H41L-1877 - Dendra: a real-time cloud-based time-series curation system

Collin Bode[1] and J. Scott Smith[1]

[1]University of California, Berkeley

November 23, 2022

Abstract

Wireless sensor networks for environmental monitoring are becoming a common tool for researchers across many of the field sciences. However, managing these systems is still an emerging issue. Internet of Things (IoT) technologies have provided many tools for creating big data solutions to these issues, but the industry is at cross-purposes with scientists. The big data approach is to collect massive amounts of data then throw out the anomalous points. For environmental monitoring, we need to archive and curate all the data as a permanent record of rapid environmental change. To achieve this, we need to combine IoT with museum curation sensibility. Dendra is cyberinfrastructure for real-time sensor data storage, retrieval, management, and curation for the field sciences. It is a cloud-based, multi-organizational system, designed to support massive permanent monitoring efforts (https://dendra.science). The name is derived from dendritic networks, such as river networks. Environmental monitoring performs in a similar manner, pulling data from the earth's surface to a single location. To curate streaming data, we developed a dynamic data versioning system. A field scientist reports invalid data from the field via mobile phone, the annotation is approved by curator, and is instantly applied to all data accessed. Data is only modified on extract. This allows us to pull data from any time in the past with the edits and calibrations of that time. Networked data logger integration works with LoggerNet, GOES satellite, and soon Iridium satellite. Dendra is hosted on NSF's XSEDE Jetstream cloud service. The system is designed as a set of microservices that interact through as set of persistent ques (NATS). Server-side javascript with Node.js is the primary development language. A data abstraction layer allows for multiple time-series databases (InfluxDB, MySQL, etc) to be accessed, even for a single instrument over time and reassembled as a single datastream. Access is via REST API & website. Dendra is used and supported by: Eel River Critical Zone Observatory (23 stations) in Mendocino, California; the University of California Natural Reserve System (25 stations); the Moore Foundation funded California Heartbeat Initiative (4 stations, 10 mobile, 40 planned).

H41L-1877 dendra a real-time cloud-based time-series curation system

https://dendra.science

mobile friendly

AGU 100

Collin Bode[1] and J. Scott Smith[2]

1. collin@berkeley.edu, Eel River CZO, UC Natural Reserve System, UC Berkeley, Berkeley, CA, USA, ORCID: 0000-0002-9654-6352
2. jscottsf@berkeley.edu, Eel River CZO, UC Natural Reserve System, UC Berkeley, Berkeley, CA, USA

What is it?

Dendra is cyberinfrastructure for real-time sensor data storage, retrieval, management, and curation for the field sciences. It is a cloud-based, multi-organizational system, designed to support massive permanent monitoring efforts.

dendra is short for *dendritic network*

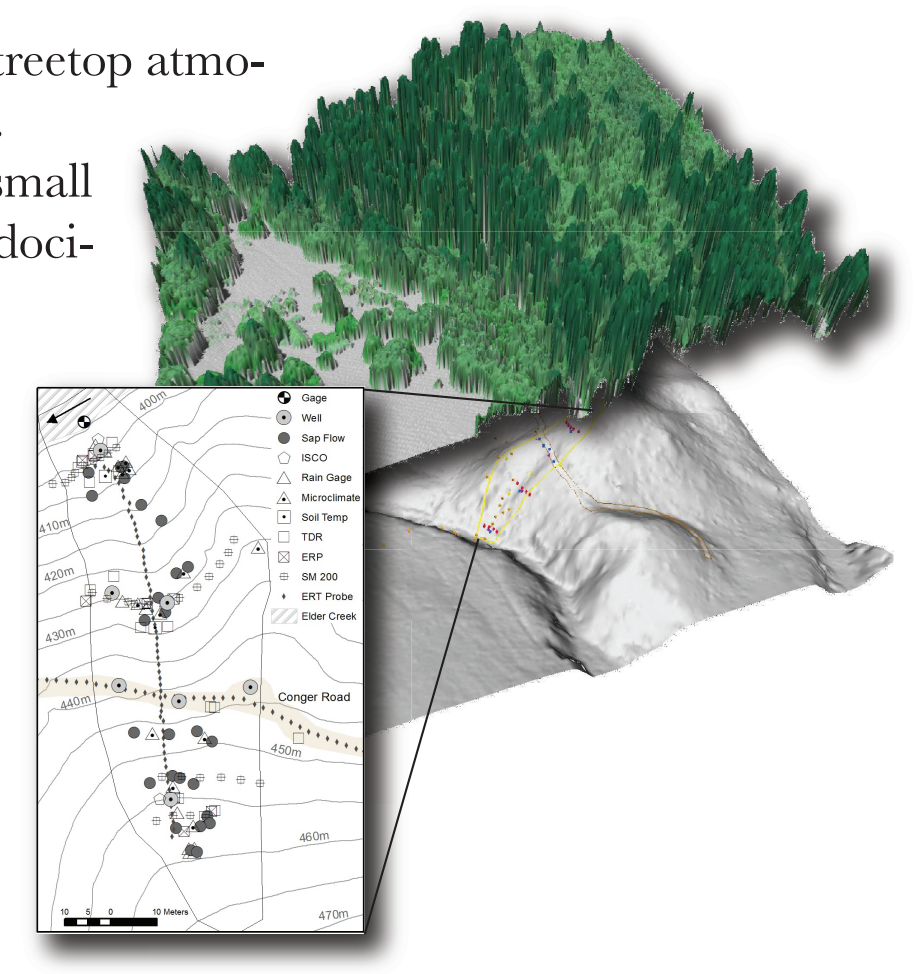
Who is it for?

Dendra is for organizations performing ongoing environmental monitoring who are interested in developing long-term high quality time-series datasets. It really shines for managing sensor observatories and mesonets.

Sensor Observatories using Dendra

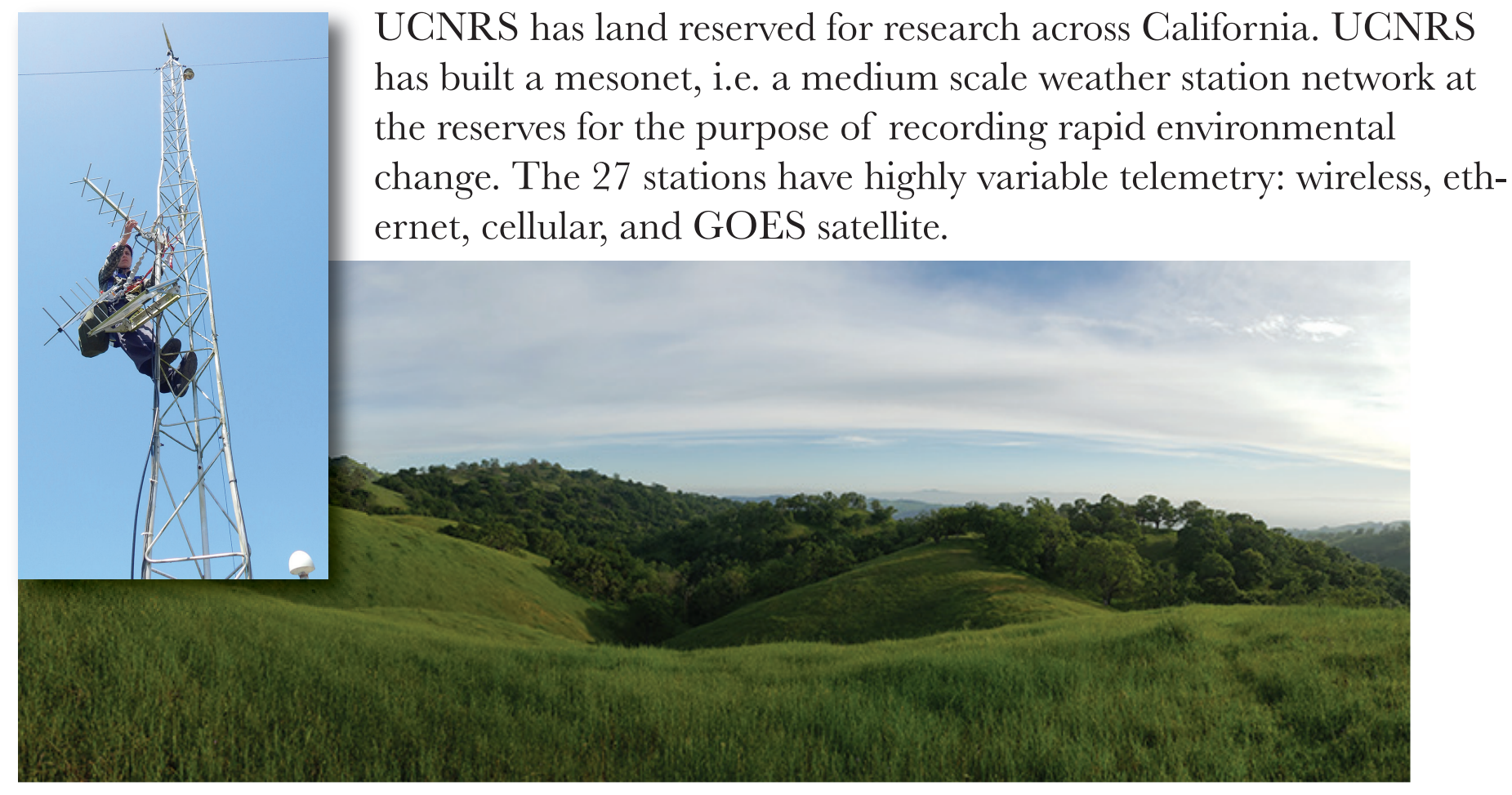
Eel River Critical Zone Observatory http://criticalzone.org/eel

