An Evaluation and Recommendation for Data Standards for Continuous Monitoring Data
Draft data standards report by the E-Enterprise Advanced Monitoring Team
October 26, 2017

Executive Summary

The Advanced Monitoring Team is a collaborative effort under the E-Enterprise Initiative to understand how to use, communicate, and ensure the quality of advanced monitoring data, especially continuous monitoring data. Environmental monitoring technology is rapidly evolving, with major implications for EPA and state environmental programs.

The team evaluated data standards based on how well they met a set of criteria, the most important of which are market maturity and ease of discovery. In addition, the team evaluated the options to see which ones met four uses cases (i.e., data access, querying, ownership and archiving) that support EPA, states and Tribes in their efforts to use continuous monitoring data.

"Data standards are the rules by which data are described and recorded. In order to share, exchange, and understand data, we must standardize the format as well as the meaning" (USGS, 2017). A data standard enables different entities to be able to communicate information between one-another without having to first process or reformat the data. Standards are important because they reduce the cost for secondary use of data, and allow for the preservation of the data thereby preserving the initial investment in the collection of that data.

Team 4 has been tasked with reviewing existing data standards defining the representation, format, definition, structure, transmission, and management of data. There are many existing data formats in use in both the public and private sector. The Team explored these formats (some of which have been adopted by broader communities as 'standards') and evaluated each of these formats against an agreed upon set of criteria. Based on these criteria, the team has also made a recommendation as to which of these standards would be the most appropriate to target as a common standard across sensors for all media.

Based on this evaluation, the team recommends that agencies use the Open Geospatial Consortium (OGC) SensorML/WaterML 2 standard. It meets the criteria best and it meets all four use cases for data use (defined below in the criteria section). However, it does not currently define all of the data elements necessary to determine data quality. As such, further work would need to be done with the OGC to help strengthen this section of the standard. It is worth noting that this standard may be applicable in some situations but not in others. For example, NetCDF ranks high for most criteria but it is intended for archiving and is therefore not ideal for querying. However, NetCDF could meet the need for archiving data where SensorML/WaterML2 may not efficiently meet this need.

Comparison of Standards to Criteria

Standard	Score (7 max)	Market Maturity	Ease of Discovery	Government Use	Data Quality Documentation	Off-the-Shelf/ Community Support	Meets Use Cases	Multimedia
OGC: SensorML/WaterML 2	6.5	H	Queryable	Broad	M	H	1, 2, 3, 4	Tested
WaterOneFlow/WaterML	6.0	M	Queryable	Broad	M	H	1, 2, 3, 4	Tested
WQX*	5.0	M	Queryable	Broad	H	M	2, 3, 4	Designed
AQcsv	4.5	M	Internal	Broad	M	M	1, 2, 3, 4	Designed
NetCDF (archive standard)	4.5	H	Archive	Broad	M	H	4	Tested
AQS*	4.0	M	Internal	Community	H	M	2, 3, 4	Single-Media
OGC: SensorThings API	3.0	L	Queryable	None	L	L	1, 2, 3 (4?)	Tested
VIPER	2.0	L	Internal	Limited	Unknown	L	1, 2, 4	Tested

*AQS and WQX are both capable of exchanging real-time (continuous) data, however, they are principally designed for sharing sample data which has different metadata requirements and a different structure. Although both of these standards could support the sharing of this data, they would not be able to handle the data in an efficient manner.

Criteria Definitions

Score: A ranking between 0-7 to indicate how well each standard met the criteria. The higher the value, the closer the standard came to meeting all of the criteria.

Market Maturity: The extent to which the standard is used outside of government and a community exists to promote and support the standard.

Ease of Discovery: The ability to discover and query data published using this standard.

Government Use: The extent to which the standard is used by U.S. government agencies.

Data Quality Documentation: The built-in capability of the standard to provide the data elements that would allow a user to determine the quality of the data.

Off-the-Shelf/Community Support: The extent to which third-party developers have developed off-the-shelf software that support or implement the standard.

Use Cases: Does the standard support the following four use cases:

1. Provide real-time access to sensor data. “Tell me the value right now, or at least the last reading.”
2. Query the data by date and parameter, and allow the retrieval of large amounts of data quickly and efficiently.
3. Allow for data ownership. Easily identify who owns the data when data are being provided by multiple different entities, and allow for the retrieval of relevant data quality information across organizations.
4. Allow for long-term archiving of high temporal-density data sets.

Multimedia: Has the standard been designed and/or tested to share both air and water data?

Introduction

The Advanced Monitoring Team is a collaborative effort under the E-Enterprise Initiative to understand how to use, communicate, and ensure quality of advanced monitoring data, especially continuous monitoring data. Environmental monitoring technology is rapidly evolving, with major implications for EPA and state environmental programs. New advanced monitoring technologies that are smaller, more portable, and less expensive than traditional methods offer unprecedented opportunities to enhance environmental protection by enhancing existing monitoring networks and our ability to reduce pollution. The collective community must proactively harness and take advantage of benefits from these sensors or risk losing the opportunity to shape how the new technologies are applied.

