Data Exchange Specifications

Becca Scully

3/1/2021

## R Markdown

# Stream Monitoring Habitat Data Exchange Specifications

Data exchange specifications are a set of guidelines and rules for using and combining information. Rigorous data exchange specifications support reuse, promote interoperability, and reduce data integration costs (Morris and Frechette 2008, Hamm 2019).

# Methods

To integrate data, we need to establish business rules to flow data from the sources to a target data sets. Data must be accessed, extricated, moved, validated, cleaned, standardized, transformed, and loaded (DAMA Dictionary). At the start of this project there is no target data set or rules to integrate stream habitat data. Therefore, to build an integrated data set we define the structure and fields to include in the target data set and define business rules to flow data from sources.

Before starting the integration work, we need to understand what data and metadata from the desperate data sets needs to be integrated. There are two approaches: 1. Look for a specific regulatory management question that needs to be answered in order to inform a decision and provide the data to answer those specific questions. (CITE EXAMPLE CAX) 2. Let the data guide the integration, identify a subset of like data that provide value and could be use together to answer a variety of management or research question

For stream habitat monitoring data we took the second approach and used the data, metadata and program experts to guide the structure and fields of the target data set. Fields refer to the column headers in the data set.

We invited four large scale, long term, well documented stream habitat monitoring programs to collaborate in defining a method for integrating data from there stream habitat monitoring programs that can be adapted to similar Federal, State and Tribal programs as needed.

#### We coordinated with:

* The Assessment, Inventory, and Monitoring (AIM) Strategy provides a framework for the Bureau of Land Management (BLM) to inventory and quantitatively assess the condition and trend of natural resources on the nation’s public lands. This program gathers information to determine ecosystem conditions and how they are changing over time. Such information is actively used by the BLM to guide and justify land uses, policy actions, and adaptive management decisions. (AIM Strategy website) LINK TO THE SCIENCEBASE ITEM
* The USFS Aquatic and Riparian Effectiveness Monitoring Program (AREMP) was developed to track changes that occurred as a result of active and passive management on the landscape. The Aquatic and Riparian Effectiveness Monitoring Program focuses on assessing the degree to which federal land management under the aquatic conservation strategy (ACS) of the Northwest Forest Plan (NWFP) has been effective in maintaining and improving watershed conditions. (USFS AREMP website) LINK TO THE SCIENCEBASE ITEM
* The USFS PacFish/InFish Biological Opinion Monitoring Program (PIBO MP) is to monitor stream and riparian habitats within the PIBO MP study area, in order to determine if the PacFish (Pacific Anadromous Fish) and InFish (Inland Fish) aquatic conservation strategies can effectively maintain or restore the structure and function of riparian and aquatic systems. (USFS PIBO website) LINK TO THE SCIENCEBASE ITEM
* EPA The National Aquatic Resource Surveys (NARS) are collaborative programs between EPA, states, and tribes designed to assess the quality of the nation’s coastal waters, lakes and reservoirs, rivers and streams, and wetlands using a statistical survey design. The NARS provide critical, groundbreaking, and nationally consistent data on the nation’s waters. (EPA website) LINK TO THE SCIENCEBASE ITEM

These four monitoring programs have structured implementation, core indicators collected with a consistent methodology, statistically valid designs, and effective data management. In addition, based on past work we know there is a subset of consistent metrics and indicators produced across these programs, creating an opportunity to combine data from multiple programs (CITE ).

# Data Model

The general structure for this type of monitoring is a dataset contains multiple locations, some locations are sampled multiple times, those sampling efforts we refer to as events, at each event, programs collect multiple measurements, producing numerous metrics (Figure 1)[]. This type of information lends itself to a relational database model so information is not repeated, this will optimize data retrieval for integrated analysis (CITE). Therefore a location will be stored once and linked to multiple events.