ERCZO studies the critical zone – the area from treetop atmosphere to bedrock – through intensive monitoring. ERCZO has over 800 sensors on 23 stations in a small first-order watershed in the Angelo Reserve, Mendocino, CA. Telemetry is mostly high speed wireless.



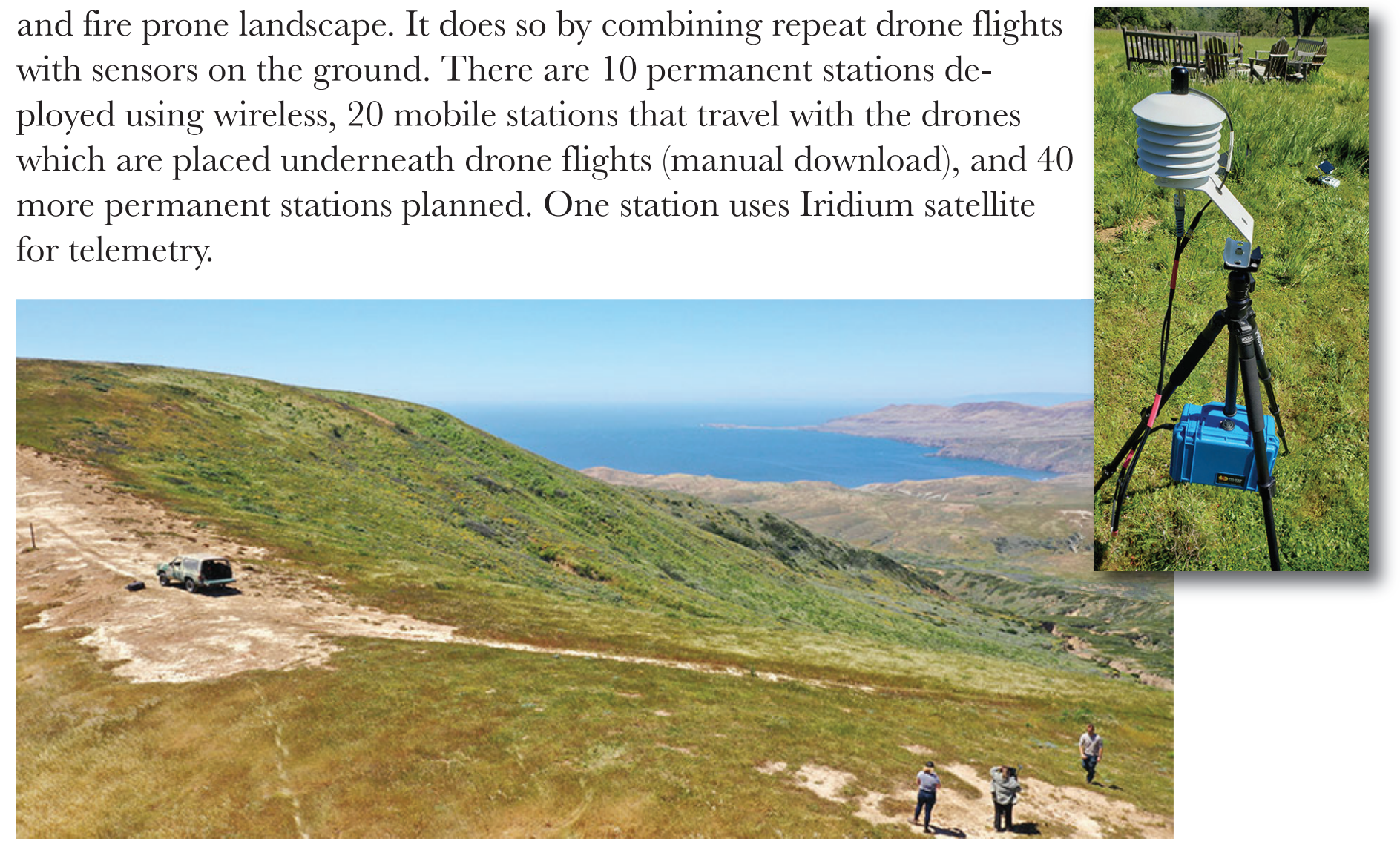
University of California Natural Reserve System https://ucnrs.org

UCNRS has land reserved for research across California. UCNRS has built a mesonet, i.e. a medium scale weather station network at the reserves for the purpose of recording rapid environmental change. The 27 stations have highly variable telemetry: wireless, ethernet, cellular, and GOES satellite.



California Heartbeat Initiative https://ucnrs.org/california-heartbeat-initiative

CHI studies plant moisture and microrefugia in California's drought and fire prone landscape. It does so by combining repeat drone flights with sensors on the ground. There are 10 permanent stations deployed using wireless, 20 mobile stations that travel with the drones which are placed underneath drone flights (manual download), and 40 more permanent stations planned. One station uses Iridium satellite for telemetry.



F.A.I.R. Compliance

Findable – Within site faceted search. REST API for site agregators. Future: implementing EarthCube GEOCODES search.

Accessible – Access control allows for granular openness while still embargoing data prior to publication or due to sensitive location, e.g. endangered plants.

Interoperable – Controlled Vocabulary agnostic. Can include new CV's as needed. API allows system-to-system atomic queries.

Re-Usable – Equipment Library, Station/Datastream metadata, and Annotations describing the history of the site, data, and condition of the equipment. Extremely rich metadata for extending the long tail of data.