The rapid evolution of monitoring technology presents a challenge to government agencies, the public, and the regulated community because of the disparate structure and amount of data produced by continuous sensors versus traditional discrete style of sampling. While previous attempts by various groups (NWQMC 2006) have been made to standardize data generated via discrete sampling, to date, standards for continuous data have not been proposed. Without a standard way of interacting with and storing this data, the possibility of having multiple incompatible approaches for publishing these data is very real. This situation does not lend itself well to interoperability, and poor data management and poor communication, due to lack of standards, can mean that a lot of opportunities for sharing data and using that data for secondary purposes will be missed.

Team 4 has been tasked with reviewing existing data standards defining the representation, format, definition, structure, transmission, and management of data. There are many existing data formats in use in both the public and private sector. The Team explored these formats (some of which have been adopted by broader communities as 'standards') and evaluated each of these formats against an agreed upon set of criteria. Based on these criteria, the team has also made a recommendation as to which of these standards would be the most appropriate to target as a common standard across sensors for all media. Many existing data platforms exist, and this report should not be read as to mean that those systems need to change in order to adopt these recommended standards, but rather should identify ways to incorporate these standards into their data sharing approaches. This report is the first of three reports. This report focuses on the core data elements and standards around those data elements. Subsequent reports will focus on the necessary metadata needed to identify the quality of the data and the needed architecture for establishing an interoperable sensor network.

What are standards?

"Data standards are the rules by which data are described and recorded. In order to share, exchange, and understand data, we must standardize the format as well as the meaning" (USGS, 2017). A data standard enables different entities to be able to communicate information between one-another without having to first process or reformat the data. There are multiple examples of data standards in use throughout the public and private sectors, as well as existing entities that focus exclusively on defining, promoting, and testing these standards. For example, the banking sector has the Open Financial Exchange Standard which defines standard data formats and standard exchange protocols for exchange financial information between financial institutions and financial applications (OFX, 2017). The existence of this standard enables people to do online banking and balance their accounts from the convenience of their favorite banking software. Without the OFX standard, a user would need to go to a bank's website, log in, download their financial transaction information, reformat that information to meet their software's specification, and then finally load that data into their software. With the OFX standard, this is a one-step process. Another example of a standard is the Internal Revenue Service's e-file standard (IRS, 2017). This standard is a government-defined standard that describes the data and the rules around that data that would enable an individual to electronically file their tax returns. The existence of this standard has enabled third-party developers to build applications that can provide

electronic filing as a service for a fee. These third-party vendors also provide added value by building software that walks a user through the creating of their tax forms in a user friendly way.

There are also several organizations (both public and private) that have engaged in efforts to formalize and define standards. IEEE, the Open Geospatial Consortium (OGC), and the National Institute of Standards and Technology are just three such examples. These groups provide the expertise, and often also convene forums of experts to define, test, and promote standards. As part of the research of this Team, a core principle that was followed was to leverage the work of these expert bodies wherever possible.

When considering standards, there are at least four different aspects of standards to consider:

1. **Data Standard.** This defines the data elements and rules used to describe data, as well as the structure and format of that data. Additional considerations can include the ontologies used to describe the data (i.e. common set of terms used to describe the parameters being monitored). The data standard is a critical first step in being able make data interoperable. Defining a data standard must go beyond defining broad concepts, but must also get down to defining specific data elements with the precise names and rules for each of these elements. The Exchange Network has served as an excellent example of this process with data standards having been defined for various data and information flows.
2. **Communication Interface (Application Program Interfaces).** To allow for computers to be able to communicate with one another (which is one of the key reasons why you develop a data standard), you must also agree upon how that communication is going to occur. The computer must know what questions can be asked, and how those questions will be answered. To make information interoperable, any computer participating in that network of shared information must be able to answer the same questions in a common way. This enables third-party developers to write applications against the network, without having to rewrite their application every time a new partner joins the network.
3. **Metadata.** Metadata describes the data that are being communicated via the data standard. Although metadata could be considered a part of the data standard, for the purpose of this report, the team has chosen to separate out portions of the metadata elements from the data standard. Part of the reason for this is in part due to the uniqueness of continuous data and how it is used. For continuous data, a few pieces of metadata can describe a large quantity of data, and as such doesn't need to be repeated for every time/value pair in the data set. The metadata are therefore used for discovering data and determining the fit-of-use of that data, and would likely be done prior to actually retrieving the data themselves. Metadata will be covered in a future report.
4. **Architecture.** The architecture of the system/network defines how the data would flow between partners to enable discovery of the data. For example, would all of the data be submitted to a central repository, and then shared out from there? Or rather, would a central index be maintained of where data are available and then data are retrieved real-time from the source. The architecture of the system depends very much on how the data will be uses, and the nature of the data (i.e. real-time data may not be a good fit for a central repository, if those data are coming from multiple sources). The proposed architecture for the system will also be covered in a future report.