We did not want to build a new data model from scratch so in the future an integrated dataset can be link into exiting data models, repository and resources. We started with the Observation Data Model 2 (ODM2). ODM2 takes features from the Open Geospatial Consortium (OGC) Observations & Measurements (O&M) standard (Cox, 2007a; Cox 2007b; Cox 2011a; Cox 2011b) and the Horsburgh et al, 2008 observation Data Model (ODM) desciring a data model for hydrologic data and the CUAHSI Hydrologic Information System. ODM2 is a generic model for observations and designed for interoperability among disciplines (<http://www.odm2.org/>). ODM2 has a “core” and multiple “extension” and established controlled vocabularies. An observation is defined as an act associated with a descript time or instant, through which a number, term or other symbol is ascribed to an event (Horsburgh et al. 2008). Observations require contextual information, location where the observation was made, date and time, the type of variable and other metadata method used for observation. This OBM2 (<https://github.com/ODM2>) is highly flexible and provides a good starting point for integrating these data sets. Based on the needs for bringing data sets together and these examples we collapsed ODM2 tables, limited the fields to one important for stream habitat monitoring data and added a few fields from the Darwin Core principles as outlined by Wieczorek et al. in 2012. The Darwin Core standard is maintained at the GitHub repository <https://github.com/tdwg/dwc>.

#### Definitions:

* Primary Key
* Feature Key
* Controlled Vocabulary

# Structure

To streamline the data storing and data retaliate we selected a relational data model based on ODM2, we simplified some relationships, removed fields that are not approperate for this datatype and added a few fields from the DarwinCore. See our version of the data schema here: ADD LINK

#### The original data schemas are documented in the ODM2 Git Repository here:

* Core schema: <http://odm2.github.io/ODM2/schemas/ODM2_Current/diagrams/ODM2Core.html>
* Results Extension: <http://odm2.github.io/ODM2/schemas/ODM2_Current/diagrams/ODM2Results.html>
* Sampling Feature Extension: <http://odm2.github.io/ODM2/schemas/ODM2_Current/diagrams/ODM2SamplingFeatures.html>

## ODM Core.Datasets Record Level Class

The Record Level Class documents the core elements of a data set, including information about the origin of the data set, who collected the data, and how to cite the data set. See details in the [Record Level table](Tables/RecordLevel_table.csv). A data set is a collection of locations, at each location a collection events, at each event a collection of metrics; for example, a program releases a data set every five years containing all the data collection locations, events and metrics occurring in the previous five years. We recommend storing metadata about the data sets in a trusted online data repository ensuring we have sufficient information about data sets’ origins. If a program does not have the resources to build a repository, we recommend using USGS ScienceBase, which is available to all. Find more information about ScienceBase here <https://www.sciencebase.gov/about/>.

#### To the Datasets table from ODME2 Metadata table we add:

* MetadataLink

#### From DarwinCore we added to the Record Level table:

* DatasetOrginization
* Modified

#### The primary key is:

* DatasetID

Datasets Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ODMTable | Term | ODM | Description | Examples | DataType |
| Datasets | datasetID | datasetID | An identifier for the set of data. May be a global unique identifier or an identifier specific to a collection or institution. | The ScienceBase ID number of the dataset | Integer |
| Datasets | type | DataseetType | The nature or genre of the resource. | This will be “Water Quality and Physical Habitat Structure” for all values in this data set. | String |
| Datasets | modified | modified | The most recent date-time on which the datasets were combind. | Recommended best practice is to use a date that conforms to ISO 8601:2004(E). | Date |
| Datasets | bibilographicCititation | Citation | A bibliographic reference for the resource as a statement indicating how this record should be cited (attributed) when used. | NA | String |
| Datasets | datasetName | SourcesDescription | The name identifying the data set from which the record was derived. | Example: EPA NARS 2018 dataset | String |

## Sampling Feature

Understanding where data are collected is critical to interpreting biological monitoring data. The Sampling Feature class describes where information are collected, see the list of terms in the [Sampling Feature table](Tables/Location_table.csv). In this data type often this is refered to as the sampling location, but to be compadable with the ODM2 we now refer to this as the Sampling Feature. Each data set contains mutiple sample features (locations). The actionID is the key to link locations to events. To view and analysis data from various sources, latitudes and longitude information must be consistent among data sets; therefore, for this data all latitude and longitudes are converted to WGS1984.