Terms that make the system work

Metadata and the function of Dendra is structured around a very simple logical heirarchy:

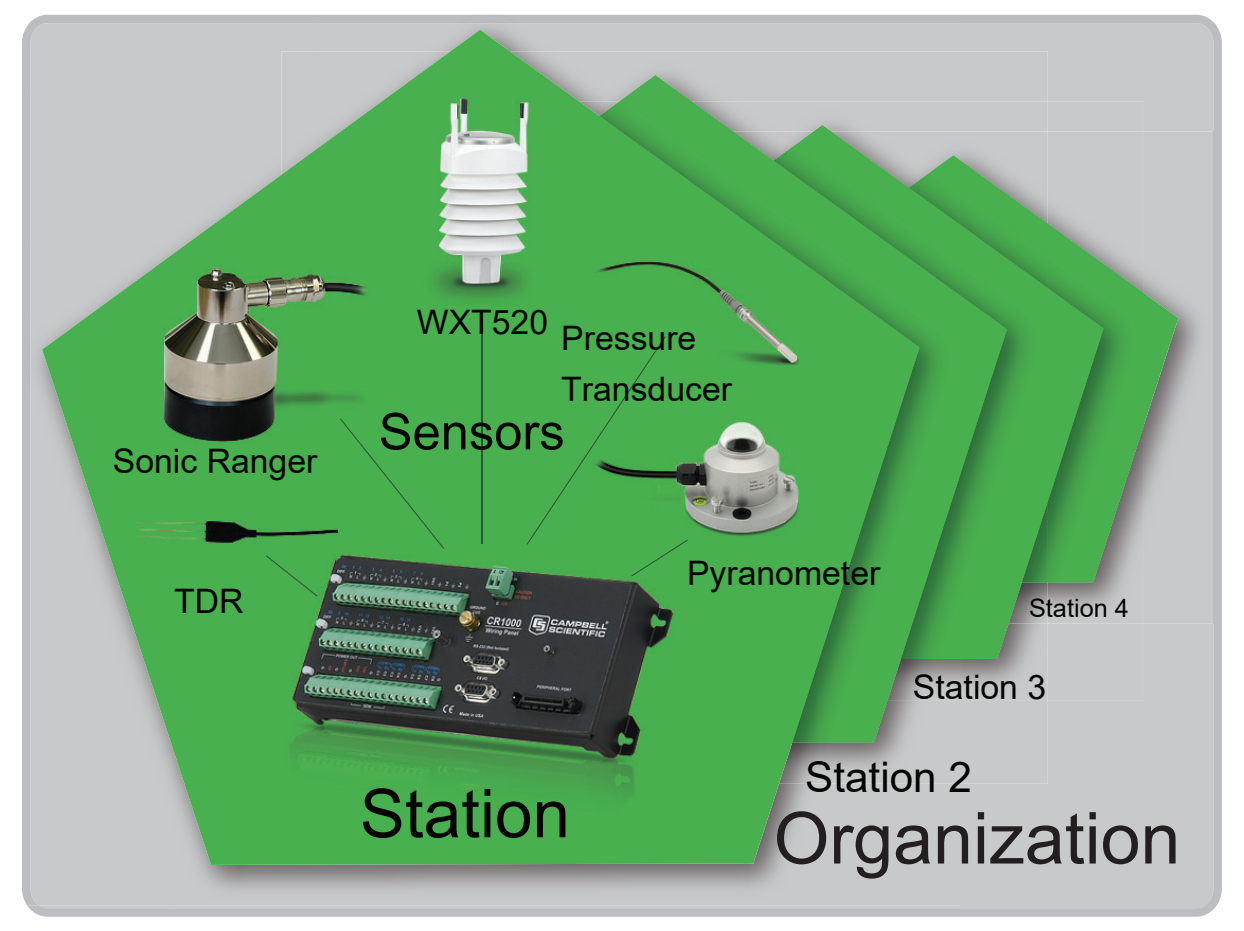
Organization – Group performing environmental monitoring.

Station – A datalogger at a location. All Stations belong to an organization.

Sensor – The type of instrument used for sensing. Each sensor may have one or more Datastreams, e.g. an R.M. Young 05103 Wind Monitor has wind speed and direction.

Datastream – Data that comes from a type of sensor at a location. Logically abstracted from sensor, so equipment swaps will not disrupt the Datastream.

Datapoints – Individual measurement pairs: timestamp and value.



Statistics

1.3 billion Datapoints
585,000 Datapoints/Day
2,700 Datastreams
1,600 Instruments
70 Sites

Controlled Vocabularies

Dendra uses controlled vocabularies for its internal operations. It is also designed to accept other organization's CVs. Current vocabularies implemented:

DQ – Data query terms
DT – Units (includes definitions for unit conversions)
DS – Internal metadata characterizing measurements
DW – NOAA Digital Weather Markup Language. Used in dashboards for trip planning.

Security & Access Control

Dendra has fine-grained access control that is designed to work specifically with how people use time-series data. Access rules are inherited from Organization to Station down to individual Datastreams. Specific rules can override the more general rules.

People
Public – no login
Members – part of Organization
Curator – always full access

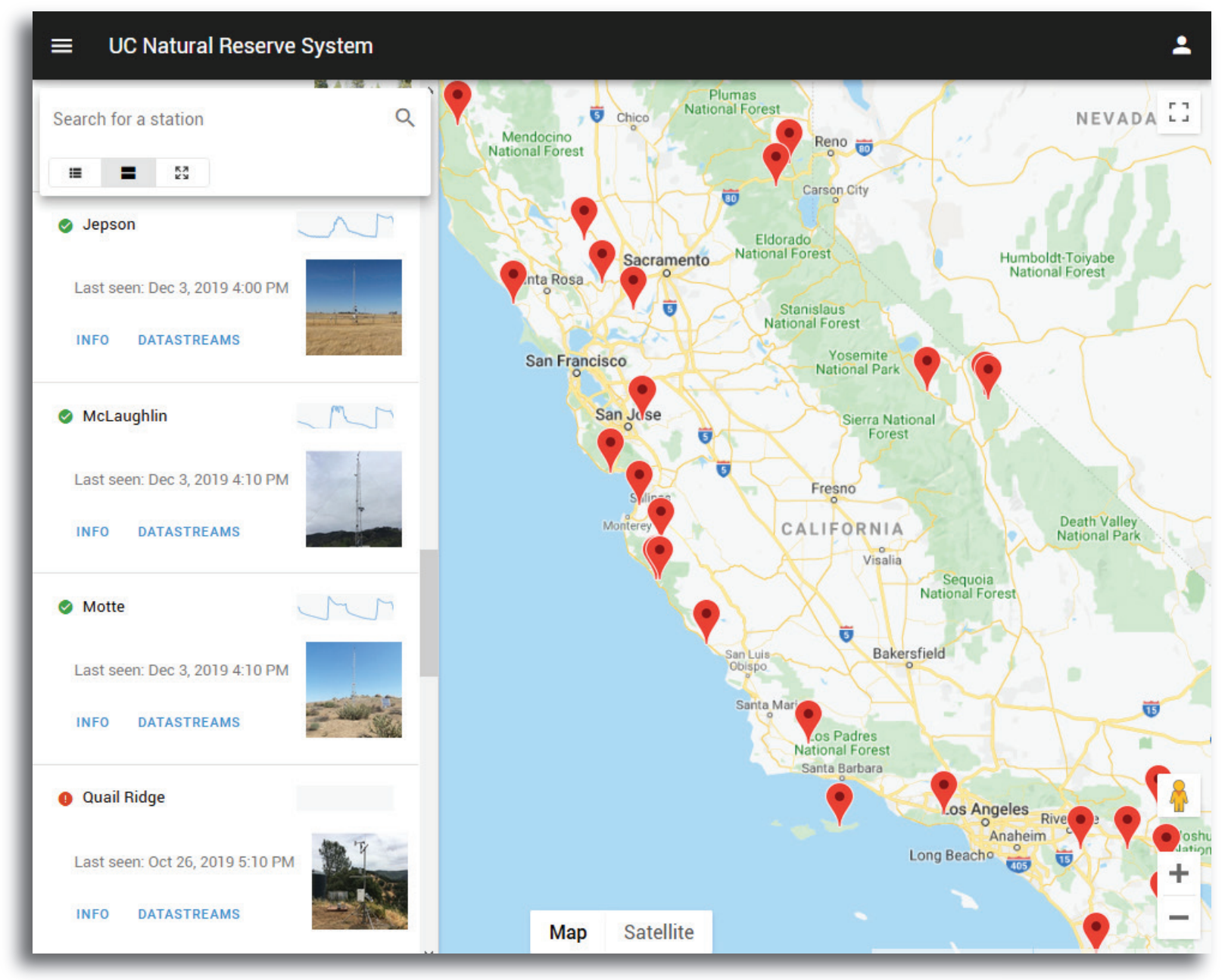
Data Access
0 – Restricted, the resource is not visible to users
1 – Metadata
2 – Graph, metadata
3 – Download, graph, metadata

Example

Organization	ERCZO	public: 1, members: 3	general policy
Station	Level 6	public: 1, members: 3	inherits policy
Datastream	L6 SapFlow	public: 0, members: 3	overrides and restricts access

Equipment Status Monitoring & Alerts

A critical need for large-scale monitoring is real-time status reporting. Dendra has a Status page showing last timestamp each station reported in, and a sparkline of battery voltage. Email alerts are sent if a logger has not reported in for 24 hours. Text alerts are planned.