Minimum Data and Metadata Elements

As part of this evaluation, the team identified a core set of data elements that any data standard would need to have. In order to effectively distinguish between metadata and data, the team identified two types of metadata:

Metadata Tier 1: Identifies what kind of data is being collected and site identification information. Set of descriptive features: Station ID, Latitude, Longitude, parameters being measured, frequency of measurements, etc. This metadata is critical to the understanding of the data, and must be transmitted with the actual data (time/value pairs).

Metadata Tier 2: Identifies the quality of data. Calibration, cleaning, maintenance, frequency of data change, and QAQC information. This metadata is valuable when discovering which data a user may be interested in, and screening out those data sets that they are not interested in. Although these data are important to understanding the quality of the data and how the data may be used, they do not necessarily always need to be included with the data. For example, a user could use Tier 2 metadata to discover what data sets they are interested in, and then once they know that, they could then download the data. This metadata was critical in making the decision, but wouldn't necessarily need to be downloaded with the data.

Continuous data vs. Discrete Data

Although many standards exist and are utilized to manage discrete data, these standards are not appropriate to describe data coming from continuous sensors as advanced monitoring data is inherently different than traditional discrete or lab sample type monitoring data. The frequency that monitoring data from these sensors is reported can be very high because there is no lab analysis needed, measurements can be taken very quickly. With telemetry techniques, these advanced monitoring systems are equipped to serve up data in almost real time.

The set of metadata that needs to accompany the high frequency time value pairs needs to be communicated on a much less frequent basis, some information only needing to be communicated once.

For the purposes of this report, the team focused exclusively on core data elements and the Tier 1 metadata and plans to evaluate Tier 2 metadata in a follow-on report. The below table identifies a number of core data element concepts. Element names are defined in each standard that was evaluated, and the actual name could vary depending on the standard. What is important is to know whether or not the standard could capture the concept. Having a common name is certainly important, and would be a critical piece of the recommended standard, but for the purpose of evaluating these standards is not necessary.

Core Data Element Concepts

Element Concept	Description
Value	The actual measured or derived value that is being reported.
Date/time	The date and time of when the measurement was collected. Time values must include time zone offsets, and should follow standard formats.
Feature	Feature being measured. This could be a stream, facility, Airshed, etc. Features can have their own geospatial definitions that does not need to be confined to a simple latitude/longitude point coordinate.
Feature Name	The name of the feature being measured. This simplifies the discovery of the data by others.
Feature Identifier	A unique identifier for the feature. This simplifies the ability to retrieve data for that feature through services.
Feature Location	In most cases, this will be a latitude/longitude coordinate pair, but could be lines or polygons.
Parameter	The actual parameter being measured or derived.
Units	The unit of measure describing the value.
Qualifier Code	A code qualifying the measure (i.e. Raw, Preliminary, Final).

Why Are Standards Needed?

With the proliferation of new technology, data are being collected in various formats with widely varying levels of data quality and documentation, making it difficult to use and evaluate. This circumstance often requires a significant effort in data wrangling to harmonize data from multiple sources into a consistent format for analysis. In an ever increasingly data driven world, this recognition has led to a call for data to adhere to the FAIR (Findable, Accessible, Interoperable, and re-usable) principle. Findable implies that data and metadata are ascribed with globally and eternally persistent unique identifiers. To be accessible means that data and metadata are retrievable by those unique identifiers via commonly used internet communication protocols. Interoperability deals with ontological and semantic consistency and involves adhering to formally defined vocabularies and utilizing terminologies across a domain. Achieving interoperability requires data and metadata standards. This greatly reduces the chance of misunderstanding the meaning of terms used in the larger domain as well as the data and metadata standards. Finally, re-usability advocates for a license to use and reuse the data as well as an assurance that the metadata is robust enough to fully describe the data and provenance thus allowing users to determine the proper use of the data.

Achieving interoperability requires data and metadata standards

The FAIR principle enables any user to discover and correctly use information with much greater certainty regarding provenance and data quality thus providing a greater level of transparency with which to assess the condition of one's air and water quality. Adherence to the FAIR principle enables efficient query, discovery, retrieval, storage, analysis, and display of data and necessarily implies the use of data and metadata standards in achieving that goal. Standards for documentation, format, definition, and structure facilitates management of data and allow numerous and diverse entities to distribute, share, and integrate data.

Given shrinking budgets for agencies, the democratization of environmental monitoring, and the trend for open publication and sharing of collected data, there is need and interest in utilizing multiple sources of data to support decision-making. Incompatible datasets, however, along with missing information can render data unusable or require extensive reformatting to make cross platform datasets viable thus wasting resources expended to generate such data (Sprague et al. 2017). Data ambiguity and the risk or misusing data can be greatly minimized through the use of a common data standard. Some of the benefits include:

1. Leveraging resources from multiple sources and using data beyond its original intent
2. Ability to more easily query and filter information and find the needed data
3. Higher degree of confidence in analysis and decision making
4. Archiving and storage of data without the fear of future incompatibility

Simply put, data cannot be consistently and accurately shared without the adoption and adherence to data and metadata standards.

Results/Findings

The team realizes that a single standard, or even a small set of standards, cannot encompass the breadth of environmental data being gathered. However, by evaluating existing systems, each with unique spatial, temporal, and media features, the team found a set of core commonalities. Those common features will lead to broad standards for information interchange.

