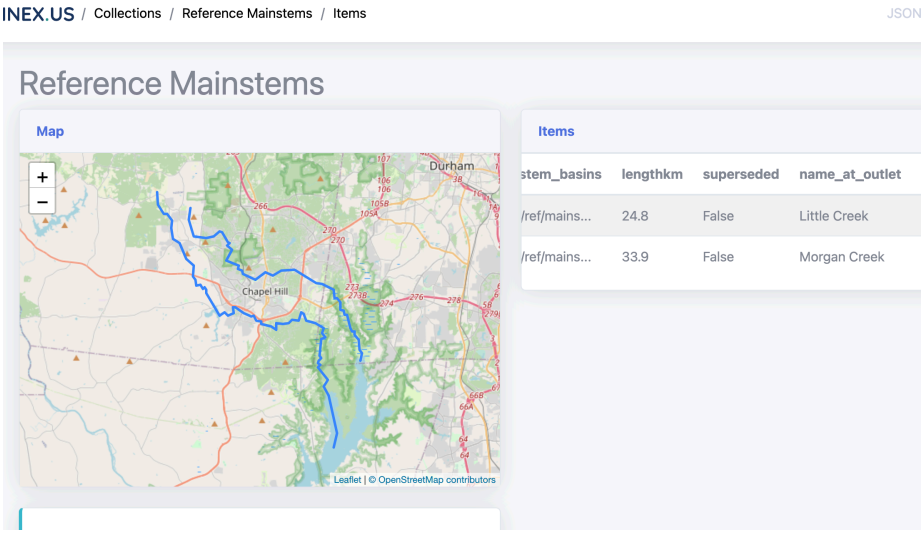

Figure 9: Observed/Modeled Use Case example mainstems of interest


```
  "hyf:HydroLocationType": "hydrometric station",
  "sameAs": {
    "@id": "https://geoconnex.us/ref/gages/1018463"
  },
  "identifier": {
    "@type": "PropertyValue",
    "propertyID": "USGS site number",
    "value": "02097517"
  },
  "name": "Morgan Creek Near Chapel Hill",
  "description": "Stream/River Site",
  "provider": {
    "url": "https://waterdata.usgs.gov",
    "@type": "GovernmentOrganization",
    "name": "U.S. Geological Survey Water Data for the Nation"
  },
  "geo": {
    "@type": "schema:GeoCoordinates",
    "longitude": -106.4707722,
    "latitude": 36.7379333
  },
  "gsp:hasGeometry": {
    "@type": "http://www.opengis.net/ont/sf#Point",
    "gsp:asWKT": {
```

```
        "@type": "http://www.opengis.net/ont/geosparql#wktLiteral",
        "@value": "POINT (-106.4707722 36.7379333)"
      },
      "gsp:crs": {
        "@id": "http://www.opengis.net/def/crs/OGC/1.3/CRS84"
      }
    },
    "hyf:referencedPosition": {
      "hyf:HY_IndirectPosition": {
        "hyf:linearElement": {
          "@id": "https://geoconnex.us/ref/mainstems/2408909"
        }
      }
    },
    "subjectOf": {
      "@type": "Dataset",
      "name": "Discharge data from USGS Monitoring Location 08282300",
      "description": "Discharge data from USGS Streamgage at Rio Brazos at Fishtail Road NR
      "variableMeasured": {
        "@type": "PropertyValue",
        "name": "discharge",
        "description": "Discharge in cubic feet per second",
        "propertyID": "https://www.wikidata.org/wiki/Q8737769",
        "url": "https://en.wikipedia.org/wiki/Discharge_(hydrology)",
        "unitText": "cubic feet per second",
        "qudt:hasQuantityKind": "qudt-quantkinds:VolumeFlowRate",
        "unitCode": "qudt-units:FT3-PER-SEC",
        "measurementTechnique": "observation",
        "measurementMethod": {
          "name": "Discharge Measurements at Gaging Stations",
          "publisher": "U.S. Geological Survey",
          "url": "https://doi.org/10.3133/tm3A8"
        }
      },
      "temporalCoverage": "2014-06-30/..",
      "ssn-system:frequency": {
        "value": "15",
        "unitCode": "qudt-units:Minute"
      },
      "distribution": [
        {
          "@type": "DataDownload",
          "name": "USGS Instantaneous Values Service"
          "contentUrl": "https://waterservices.usgs.gov/nwis/iv/?sites=2408909&parameterCd=
```

```
        "encodingFormat": [
          "text/tab-separated-values"
        ],
        "dc:conformsTo": "https://pubs.usgs.gov/of/2003/ofr03123/6.4rdb_format.pdf"
      },
      {
        "@type": "DataDownload",
        "name": "USGS SensorThings API",
        "contentUrl": "https://labs.waterdata.usgs.gov/sta/v1.1/Datastreams('0adb31f7852e
        "encodingFormat": [
          "application/json"
        ],
        "dc:conformsTo": "https://labs.waterdata.usgs.gov/docs/sensorthings/index.html"
      }
    ]
  }
}
```