Equipment Library

End users care about measured environmental variables. Datastreams characterize that need. To evaluate the quality of the these measurements, the end user needs to know the instrument type and its condition.

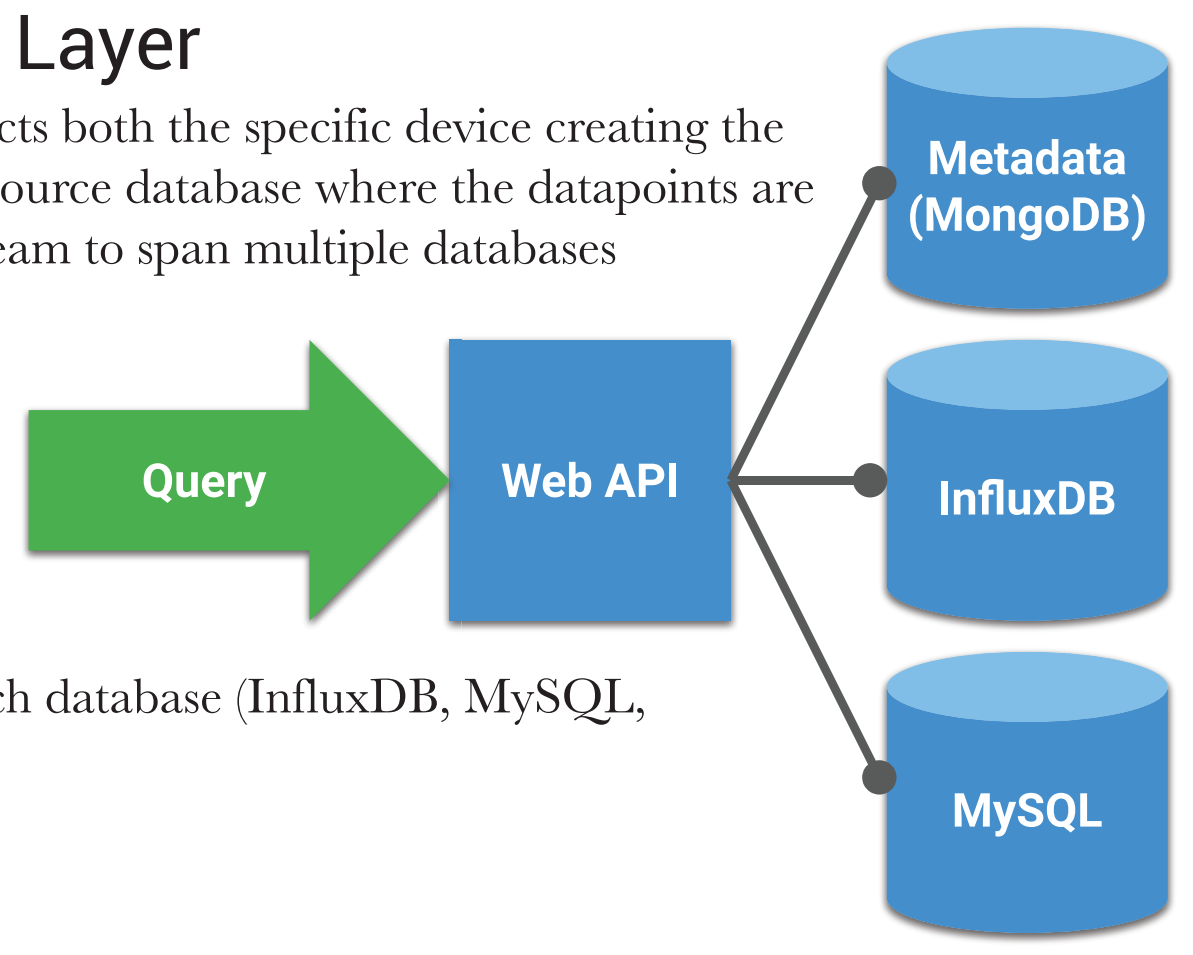
For facilities management, the instrument type and its condition are also critical metadata for proper operation of the observatory. Managers need to know how to calibrate, and what to replace if it breaks.

Using Internet of Things (IoT) terminology, Dendra has a library of "thing-types", or models of equipment. Each Station and Datastream has a *Thing* associated with it (datalogger, sensor, etc). A Thing or piece of equipment can be removed from one Station, recalibrated, and installed at another location. The equipment can be tracked separately from the Datastreams. Annotations are used to perform this tracking.

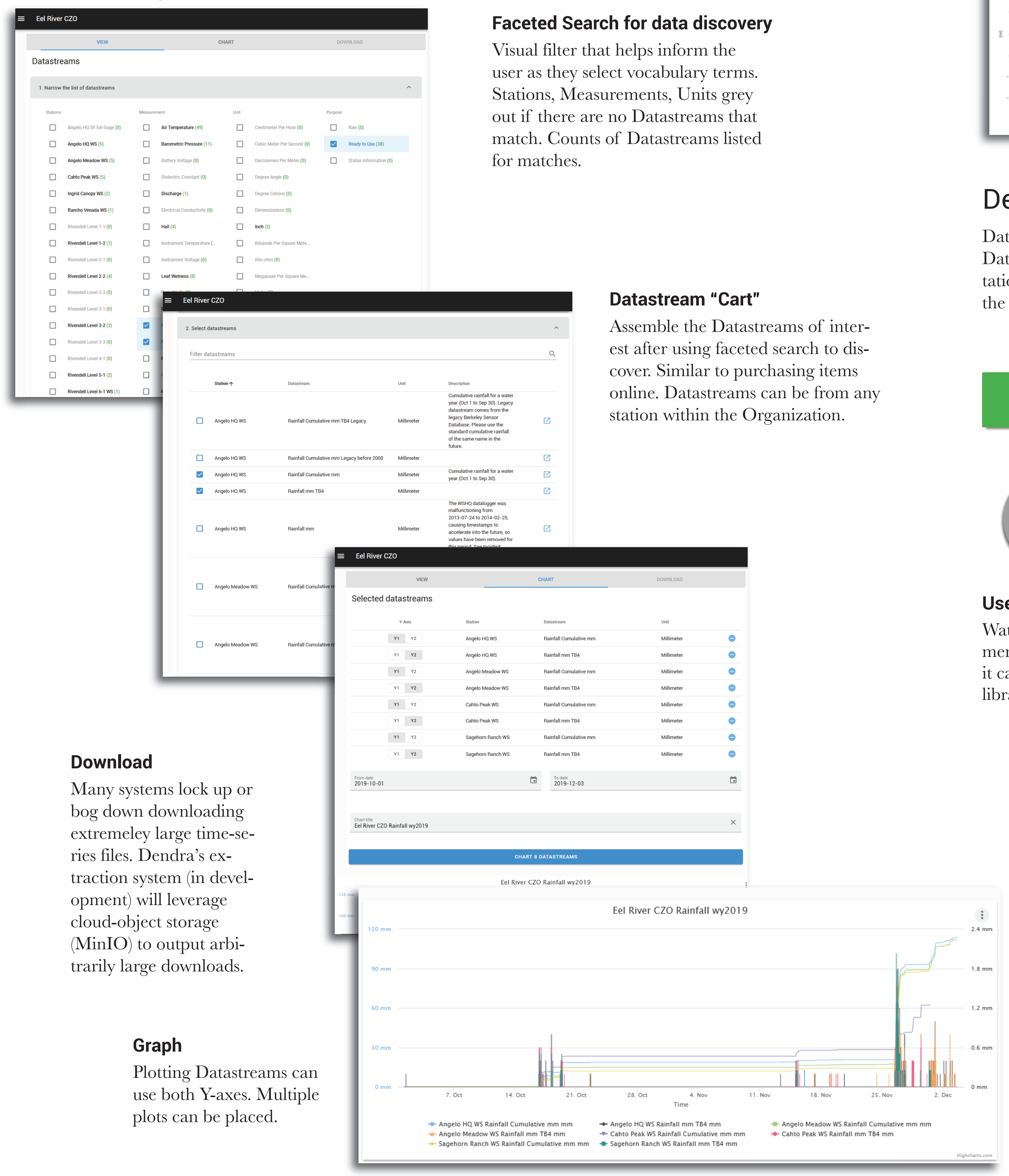
Data Abstraction Layer

Datastream metadata abstracts both the specific device creating the measurements, but also the source database where the datapoints are stored. This allows a datastream to span multiple databases seamlessly.