This effort requires strong and broad partnerships between data providers and users, including EPA Program offices and their state and Tribal equivalents, ORD, the Office of Environmental Information

(OEI), EPA Regions, citizen scientists, researchers, environmental groups, cloud providers, sensor manufacturers, other federal agencies, and academic institutions**.**

As part of this review, the team reviewed several existing standards, including international, open standards such as the Open Geospatial Consortium SensorML/WaterML2 standard, NetCDF, AQcsv, WaterOneFlow/WaterML, AQS, WQX, VIPER and OGC SensorThingsAPI (see Appendix A for a description of all of the standards that were reviewed). The team also identified criteria by which these standards were evaluated (see Criteria section below).

Recommended Approach

The team evaluated data standards based on how well they met a set of criteria, the most important of which are market maturity and ease of discovery. In addition, the team evaluated the options to see which ones met four uses cases (i.e., data access, querying, ownership and archiving) that support EPA, states and Tribes in their efforts to use continuous monitoring data.

Based on this evaluation, the team recommends that agencies use the Open Geospatial Consortium (OGC) SensorML/WaterML 2 standard. It meets the criteria best and it meets all four use cases for data use (defined below in the criteria section). However, it does not currently define all of the data elements necessary to determine data quality. As such, further work would need to be done with the OGC to help strengthen this section of the standard. It is worth noting that this standard may be applicable in some situations but not in others. For example, NetCDF ranks high for most criteria but it is intended for archiving and is therefore not ideal for querying. However, NetCDF could meet the need for archiving data where SensorML/WaterML2 may not efficiently meet this need.

The biggest impact of adopting this standard would be on existing systems that are already publishing continuous data that do not conform to this standard. These systems already provide robust data access with applications built off those existing access methods. The team recommends that the agency adopt these standards in a step-wise approach with newer systems/data flows adopting these standards as they are developed, and existing systems building out new services that meet these standards as funding allows. Agencies should also explore existing off-the-shelf software that can support these standards as a means of enabling the publishing of these data using the standards.

To implement these standards, the team recommends that the metadata gap with the SensorML/WaterML 2 be addressed. The team will explore options for this as part of a follow-on report to this report. With a recommended metadata model, the team would be ready to begin piloting efforts to share/publish data in these formats. One pilot (the Interoperable Watersheds Network) has already been completed, and lessons learned from that effort should inform follow-on activities, including looking at cross-media data publishing and working with Team 3 to develop ways to communicate the data.

Limitations of Recommendation

The team evaluated multiple approaches for sharing continuous data. No one approach is inherently wrong, and all of the approaches meet the specific needs for which they were designed. This report should not be read as an evaluation of the overall value of any of the standards evaluated here. Rather what this report seeks to do is to provide a systematic way of recommending an approach for EPA, state, and tribal agencies to have their data be interoperable. Further work needs to be done in this sphere, including the development of a metadata model for describing data quality and developing an architecture for how an interoperable system would be developed. The overall approach should consider all three of these components as a whole, and the success of this effort is dependent upon all three components.

Criteria Used for Evaluation

To provide a framework by which to evaluate existing standards and approaches, the team developed a suite of evaluation criteria to help distinguish between the various pros and cons of each standard. For the purposes of the evaluation, it was determined that all of these standards that are part of this review met the basic requirement of a data standard in that they each define a suite of data elements and rules by which information can be shared. Table 1 below shows the rankings for each of the standards that were evaluated based on the below criteria.

Market Maturity: The extent to which the standard is used outside of government and a community exists to promote and support the standard.

Ease of Discovery: The ability to discover and query data published using this standard.

Government Use: The extent to which the standard is used by U.S. government agencies.

Data Quality Documentation: The built-in capability of the standard to provide the data elements that would allow a user to determine the quality of the data.

Off-the-Shelf/Community Support: The extent to which third-party developers have developed off-the-shelf software that support or implement the standard.

Use Cases: Does the standard support the following four use cases:

5. Provide real-time access to sensor data. "Tell me the value right now, or at least the last reading."
6. Query the data by date and parameter, and allow the retrieval of large amounts of data quickly and efficiently.
7. Allow for data ownership. Easily identify who owns the data when data are being provided by multiple different entities, and allow for the retrieval of relevant data quality information across organizations.
8. Allow for long-term archiving of high temporal-density data sets.

Multimedia: Has the standard been designed and/or tested to share both air and water data.

Table 1. Comparison of Standards to Criteria

Standard	Score (7 max)	Market Maturity	Ease of Discovery	Government Use	Data Quality Documentation	Off-the-Shelf/ Community Support	Meets Use Cases	Multimedia
OGC: SensorML/WaterML 2	6.5	H	Queryable	Broad	M	H	1, 2, 3, 4	Tested
WaterOneFlow/WaterML	6.0	M	Queryable	Broad	M	H	1, 2, 3, 4	Tested
WQX*	5.0	M	Queryable	Broad	H	M	2, 3, 4	Designed
AQcsv	4.5	M	Internal	Broad	M	M	1, 2, 3, 4	Designed
NetCDF (archive standard)	4.5	H	Archive	Broad	M	H	4	Tested
AQS*	4.0	M	Internal	Community	H	M	2, 3, 4	Single-Media
OGC: SensorThings API	3.0	L	Queryable	None	L	L	1, 2, 3 (4?)	Tested
VIPER	2.0	L	Internal	Limited	Unknown	L	1, 2, 4	Tested

*AQS and WQX are both capable of exchanging real-time (continuous) data, however, they are principally designed for sharing sample data which has different metadata requirements and a different structure. Although both of these standards could support the sharing of this data, they would not be able to handle the data in an efficient manner.