#### To the Sample Feature Table from ODME2 SampleFeature.Site we added:

* Latitude
* Longitude
* SpatialReferenceID

For the integrated data sets all locations Latitude, Longitude are transformed into WGS1984 in Decimal Degrees but in the original data sets the locations are in a variety of spatial references based on each program’s needs, therefor we added the verbatim fields to track the data before transformation.

#### From DarwinCore we added to the Sample Feature table:

* VerbatimLatitude
* VerbatimLongitude
* VerbatimSpatialReferenceID

#### The primary key is:

* SamplingFeatureID

#### Added a foreign key:

* datasetID to link the Dataset table to the Sample Features table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ODMTable | Term | ODM | Description | Examples | DataType |
| SamplingFeature | locationID | SamplingFeatureID | This is the location identification for the integrated data set the value is the concatenation of the verbatimlocationID and the institutionCode. | Example) 5483AIM, 88963AREMP, WtR563EPA | String |
| SamplingFeature | verbatimlocationID | SiteCode | A code or number that identifies a unique sampling location. A site is a stream segment with a fixed starting and ending location for sampling. Code will remain the same across all site visits. | Recommended best practice is to use the EPSG code of the SRS, if known. Otherwise use a controlled vocabulary for the name or code of the geodetic datum, if known. Otherwise use a controlled vocabulary for the name or code of the ellipsoid, if known. If none of these is known, use the value unknown | String |
| SamplingFeature | verbatimLatitude | verbatimLatitude | The verbatim original latitude of the Location. The coordinate ellipsoid, geodeticDatum, or full Spatial Reference System (SRS) for these coordinates should be stored in verbatimSRS and the coordinate system should be stored in verbatimCoordinateSystem. | For this dataset we use the botton of the reach or the middle of the reach depending on the program. | Numeric |
| SamplingFeature | verbatimLongitude | verbatimLongitude | The verbatim original longitude of the Location. The coordinate ellipsoid, geodeticDatum, or full Spatial Reference System (SRS) for these coordinates should be stored in verbatimSRS and the coordinate system should be stored in verbatimCoordinateSystem. | For this dataset we use the botton of the reach or the middle of the reach depending on the program. | Numeric |
| SamplingFeature | verbatimWaterbody | samplingFeatureName | The water body name from the original data set. | For this data set this field is often refered to as Stream Name. | String |
| SamplingFeature | verbatimSRS | verbatimSpatialReferenceID | The ellipsoid, geodetic datum, or spatial reference system (SRS) upon which coordinates given in verbatimLatitude and verbatimLongitude, or verbatimCoordinates are based. | NA | String |
| SamplingFeature | StateProvince | State | The name of the next smaller administrative region than country (state, province, canton, department, region, etc.) in which the Location occurs. | NA | String |
| SamplingFeature | decimalLatitude | Latitude | The geographic latitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are north of the Equator, negative values are south of it. Legal values lie between -90 and 90, inclusive. | NA | Numeric |
| SamplingFeature | decimalLongitude | Longitude | The geographic longitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are east of the Greenwich Meridian, negative values are west of it. Legal values lie between -180 and 180, inclusive. | NA | Numeric |
| SamplingFeature | geodeticDatum | SpatialReferenceID | The ellipsoid, geodetic datum, or spatial reference system (SRS) upon which the geographic coordinates given in decimalLatitude and decimalLongitude as based. | NA | String |

## Core Action

The Action describes an action that occurs at a specific time frame see the [Core Action table](Tables/Event_table.csv) for the terms. In this data type often this is refered to as the sampling event, but to make our data comparable with the ODM2 we adopoted the term action. To assess the status and trend of a resource as a response to management actions, stream habitat monitoring programs often implement a rotating panel design, meaning that the project returns to a single location multiple times during the study duration. Therefore, a data set will contain numerous locations, and each location can include numerous events.