Abstraction Process
1. Parse query
2. Get metadata config
3. Prepare query steps
4. Fetch Datapoints from each database (InfluxDB, MySQL, Postgres, OpenTSDB, etc.)
5. Stitch and transform
6. Return data to caller



Search, Graph & Download


Eel River CZO
VIEW
CHART
DOWNLOAD
Datastreams
1. Narrow the list of datastreams
Eel River CZO
2. Select datastreams
Filter datastreams
Eel River CZO
VIEW
CHART
DOWNLOAD
Selected datastreams
From date 2019-10-01
To date 2019-12-03
Chart title Eel River CZO Rainfall wy2019
CHART 8 DATASTREAMS
Eel River CZO Rainfall wy2019
120 mm
90 mm
60 mm
30 mm
0 mm
2.4 mm
1.8 mm
1.2 mm
0.6 mm
0 mm
7. Oct
14. Oct
21. Oct
28. Oct
4. Nov
11. Nov
18. Nov
25. Nov
2. Dec
Time
Angelo HQ WS Rainfall Cumulative mm mm
Angelo HQ WS Rainfall mm TB4 mm
Angelo Meadow WS Rainfall Cumulative mm mm
Angelo Meadow WS Rainfall mm TB4 mm
Cahto Peak WS Rainfall Cumulative mm mm
Cahto Peak WS Rainfall mm TB4 mm
Sagehorn Ranch WS Rainfall Cumulative mm mm
Sagehorn Ranch WS Rainfall mm TB4 mm
Highcharts.com

Stations	Measurement	Unit	Purpose
☐ Angelo HQ SF Eel Gage (0)	☐ Air Temperature (49)	☐ Centimeter Per Hour (0)	☐ Raw (0)
☐ Angelo HQ WS (5)	☐ Barometric Pressure (11)	☐ Cubic Meter Per Second (0)	☑ Ready to Use (38)
☐ Angelo Meadow WS (5)	☐ Battery Voltage (0)	☐ Decisiemen Per Meter (0)	☐ Status Information (0)
☐ Cahto Peak WS (5)	☐ Dielectric Constant (0)	☐ Degree Angle (0)	
☐ Ingrid Canopy WS (2)	☐ Discharge (1)	☐ Degree Celsius (0)	
☐ Rancho Venada WS (1)	☐ Electrical Conductivity (0)	☐ Dimensionless (0)	
☐ Rivendell Level 1-1 (0)	☐ Hail (4)	☐ Inch (2)	
☐ Rivendell Level 1-2 (1)	☐ Instrument Temperature (...	☐ Kilojoule Per Square Mete...	
☐ Rivendell Level 2-1 (0)	☐ Instrument Voltage (0)	☐ Kilo-ohm (0)	
☐ Rivendell Level 2-2 (4)	☐ Leaf Wetness (8)	☐ Megajoule Per Square Me...	
☐ Rivendell Level 2-3 (0)			
☐ Rivendell Level 3-1 (0)			
☐ Rivendell Level 3-2 (3)			
☐ Rivendell Level 3-3 (0)			
☐ Rivendell Level 4-1 (0)			
☐ Rivendell Level 5-1 (2)			
☐ Rivendell Level 6-1 WS (1)			

	Station ↑	Datastream	Unit	Description
☐	Angelo HQ WS	Rainfall Cumulative mm TB4 Legacy	Millimeter	Cumulative rainfall for a water year (Oct 1 to Sep 30). Legacy datastream comes from the legacy Berkeley Sensor Database. Please use the standard cumulative rainfall of the same name in the future.
☐	Angelo HQ WS	Rainfall Cumulative mm Legacy before 2008	Millimeter	
☑	Angelo HQ WS	Rainfall Cumulative mm	Millimeter	Cumulative rainfall for a water year (Oct 1 to Sep 30).
☑	Angelo HQ WS	Rainfall mm TB4	Millimeter	
☐	Angelo HQ WS	Rainfall mm	Millimeter	The WSHQ datalogger was malfunctioning from 2013-07-24 to 2014-02-25, causing timestamps to accelerate into the future, so values have been removed for this period. See incident
☐	Angelo Meadow WS	Rainfall Cumulative		
☐	Angelo Meadow WS	Rainfall Cumulative		

Y-Axis	Station	Datastream	Unit
Y1	Angelo HQ WS	Rainfall Cumulative mm	Millimeter
Y2	Angelo HQ WS	Rainfall mm TB4	Millimeter
Y1	Angelo Meadow WS	Rainfall Cumulative mm	Millimeter
Y2	Angelo Meadow WS	Rainfall mm TB4	Millimeter
Y1	Cahto Peak WS	Rainfall Cumulative mm	Millimeter
Y2	Cahto Peak WS	Rainfall mm TB4	Millimeter
Y1	Sagehorn Ranch WS	Rainfall Cumulative mm	Millimeter
Y2	Sagehorn Ranch WS	Rainfall mm TB4	Millimeter

Faceted Search for data discovery
Visual filter that helps inform the user as they select vocabulary terms. Stations, Measurements, Units grey out if there are no Datastreams that match. Counts of Datastreams listed for matches.

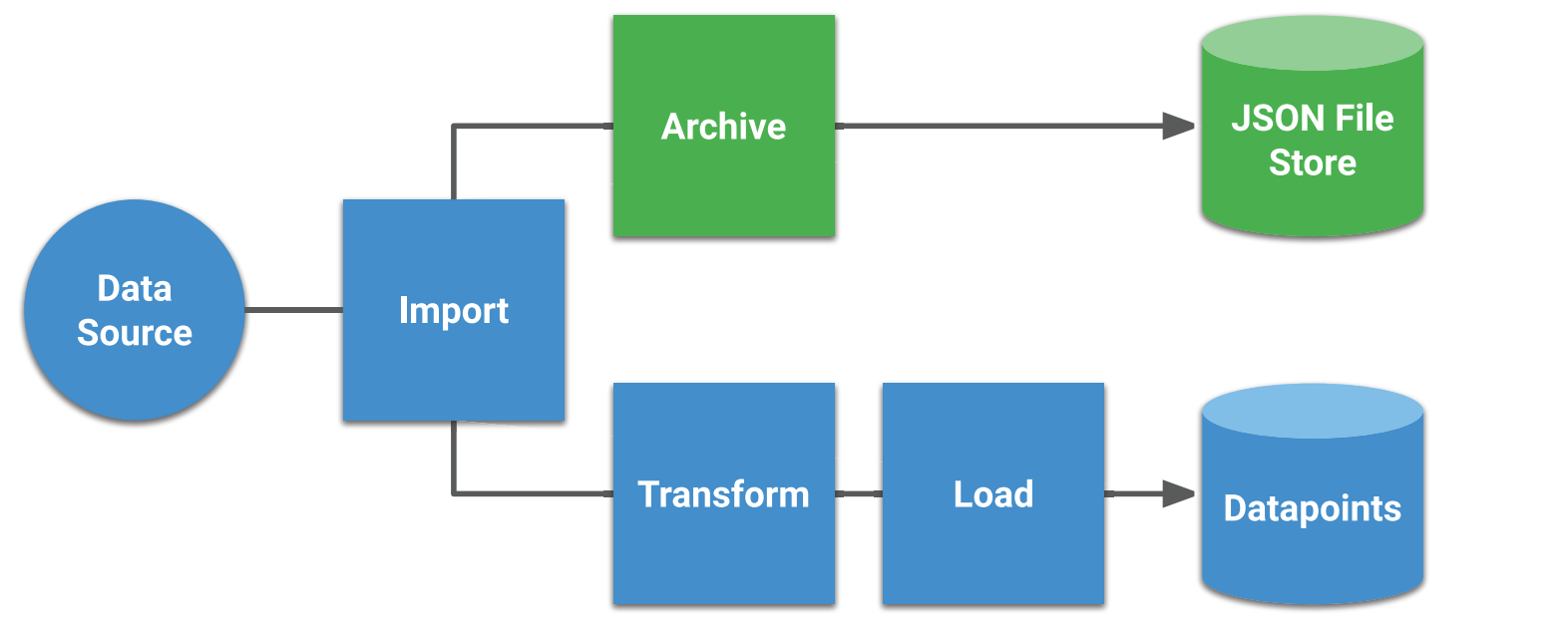
Datastream "Cart"
Assemble the Datastreams of interest after using faceted search to discover. Similar to purchasing items online. Datastreams can be from any station within the Organization.