Appendix A: Researched System Descriptions

As part of the evaluation of the standards, the team looked at existing systems and implementations. This helped inform the evolution. Below is a description of each of the standards and implementations of those standards that were evaluated.

Standards Evaluated

OGC: SensorML/ WaterML 2

The Open Geospatial Consortium (OGC, 2017) is an international standards setting body that works to define standards and protocols for communicating geospatial data. Their largest areas of focus have been on mapping data. Some other example standards from OGC include Web Mapping Service, Web Feature Service, and Keyhole Markup Language (KML). SensorML, WaterML and SensorThings API are part of OGC's Sensor Web Enablement suite (SWE), designed to support the Internet of Things (IoT). SensorML describes the sensor and metadata around the sensor. WaterML 2 describes the feature being measured and contains the actual measurements. SensorThings API provides a framework to interconnect IoT devices, data, and applications over the Web.

OGC: SensorThingsAPI

SensorThings API is an OGC standard that provides ways to communicate sensor locations, sensor and data parameters and sensor instruction sets. The standard is open and it applies an easy-to-use REST-like style. It is relatively new so few organizations, such as the University of Calgary, have implemented it.

NetCDF

The Network Common Data Format (NetCDF) (UCAR, 2017) was developed by the University Corporation for Atmospheric Research. It is used extensively by NOAA, USGS, and NASA for archiving datasets for later discovery and for sharing data among researchers. Data contained in NetCDF format are self-describing and can be accessed using a number of open-source platforms. It is often used to publish/share gridded data sets (i.e. raster data), but can also share large collections of sensor data. It is not designed for querying pieces or subsets of a dataset, but rather retrieving an entire data set and then determining which parts of the data you want to use.

WQX

The Water Quality Exchange (WQX) (EPA, 2017) was developed by EPA to enable the sharing of water quality sampling data between other federal agencies, states, tribes, and watershed groups. It provides a complete metadata profile for describing data quality and the methods used to collect and analyze the data. It is occasionally used to share sensor data, but because the data model is designed for sample data, it is not an efficient model for sharing sensor data. EPA's current recommendations for sensor data is that partners summarize the data into daily averages, and then share the averages instead of the incremental measurements.

For water quality, WQX (Water Quality Exchange) is a data standard for submitting water data into EPA's STORET (Storage and Retrieval) system, primarily generated via discrete sampling. Prior to submitting data, an entity needs to follow the data standard for submission. Meaning all required fields, parameter and unit naming conventions, must be part of the data submission to successfully upload data into STORET.

For this upload to happen, a series of complicated technology-based procedures need to occur. Data submissions to STORET can occur through state nodes, node clients, or a lighter based tool called WQX web. What all three have in common is file transfers are handled through XML code. Text files are converted to XML code so that the individual parts of the file containing information can successfully populate an underlying database. The communication standard is all about the flow of the data and the mechanisms in place to make that happen.

WaterOneFlow/WaterML 2

The Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) developed a family of the Web services called WaterOneFlow to enable the sharing of water data between multiple entities (CUAHSI, 2017). These standards predate the OGC standards and served as starting point for the development of the current WaterML 2 OGC standard. These standards are used broadly by the academic community, and there is a large open-source community that is developing tools and applications to work with data in these formats.

VIPER

VIPER (EPA Emergency Response Team, 2017) is a custom protocol used by EPA's Emergency Response Team to enable the management of real-time data collected during emergency response. The VIPER protocol enables data to be transmitted from the sensors in the field back to a SQL Server database where the data can then be published, shared, and visualized.

AQcsv

The AQCSV data format was developed to support the EPA AirNow real time data system. The format is a simple text record with comma delimiters between each field. Each record contains one data value for a given site, time, and parameter. Other supporting fields in the record describe the time zone, units, location, and other information for that data value.

AQCSV was developed with the AQS XML data format as a model, so that the two systems could maintain a high level of compatibility. AQCSV supports AQS standards for parameter codes and pollutant occurrence codes (POC). AQCSV can accept any AQS parameter from multiple sources such as speciated lab data, mobile monitor data, and even sub-hourly data.

By adopting the AQCSV format, AirNow supports the backfill of AQS approved data into AirNow when available. AQCSV was also designed to support international data submission, by providing a simple, well-supported input format. The format has been implemented by all major environmental monitoring equipment and software companies, making it simple for air quality professionals to send data to AirNow.