Meanwhile the NOAA National Water Model could publish its model at the scale of NHDPlusV2 comid flowpaths with its own JSON-LD. Hypothetically, this would be for the model run for NHDPlusV2 comid 8896260:

```
{
  "@context": {
    "@vocab": "https://schema.org/",
    "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
    "dc": "http://purl.org/dc/terms/",
    "qudt": "http://qudt.org/schema/qudt/",
    "qudt-units": "http://qudt.org/vocab/unit/",
    "qudt-quantkinds": "http://qudt.org/vocab/quantitykind/",
    "gsp": "http://www.opengis.net/ont/geosparql#",
    "locType": "http://vocabulary.odm2.org/sitetype",
    "odm2var":"http://vocabulary.odm2.org/variablename/",
    "odm2varType": "http://vocabulary.odm2.org/variabletype/",
    "hyf": "https://www.opengis.net/def/schema/hy_features/hyf/",
    "skos": "https://www.opengis.net/def/schema/hy_features/hyf/HY_HydroLocationType",
    "ssn": "http://www.w3.org/ns/ssn/",
    "ssn-system":  "http://www.w3.org/ns/ssn/systems/"
  },
  "@id": "https://geoconnex.us/noaa/nwm/reachID/8896260",
  "@type": [
    "hyf:HY_HydrometricFeature",
  ],
  "hyf:HydroLocationType": "hydrometric station",
```

```
    "sameAs": {
      "@id": "https://geoconnex.us/nhdplusv2/comid/8896260"
    },
    "name": "Morgan Creek Near Chapel Hill",
    "description": "Stream/River Site",
    "provider": {
      "url": "water.noaa.gov",
      "@type": "GovernmentOrganization",
      "name": "NOAA National Water Model"
    },
    "hyf:referencedPosition": {
      "hyf:HY_IndirectPosition": {
        "hyf:linearElement": {
          "@id": "https://geoconnex.us/ref/mainstems/2408909"
        }
      }
    },
    "subjectOf": {
      "@type": "Dataset",
      "name": "Stream Flow Short-Range forecast from ReachID 8896260",
      "variableMeasured": {
        "@type": "PropertyValue",
        "name": "Stream Flow",
        "description": "Stream Flow in cubic feet per second",
        "propertyID": "https://www.wikidata.org/wiki/Q8737769",
        "url": "https://en.wikipedia.org/wiki/Discharge_(hydrology)",
        "unitText": "ft3/ second",
        "qudt:hasQuantityKind": "qudt-quantkinds:VolumeFlowRate",
        "unitCode": "qudt-units:FT3-PER-SEC",
        "measurementTechnique": "model",
        "measurementMethod": {
          "name": "NOAA National Water Model Short Range Forecast",
          "publisher": "NOAA",
          "url": "https://github.com/awslabs/open-data-docs/tree/main/docs/noaa/noaa-nwm-pd
        }
      },
      "temporalCoverage": "2016-06-30/..",
      "ssn-system:frequency": {
        "value": "1",
        "unitCode": "qudt-units:Hour"
      },
      "distribution": [
        {
          "@type": "DataDownload",
```

```
          "name": "NOAA NWM"
          "contentUrl": "https://water.noaa.gov/map?center_x=-8801823.929280883&center_y=42
          "encodingFormat": [
            "text/csv"
          ],
          "dc:conformsTo": "https://pubs.usgs.gov/of/2003/ofr03123/6.4rdb_format.pdf"
        }
      ]
    }
  }
```

The Geoconnex system in this example scenario would have harvested both of these documents, and thus there would be two monitoring locations referenced to the same mainstem via the `hyf:linearElement` property from HY_Features. Each of the datasets cover the same variable (streamflow) over overlapping `temporalCoverage`. The USGS dataset uses a `measurementTechnique` tag of `observation` and the NOAA dataset uses a `measurementTechnique` tag of `model`, while both have more detailed references to the specific data generating procedures specified in the `measurementMethod` block. `measurementTechnique` is designed to be a high level tag to delineate between broad categories of data generation methods for data cataloging and filtering purposes, with a limited codelist.