Download
Many systems lock up or bog down downloading extremeley large time-series files. Dendra's extraction system (in development) will leverage cloud-object storage (MinIO) to output arbitrarily large downloads.

Graph
Plotting Datastreams can use both Y-axes. Multiple plots can be placed.

Simultaneous Archival & Loading

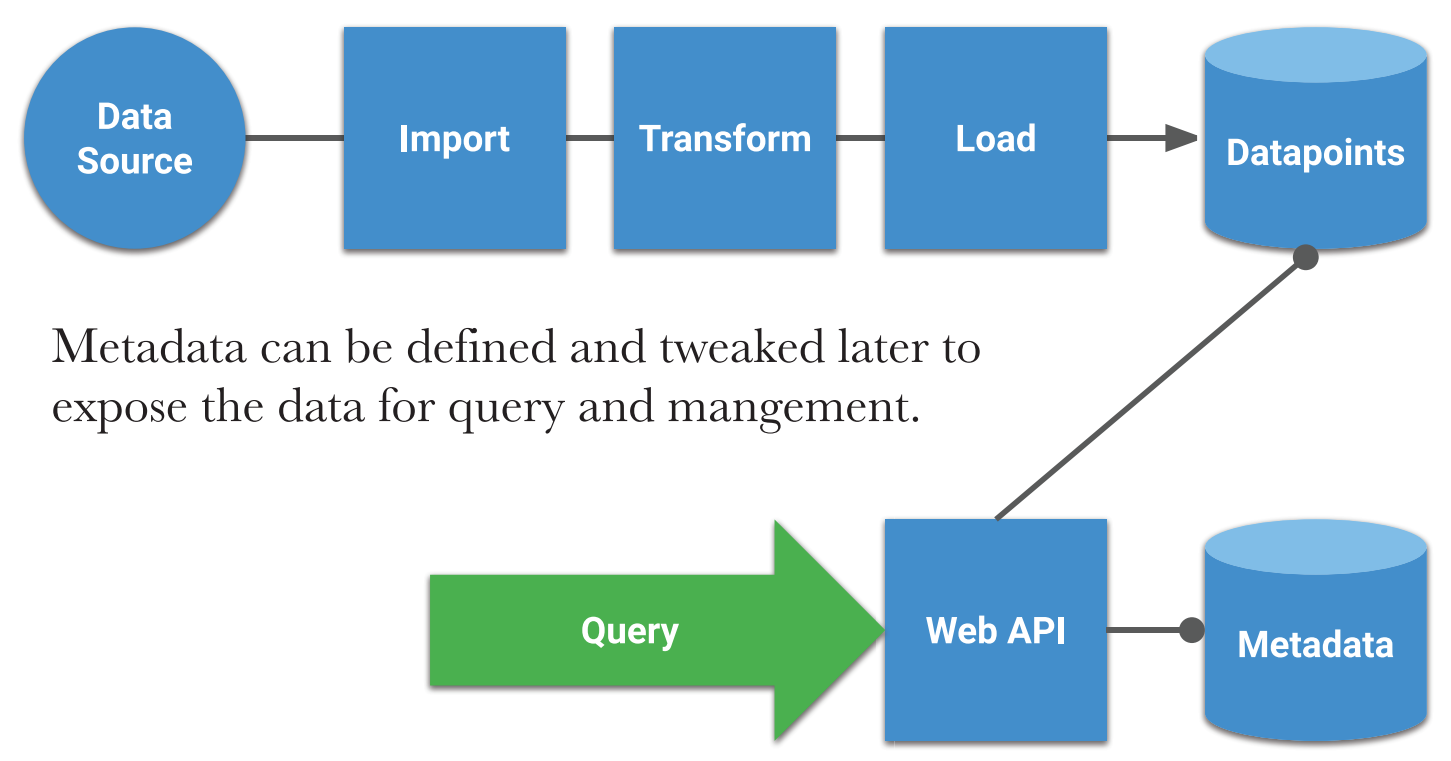
Source data is stored unmodified as individual JSON documents. This allows original records to be reloaded and reprocessed if needed.



A parallel pipeline transforms and loads source data into a time-series database for data consumers.

Naive Data Loading

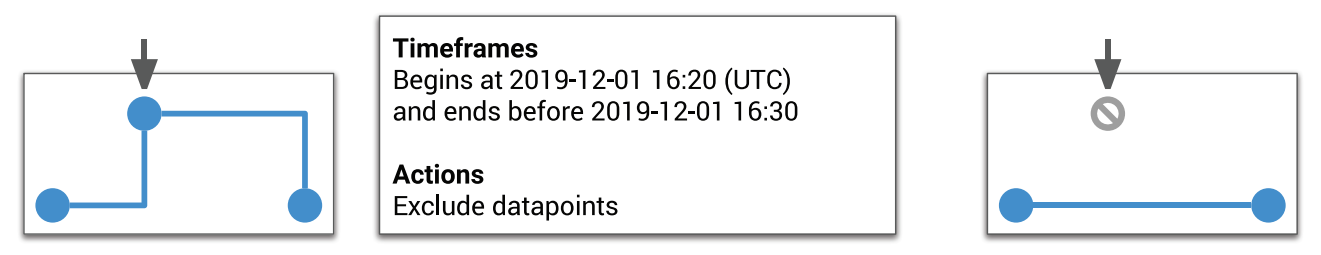
Metadata is not required upfront and is decoupled from data loading. This reduces the amount of configuration needed to set up a new data source for loading.



Metadata can be defined and tweaked later to expose the data for query and mangement.

Dynamic Dataset Versioning (Annotations)