AQS

The Air Quality System (AQS) houses the official, regulatory ambient air quality data for the US Environmental Protection Agency (EPA). Air quality measurements from thousands of monitoring stations around the country, spanning many decades, can be accessed there. AQS also serves as the ongoing repository for ambient monitoring networks.

Data reaches AQS from monitoring networks operated by state, local, and tribal agencies. Those agencies use the AQS XML schema as the primary data flow for data submission to EPA. Other submitters include Tribal consortia, other Federal agencies, analytical laboratories, and contractors.

AQS's primary purpose is to support EPA's regulatory mission, by hosting ambient air quality monitoring data that determines compliance with the Clean Air Act and amendments. A secondary purpose is to serve

as the air quality data repository for the research community, within EPA, academia, and the health effects research community.

AQS accepts approximately 90 million data measurements per year, facilitates the quality assurance of these values, calculates summaries at various time scales (sub-daily, daily, quarterly, and annual), and serves out about 50,000 reports per year. It is an N-tiered Oracle application with approximately 70 forms and 35 reports with 700 users.

Example Implementations of the Standards

Water Quality Portal (Uses WQX Standard)

This implementation of WQX enables the sharing of over 300 million water quality sampling results from over 2.5 million locations. It is developed and managed by USGS, EPA and the National Water Quality Monitoring Council. Standard services and the WQX data model enable this data sharing between over 400 different data partners. The Water Quality Portal is available at: https://waterqualitydata.us.

Interoperable Watershed Network (Uses OGC standards: SensorML, WaterML2)

The IWN was a demonstration/test project under E-Enterprise that tested the OGC standards to determine if they would perform sufficiently for sharing sensor data. The pilot tested the sharing of data from 8 different partners from over 15,000 sensors. The standards performed well. The IWN also leveraged off-the-shelf software to enable the data sharing network**.** The demonstration tool for IWN is available at: http://54.210.62.171/.

CUAHSI (Uses WaterOneFlow/WaterML)

The CUAHSI network has a similar design to the IWN project except that they use their own custom-developed standards and services. The CUAHSI catalog ingests data from over 90 partners throughout the world, with a large focus on academic partners. CUAHSI has developed a suite of open-source tools that support the publishing of WaterOneFlow services through their Observations Data Model (ODM) system which can be deployed by any partner to activate a node on their network and serve as a Hydrologic Information Service (HIS). CUAHSI has also developed open-source tools for consuming data from the network as well. The CUAHSI community is a broad and vibrant community. CUAHSI is beginning to evaluate updating their network to include the OGC standards.

USGS National Water Information System (Uses WaterML, WaterML2, and Custom Services)

HYDROML is an XML format developed for the transport and archival of hydologic data. HYDROML was originally developed by the National Water Information System (NWIS) office of the United States Geological Survey for the purposes of importing, exporting, and archiving hydrologic data from and to the NWIS data base. AquariusHYDROML is a modification of the original work to allow it to be used with the Aquarius software. AquariusHYDROML is a variation of HYDROML and not an extension. The types of hydrologic data include site information, computation instructions, corrections, ratings, shifts, time-series data including unit values and daily value statistics, peak flows, and site visit measurements. There are many different uses possible for HYDROML and not all data elements are used in every case. In fact many data elements will not be used depending upon the data content desired or the usage of the XML file. There are many different usage examples that are possible but the combinations are too numerous to list here.

MARACOOS (Mid-Atlantic Regional Association Coastal Ocean Observing System) (Uses netCDF, WaterML)

MARACOOS follows a National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Information (NCEI) 2.0 templates and CF1.6 compliant standards. Any data partners submitting data through MARACOOS must meet these standards. NOAA has developed netCDF templates to encourage long-term preservation, scientific quality control, and multiple data re-use beyond its original intent. NOAA does note that these templates do not represent an attempt to create a new standard, and they are not absolutely required for archiving data at NCEI. However, they hope that users see the benefits in structuring data following these conventions.

In addition, there is a ISO 19115-2 metadata standard that is followed that focuses more intently on the geographic and spatial content of the data. Especially imagery and gridded based data. This is a standard by International Organization for Standardization (ISO).

To test data quality, these standards are compared against the 'Quality Assurance of Real-Time Oceanographic Data' (QARTOD) methodology. QARTOD is part of the Integrated Ocean Observing System (IOOS) which is affiliated with NOAA. QARTOD has more of a focus on Real-Time Quality Control of Data and High Frequency (HF) Radar Surface Current Data. Users submitting their data must at least follow the QARTOD standard or have something more rigorous in place.

AirNow

The AirNow realtime data system receives data from nearly 4000 regulatory monitors in the US, as well as data from Canada and Mexico. More recently, the AirNow system has been extended to accept data from international partners in SE Asia, beginning with Taiwan. Partners in SE Asia hope to use the AirNow infrastructure to establish regional data sharing. Further, AirNow was selected by the Department of State to host data from its growing Embassy Monitoring Program, in which regulatory-grade monitors are deployed at embassies and consulates worldwide.

A critical part of the AirNow infrastructure, the AQCSV data format is a simple text record with comma delimiters between each field. Each record contains one data value for a given site, time, and parameter. Other supporting fields in the record describe the time zone, units, location, and other information for that data value.