The following SPARQL query would with precision identify both datasets about streamflow, one with observed and one with modeled data, about one of the two mainstems from a larger universe of data for a variety of variables from many monitoring locations that are within in the initial area of interest, within a time period of interest.

```
PREFIX hyf: <https://www.opengis.net/def/schema/hy_features/hyf/>
PREFIX qudt: <http://qudt.org/schema/qudt/>
PREFIX qudt-quantkinds: <http://qudt.org/vocab/quantitykind/>
PREFIX ssn: <http://www.w3.org/ns/ssn/>
PREFIX schema: <https://schema.org/>

SELECT ?dataset ?dataDownloadURL ?measurementTechnique ?measurementMethodURL ?dataProvide
WHERE {
    ?dataset a schema:Dataset ;
             schema:variableMeasured ?variable ;
             schema:temporalCoverage ?temporalCoverage ;
             schema:provider ?dataProviderNode ;
             schema:distribution ?distribution .

    ?distribution schema:contentUrl ?dataDownloadURL ;
                  schema:encodingFormat ?encodingFormat .
```

```
17
18      ?variable qudt:hasQuantityKind qudt-quantkinds:VolumeFlowRate ;
19                ssn:measurementTechnique ?measurementTechnique ;
20                ssn:measurementMethod ?measurementMethod .
21
22      ?measurementMethod schema:url ?measurementMethodURL .
23
24      ?temporalCoverage schema:startDate ?timeRangeStart ;
25                        schema:endDate ?timeRangeEnd .
26
27      ?dataProviderNode schema:name ?dataProvider ;
28                        schema:url ?dataProviderURL .
29
30      FILTER (
31          str(?measurementTechnique) = "model" || str(?measurementTechnique) = "observation
32      )
33
34      FILTER (
35          str(?timeRangeStart) <= "2023-09-01"^^xsd:date &&
36          str(?timeRangeEnd) >= "2016-06-30"^^xsd:date
37      )
38  }
```

The resulting table would look like:

dataset	dataDownloadURL	measurementTechnique	measurementTechnique	measurementMethod	dataProvider	dataProviderURL	timeRangeStart	timeRangeEnd
Stream Flow Short-Range forecast from ReachID 8896260	NOAA NWM Data	model	NOAA National Water Model Documentation	NOAA National Water Model	NOAA National Water Model Website	2016-06-30		(ongoing)

dataset	dataDownloadURL	measurementType	measurementTechnique	MetadataProvider	dataProviderURL	timeRangeStart	timeRangeEnd
Discharge data from USGS Monitoring Location 08282300	USGS Instantaneous Values Service, USGS SensorThings API	observation	Discharge Measurements at Gaging Stations	U.S. Geological Survey Water Data for the Nation	USGS Water Data for the Nation Website	2014-06-30	(ongoing)

6.2.4 Next steps

This use case demonstrates the value of the Geoconnex approach, since highly relevant data could be precisely found without prior knowledge of any contributor data system. However, the system does depend on consistent and widespread method specification and feature of interest tagging. Indexes for methods and features of interest and easy, automated ways to add these metadata elements to source data systems will likely be necessary for uptake by data providers, and should be investigated for feasibility in future phases of the Geoconnex project.

7 Governance

7.1 Functional Requirements Research

From November 2022 to June 2023, the team engaged Federal, state, tribal, local, and NGO data providers through a variety of channels, including personal communications, conference presentations, webinars, and Internet of Water Coalition activities to solicit advice regarding what governance structures would be appropriate to improve the participation of data providers in the Geoconnex system. This advice was synthesized into a set of key preliminary functions for a Geoconnex governance framework:

- Define and refine the identification and stewardship of essential reference features.
- Establish metadata requirements for various data types, including time series, discrete sample data, remotely sensed data, statistical and administrative records data.
- Establish location metadata requirements for key data sets, including but not limited to: surface and groundwater monitoring water use and diversion locations; hydrologic cataloging features, e.g. watershed boundaries, HUCs, aquifers, interface with coastal data, administrative boundaries for

PWS and irrigation districts, groundwater management areas, conservation districts, census data, federal regions for EPA, USGS, USBR, USACE, NOAA, and other agencies.
- Create and oversee a data submission process: Review system for structural bias and other questions related to diversity, equity and inclusion (DEI), in collaboration with external groups. Encourage participation in the system.