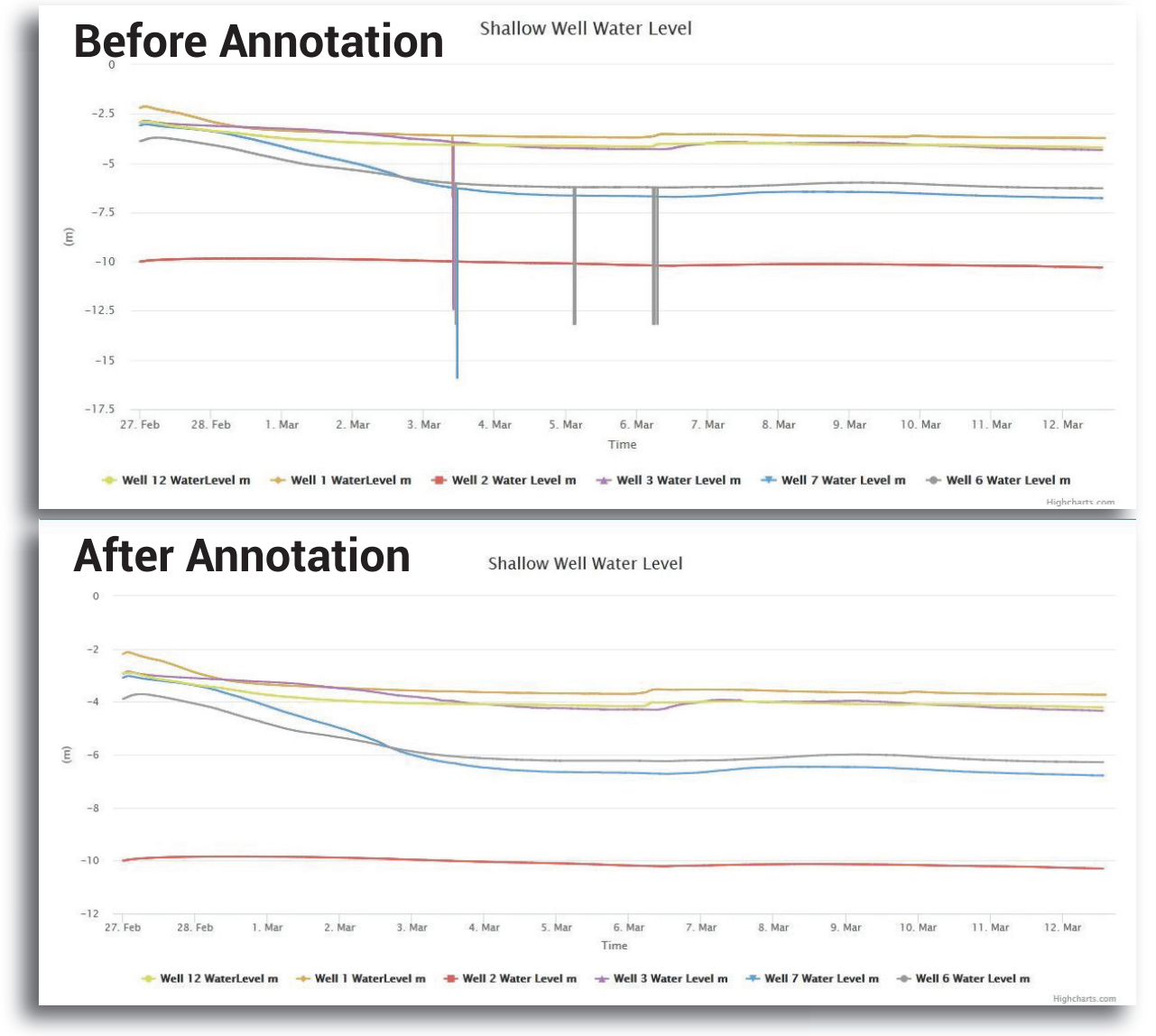
Dendra has the ability to dynamically version datasets, making it similar to a "GitHub" for time-series data. This is done through Annotations. Annotations are used to record events that happen to equipment, the environment, or data. Annotations are only applied when data is extracted. Source data is unchanged. This allows the user to "roll back" changes to an earlier time, if needed.



Timeframes
Begins at 2019-12-01 16:20 (UTC)
and ends before 2019-12-01 16:30

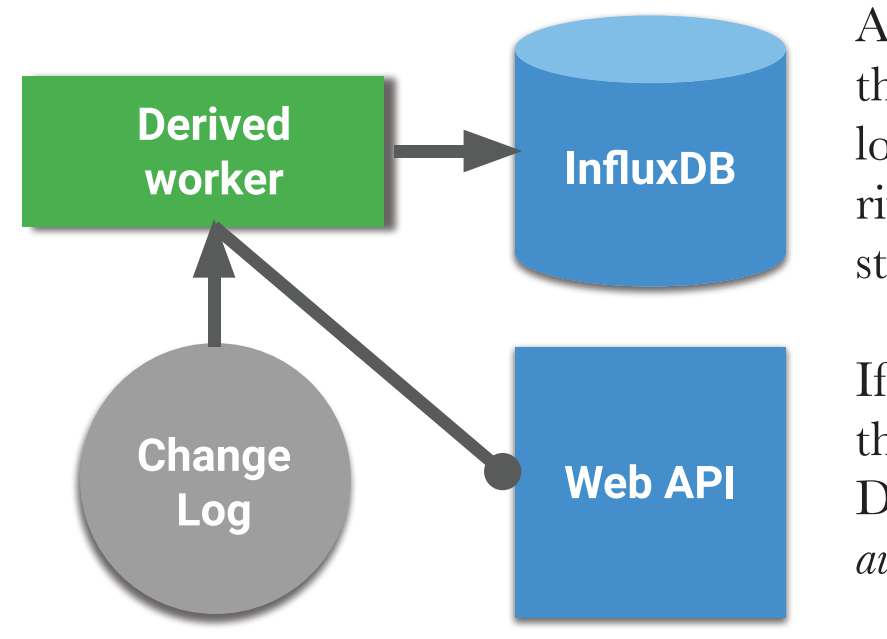
Actions
Exclude datapoints

Annotations
– Flag Datapoints
– Exclude Datapoints
– Evaluate calculations on data
– Only applied on retrieval
– Can be downloaded with data



Derived Datastreams

Datastreams are a logical abstraction of sensor data. Derived Datastreams take this a step further by providing a continuous computation engine that generates Datapoints using sensor Datastreams as the input.

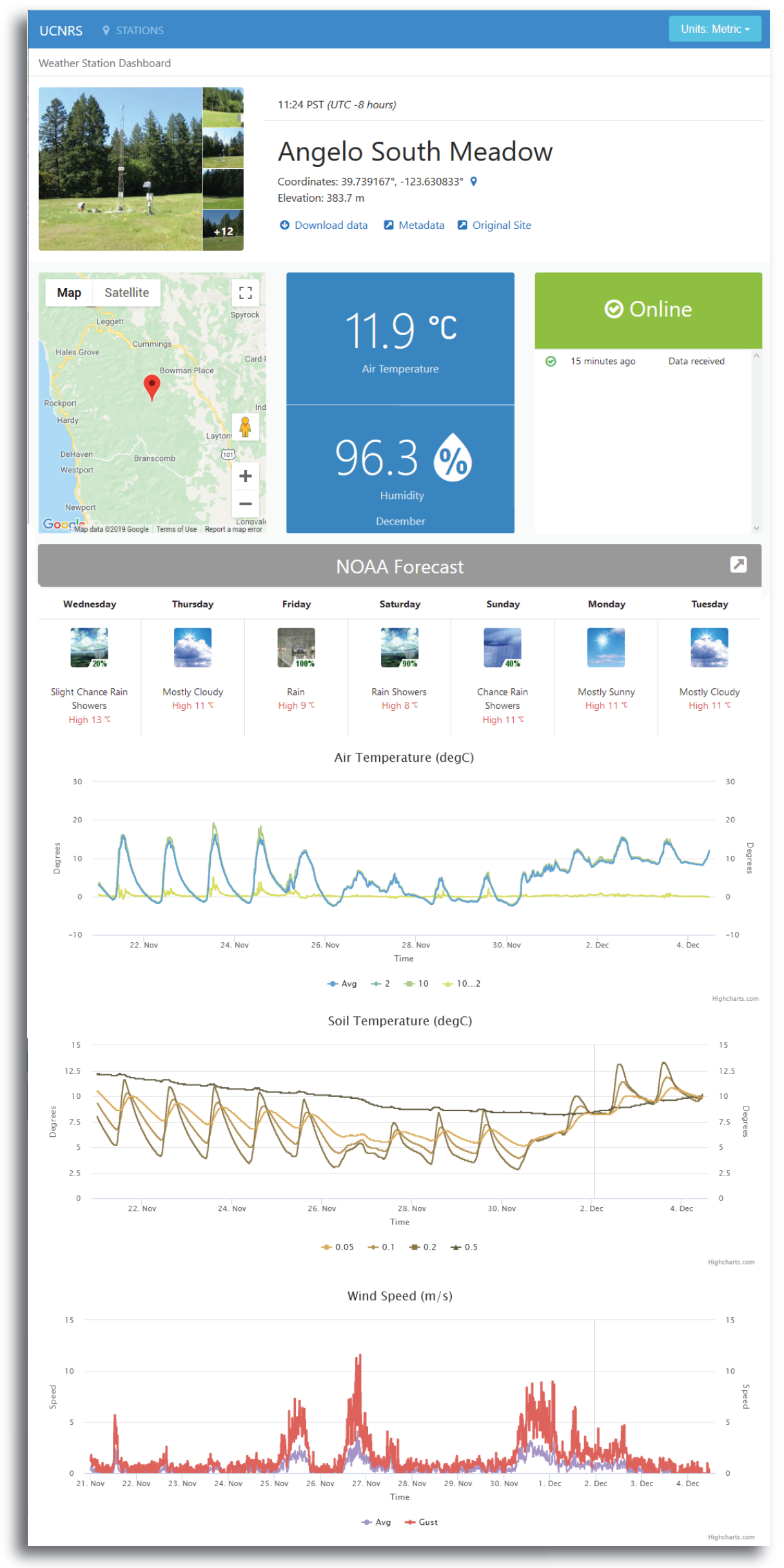


Derived Processing
As sensor data is loaded into the database, the changes are logged which trigger a Derived worker to compute and store new Datapoints.