AQCSV is a proprietary, but open and well-supported, format. The AirNow team furnished the AQCSV specifications to all monitoring equipment companies and most, if not all, have built-in AQCSV modules that can be used to flow data into AirNow.

NOAA IOOS

The Integrated Ocean Observing System's (IOOS) goal is to maximize access to data and produce information that promotes efficient and accurate decision making. IOOS supports National Oceanic Atmospheric Administration (NOAA) directly. IOOS is a national partnership that requires a series of standards, formats, and protocols depending on the type of data being submitted by a partner. IOOS data providers are expected to supply metadata using one of the below standards that feeds the 'IOOS Catalog' for Public search and discovery. In general, partners should follow the guidelines in the NOAA Data Documentation Procedural Directive when determining appropriate meta data standards for data submission.

Appendix B: Methodology

The team developed four use cases that described the primary activities that states, tribes, and EPA could use the standards for, namely:

1. Provide real-time access to sensor data. “Tell me the value right now, or at least the last reading.”
2. Query the data by date and parameter, and allow the retrieval of large amounts of data quickly and efficiently.
3. Allow for data ownership. Easily identify who owns the data when data are being provided by multiple different entities, and allow for the retrieval of relevant data quality information across organizations.
4. Allow for long-term archiving of high temporal-density data sets.

Then the team contacted environmental agencies and other organizations to learn what standards they used. The team held conference calls with the standards bodies to learn about the standards and their adoption.

The team then evaluated each standard based on a defined set of seven criteria. Each standard was provided a score based on the evaluation criteria below. An overall score (between 0-7) was also derived for each standard. The score was derived by assigning a numeric score between 0-1 for each criteria. A standard received a score of ‘1’ if it received a rating of ‘High’ or other top-tier rating for that particular criteria. A mid-level rating resulted in a score of ‘0.5’ and a low-level rating resulted in a score of ‘0’. These individual scores were then summed to derive an overall score. Below is a description of the criteria and their corresponding rating definitions.

Market Maturity: The extent to which the standard is used outside of government and a community exists to promote and support the standard.

> **High (H):** The standard is already being implemented/used by broad sections of the market outside of the government with a corresponding standard setting body having provided accreditation for the standard.
> **Medium (M):** The standard is used within a specific community, but has not had broader adoption outside of that community or been accredited by a standard setting body.
> **Low (L):** The standard is being implemented/used in limited instances, often specific to one or a few entities.

Ease of Discovery: The ability to discover and query data published using this standard.

> **Queryable:** The standard is designed to accommodate the querying of parts of a data set (i.e. data within a date range) or select parameters or sites. The standard is designed for data sharing for both inbound and outbound communication. Data shared using this standard follow the FAIR principle in that they are ‘Findable’, ‘Accessible’, ‘Interoperable’, and ‘Reusable.’
> **Internal:** The standard does not provide a protocol for external querying, but is rather a standard/protocol for the communication of data from a sensor to a sensor owner or for submitting data to a central system (the external communication is handled by other means and not by the standard).
> **Archive:** The standard provides the ability to archive large data sets efficiently, and allows for those data sets to be retrieved efficiently in their entirety, but does not allow the querying of specific pieces of the data.

Government Use: The extent to which the standard is used by U.S. government agencies.

> **Broad:** The standard is used by three or more federal agencies for publishing data.
> **Community:** The standard is used within a specific community (i.e. between states and EPA), but has not had broader adoption beyond that community by other federal agencies.
> **Limited:** The standard is used by only one government entity for specific purposes.
> **None:** The standard is not used in the government sector.

Data Quality Documentation: The built-in capability of the standard to provide the data elements that would allow a user to determine the quality of the data.

> **High (H):** The standard captures full metadata on data quality, including: methods used for collection and analysis, equipment used, precision and accuracy of the measurements, data quality flags, and other relevant QA/QC information. The standard contains all the necessary metadata elements as-is without the need for modification.
> **Medium (M):** The standard captures some metadata about data quality which may include, at a minimum, data quality flags, but may not have a standard approach for capturing other metadata such as methods and equipment. The standard has the potential to be extended to define these data elements, but more work would need be done in this area.
> **Low (L):** The standard captures limited metadata, and in some cases may only provide the final values. Data quality is ensured via other mechanisms, and is assumed within the data standard, but is not directly discoverable in the standard.

Off-the-Shelf/Community Support: The extent to which third-party developers have developed off-the-shelf software that support or implement the standard.

> **High (H):** Third-party vendors or communities have emerged to develop, market, and support tools that enable this standard with only limited government involvement. A high rating indicates that there are several off-the-shelf products available that could be used to implement this standard either for a cost or as an open-source option.
> **Medium (M):** Some third-party applications exist, but are serving exclusively a government market. Without a government market driving the development of these tools it is uncertain if the market for these tools would continue to exist.
> **Low (L):** Only custom-developed tools exist that were either developed by the private sector or government to implement that standard for one implementation. A re-use or support community has not yet emerged.