To validate these requirements and inform a draft governance plan, survey of technical experts and Geoconnex superusers identified through the Internet of Water Coalition, followed by a virtual convening of the IoW Coalition's Geoconnex Working Group. This feedback was synthesized into a proposed governance plan to help guide future work to ensure that the Geoconnex system serves the needs of the Internet of Water community.

7.1.1 Survey Questions and Results

The following survey questions were posed to 7 experts who agreed to participate in the survey. The consensus syntheses of their responses following the convened discussion are summarized below each question, including some excerpted quotes from survey respondents.

1. **Is a governance mechanism for Geoconnex.us required?** Respondents recommended "Yes" to this question. "It is always good to have a mechanism for internal guidance and interfacing with the larger community, especially if the effort is cross-jurisdictional."

2. **What stakeholders should be involved in the leadership of this governance mechanism?** Respondents suggested that stakeholders should include a cross-section of large, medium, and small data providers, technical experts, API users, and end users, across the full data life cycle. "Data providers (regulated and unregulated data providers, local water systems, etc.), data stewards (agencies), and data users (academia, NGO advocacy groups, public, govt)."

3. **Should the governance mechanism be voluntary and informal with links to government agencies, or should it be formalized and led by a government agency?** Respondents generally favored a voluntary and informal structure, with links to government. "Perhaps there needs to be a core governance team and some offshoots that allow nimble communication on technical matters, innovation, etc."

4. **How large a governance mechanism is required for decision-making, and should it have an expert character or a representative character?** Respondents expressed a preference for an inclusive, but small, mechanism for the core business of setting up and defining Geoconnex. Respondents also suggested that results be reviewed by a wider audience. "If you think about the larger goal which in my mind that all

water data has a Geoconnex PID then I think you need a somewhat larger governance with clear structures for how decisions are made within the system."

5. **What mechanisms are necessary to engage a broader range of stakeholders?** Respondents suggested funding support, strong individual leadership, working examples by sector and agency, virtual webinars, and targeted outreach to specific audiences that would have an interest in certain specific reference features.
6. **What steps should be taken to ensure diversity, equity, and inclusion in this governance mechanism?** Respondents answered varied widely. Some suggested centering the user needs of a set of underrepresented, over-burdened communities in the context of a larger DEI strategy and ensure representation in the IoW Coalition of organizations that represent those communities.
7. **Is a new governance mechanism(s) necessary, or can existing community governance mechanisms be used for this purpose?** If so, which ones?: Most respondents agreed that a new mechanism is needed because of the uniqueness of the system.
8. **Should the governance mechanism sunset after its work is complete, or will there be an ongoing need for system governance?** If the latter, at what time interval should the governance mechanism be reviewed? Most respondents recommended a period of 1 to 3 years for the new governance mechanism, with a review after that period.

7.2 Proposed Governance Plan

The following two-part framework is proposed to support Geoconnex governance:

7.2.1 The Geoconnex Working Group

A voluntary, informal, technical working group of experts convened by the Center for Geospatial Solutions (CGS) through the Internet of Water (IoW) Coalition under the auspices of a cooperative agreement with the U.S. Geological Survey (USGS). The working group will consist of representatives or liaisons from the non-profit, academic, and private sector to develop recommendations concerning the functional questions of Geoconnex governance over a period of three years. Public agency representatives from federal, state, local, tribal, and territorial public agencies may participate as liaisons and contribute to discussions but will not contribute to the consensus of the working group. The recommendations of the working group will be synthesized and published as a draft technical report of the Center for Geospatial Solutions at the Lincoln Institute of Land Policy, and submitted to the full IoW Coalition and other forums for

review and comment. The forums invited to review the recommendations will include the Earth Science Information Partnership (ESIP), the federal roundtable on water data coordination sponsored by DOE, the CUAHSI network, and professional societies of water data users including AWRA, NAQWA, NWQMC. Following the comment period, CGS will reconcile the community comments for final publication and submission to USGS.

7.2.2 USGS-EPA Joint Committee on Geoconnex

A joint committee of 3 representatives of USGS and 3 representatives of EPA Office of Water, convened and co-chaired by CGS under the auspices of its cooperative agreement with USGS. The Joint Committee will consider the recommendations of the Geoconnex Working Group and finalize decisions concerning governance of the system, to be published by the U.S. Geological Survey.