If an Annotation is created that impacts the source Datastream, *the derived data is automatically recomputed.*

Use Case
Water year cumulatives are derived from sampled rainfall measurements. A derivation method wyCumulative is coded and installed so it can be used to define new Derived Datastreams. We are building a library of derivation methods.

Station Dashboards



REST API

Our API (Application Programming Interface) is a core building block of Dendra. Both applications and users can utilize it to access data and manage resources within the system.

The API is publicly accessible over the web, and gives users a consistent and reliable way to access Dendra's data and features across a variety of programming languages. This is regardless of how the system is hosted, or where the data is stored.

API v1 was available when Dendra first launched. This provided basic management and the first iteration of our data abstraction layer, which supported multiple time-series data stores right away. Full documentation for API v1 is available online at https://dendrascience.github.io/dendra-json-schema/

API v2 is now in beta, and improves upon the basic features by adding fine-grained access control to data, as well as enabling functional annotations that can impact Datapoints as they are retrieved.

Tech Stack

Web Application
CASL (authorization library)
Highcharts
Feathers (real-time REST API)
Math.js
Moment.js
Nuxt.js (web app framework)
Vuetify (Material Design framework)
Vue.js (UI framework)
JavaScript

Middleware
NATS Streaming (message queuing)
MinIO (cloud object storage)

Tools
GitHub (source control)
Trello (project and issue tracking)

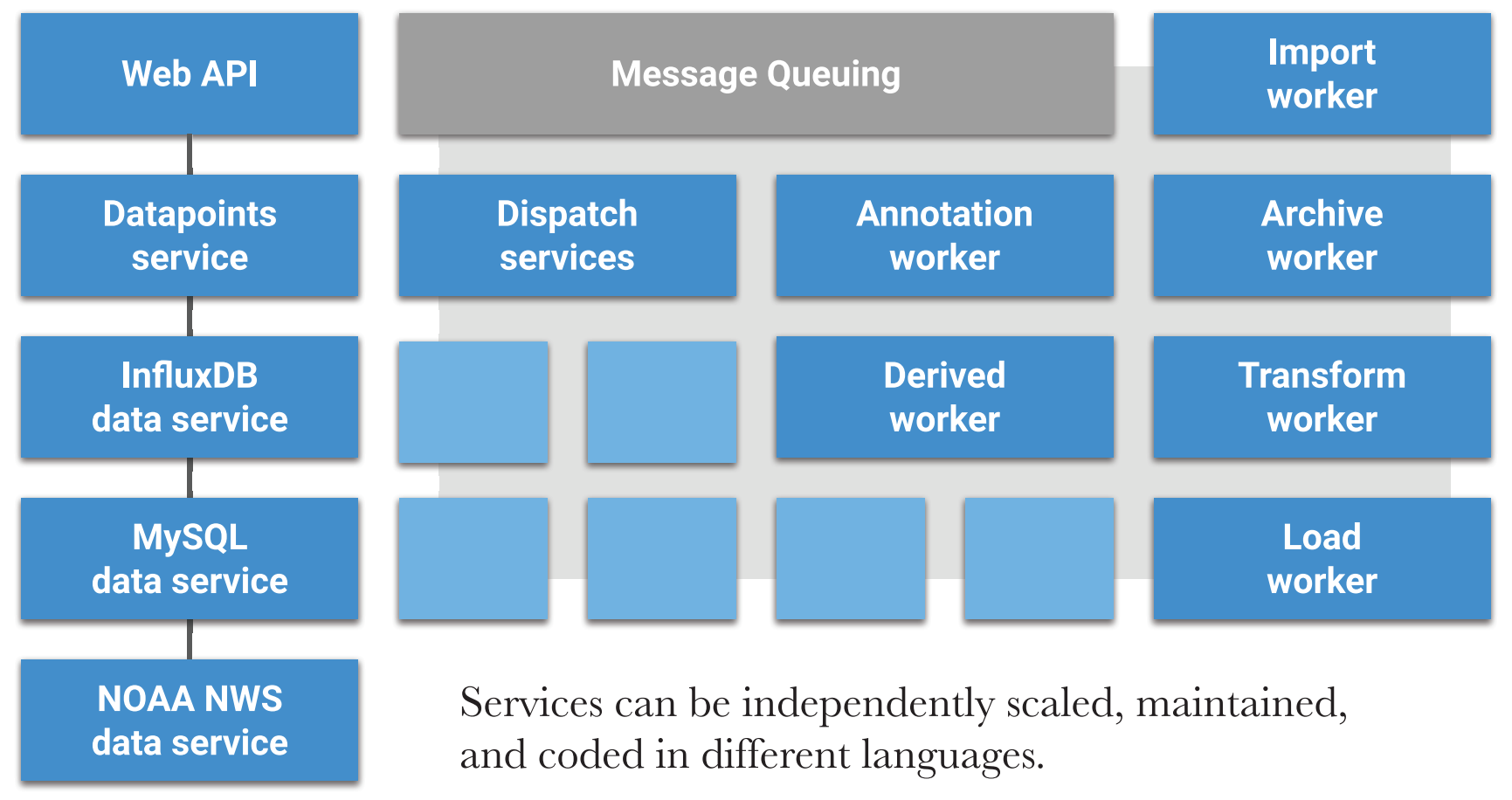
Databases
InfluxDB (time-series database)
MySQL (legacy time series data)
MongoDB (NoSQL metadata)

Backend services
CASL
Feathers (real-time REST API)
Pond.js (time-series data library)
Math.js
Moment.js
JSONata (query and transformation)
JSON Schema
Node.js (server-side JavaScript)

Infrastructure
Kubernetes/K8s (Linux container orchestration)
Docker (container platform)
Traefik (cloud edge router)
OpenStack (via XSEDE)

Microservices

Dendra has a microservices architecture, which atomizes system functions into smaller, composable pieces that work together.



Services can be independently scaled, maintained, and coded in different languages.

Infrastructure

Dendra is hosted on XSEDE **Jetstream**, which is an NSF-funded, user-friendly cloud environment designed to give researchers access to interactive computing and data analysis resources on demand.

Our system services run in multiple containers that provide isolation and portability while efficiently sharing the host operating system. Containers are scaled across multiple server nodes, and are managed by Kubernetes (K8s), which is an open-source platform that automates Linux container operations.

Presently our entire codebase is JavaScript, which runs in browsers for the frontend, and in Node.js on the backend. This allows for a high degree of consistency, portability and reuse across tiers.

Metadata is stored as JSON documents in MongoDB. Datapoints are currently stored by default in InfluxDB. Our data abstraction layer allows us to access Datapoints in other data stores, given an appropriate provider service is built.

XSEDE Extreme Science and Engineering Discovery Environment

Collaboration & Support

Eel River Critical Zone Observatory (ERCZO; NSF-EAR1331940)
University of California Natural Reserve System (weather station budget)
Gordon and Betty Moore Foundation's California Heartbeat Initiative (7063)
XSEDE Supercomputing Facilities, Jetstream (NSF-EAR190001)