Use Cases: Does the standard support the following four use cases:

1. Provide real-time access to sensor data. “Tell me the value right now, or at least the last reading.”
2. Query the data by date and parameter, and allow the retrieval of large amounts of data quickly and efficiently.
3. Allow for data ownership. Easily identify who owns the data when data are being provided by multiple different entities, and allow for the retrieval of relevant data quality information across organizations.
4. Allow for long-term archiving of high temporal-density data sets.

Multimedia: Has the standard been designed and/or tested to share both air and water data?

Tested: The standard has been designed and tested for the sharing of real-time data for both air and water.
Designed: The standard has been designed, but not yet tested for the sharing of real-time data for both air and water.
Single-Media: The standard is designed specifically for sharing data for either air or water, but not both.

Appendix C: Related Projects

Additionally, the team evaluated other approaches that are not yet standards based, but could inform future use-cases for interoperable sensor data. These projects were not evaluated as part of this report, but may provide insight for the Architecture report that will be completed later.

CoCoRaHS (Community Collaborative Rain, Hail and Snow Network)

CoCoRaHS is a grassroots volunteer network of backyard weather observers who measure and map precipitation (rain, hail and snow) in their local communities. The system receives approximately 11,000 observations per day and observers have collected approximately 38 million records over the past 10 years.

CoCoRaHS data is transmitted from a website or a mobile app using http and https protocols. They use data standards that are very similar to the National Weather Service (NWS) Cooperative Observer Program (COOP program). These differences are known by the NWS personnel who use the data, but it is not an issue because COCORaHS can map all of their data to NWS specifications.

For COCORaHS, the important thing is to be able to map their data into the formats that their data users need. For instance, the WaterML standard provides additional schema definitions for hydrology data with XML being the data format. However, the COCORaHAS data users have not asked for data in this standard because it would be harder for them to ingest on their end than a simple CSV or XML file. If COCORaHS had users who requested WaterML then they would support it.

JSON will be the primary data format for CoCoRaHS in the near future and they will continue to support the existing export utilities. They have already started supporting GeoJSON for one data feed based on need.

Regarding the combination of data, the larger data aggregators like the Applied Climate Information System, PRISM Climate Group and Global Historical Climatology Network have not provided standards per se, but their data formats become de facto standards. Once data users have a library to access data from an aggregator they can access a range of data sets. Unfortunately for the aggregators, the data sources they collect are usually unique so they have to customize each data ingest process.

Waggle/Plenar.io

Waggle is an open sensor platform developed at Argonne National Laboratory (ANL) that forms the basis of their Array of Things project. Array of Things is a smart city effort that seeks to deploy and integrate a variety of data sources and sensors to enable smart city functionality and decision making. The concept is to utilize wide-area sensor deployment to feed data to a cloud infrastructure to enable analytics, visualization, and finally automated decision making. An example might be to utilize data from air quality sensor network along with meteorological data to automatically optimize traffic signaling to reduce congestion which in turn should reduce air pollution exposures. While many embedded sensor systems are built with proprietary OS and software stack, Waggle looks to an open solution that incorporates modularity, extensibility, and security. Their immediate plan over the next few years is the deployment of 500 sensors in the Chicago area. To complement waggle, their open hardware platform, the Array of Things is utilizing an open “back end” platform called plenar.io (plenario). Plenario aims to improve the usually large effort in data wrangling needed when dealing with data from disparate sources. It also uses a spatial and temporal index making it easier to query across data sets along spatio-temporal parameters. Plenario plans to accomplish this by employing an automated ETL (extract,

transform, and load) builder. Plenario's open code is maintained on GitHub. Plenario is funded by the National Science Foundation.

References

CUAHSI. (2017). *Hydrologic Information Service: Standards*. Retrieved March 14, 2017, from Consortium of Universities for the Advancment of Hydrologic Science, Inc. (CUAHSI): http://www.opengeospatial.org/standards/sos

EPA. (2017). *Water Quality Exchange*. Retrieved March 14, 2017, from https://www.epa.gov/waterdata/storage-and-retrieval-and-water-quality-exchange

EPA Emergency Response Team, E.-E. (2017). *Emergency Response Team: VIPER*. Retrieved March 14, 2017, from EPA On-scene Coordinators: https://www.epa.gov/waterdata/storage-and-retrieval-and-water-quality-exchange

IRS. (2017). *IRS e-file Rules and Requirements*. Retrieved from IRS e-file Rules and Requirements: https://www.irs.gov/uac/irs-e-file-rules-and-requirements

OFX. (2017). *Open Financial Exchange*. Retrieved from Open Financial Exchange: http://www.ofx.net/

OGC. (2017). *Open Geospatial Consortium*. Retrieved March 14, 2017, from OGC: Sensor Observation Service: http://www.opengeospatial.org/standards/sos

UCAR. (2017). *University Corporation for Atmospheric Research*. Retrieved March 14, 2017, from Unidata: NetCDF: https://www.unidata.ucar.edu/software/netcdf/docs/faq.html#whatisit

USGS. (2017). *Data Standards*. Retrieved from https://www2.usgs.gov/datamanagement/plan/datastandards.php