### This implementation

#### To the Core Action table from ODME2 Core Feature Action we added:

* SampleFeatureID

#### From DarwinCore we added to the Record Level table:

* Added VerbatimActionID based on the DarwinCore, because when implementing these data exchange specification we found that there are duplicate ActionIDs between two or more of the source data sets. #### The primary key is:
* ActionID

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ODMTable | Term | ODM | Description | Examples | DataType |
| Action | verbatimEventID | verbatimActionID | Unique number that identifies one sample of a particular site. | In habitat datasets this is often referred to as the ReachID. This ID identifying a unique visit to a location, while the SiteID corresponds to the LocationID. We use eventID to corresponds to Darwin Core standard. | String |
| Action | eventID | ActionID | An identifier for the set of information associated with an Event (something that occurs at a place and time). May be a global unique identifier or an identifier specific to the data set. | NA | String |
| Action | samplingProtocol | samplingProtocol | The name of, reference to, or description of the method or protocol used during an Event. | UV light trap, “mist net”, “bottom trawl”, “ad hoc observation”, “point count”, “Penguins from space: faecal stains reveal the location of emperor penguin colonies, <http://dx.doi.org/10.1111/j.1466-8238.2009.00467.x>”, “Takats et al. 2001. Guidelines for Nocturnal Owl Monitoring in North America. Beaverhill Bird Observatory and Bird Studies Canada, Edmonton, Alberta. 32 pp.”, “<http://www.bsc-eoc.org/download/Owl.pdf>” | Integer |
| Action | EventDate | BegainDataTime | The date-time or interval during which an Event occurred. For occurrences, this is the date-time when the event was recorded. Not suitable for a time in a geological context. Recommended best practice is to use an encoding scheme, such as ISO 8601:2004(E). Format YYYY-MM-DD | 23078 | Date |
| Action | year | Year | The four-digit year in which the Event occurred, according to the Common Era Calendar. | 2008 | Integer |
| Action | fieldNotes | NA | One of a) an indicator of the existence of, b) a reference to (publication, URI), or c) the text of notes taken in the field about the Event. | notes available in Grinnell-Miller Library | String |

## Core Results

A metric is a value resulting from the reduction or processing of measurements taken at an event based on the procedures defined by the response design. Programs derive a variety of metrics from a single measurement. For stream habitat data at each event, programs take multiple types of measurements and produce various metrics from one measurement; for example, the measurement for pools produces both percent pools and pool frequency. Events are associated with measurements by the eventID, see the [Measurement Or Fact Table](Tables/MeasurementOrFact_table.csv) for the full definitions of terms.

From ODME2 Results.MeasurementsResultsValues in this implementation we merged the following into the Sample Feature table \* MethodID \* MeasurementRemark \* DataValue

Added a foregin key \*ActionID

### This implementation

#### To the Core Results from Results. MeasurementsResultsValues to Core Resoutls added:

* MethodID
* MeasurementRemark
* DataValue

#### The primary key is:

* ResultsID

#### Added foregin key:

* ActionID

#### Controlled Vocabularies

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ODMTable | Term | ODM | Description | Examples | DataType |
| Results | ValueID | ValueID | Unique integer identifier for each |  |  |
| data value. | 5 | Interger |  |  |  |
| Results | measurementType | NA | The nature of the measurement, fact, characteristic, or assertion. Recommended best practice is to use a controlled vocabulary. | tail length, “temperature”, “trap line length”, “survey area”, “trap type” | String |
| Results | measurementID | VariableID | An identifier for the MeasurementOrFact (information pertaining to measurements, facts, characteristics, or assertions). May be a global unique identifier or an identifier specific to the data set. | NA | Numeric |
| Results | measurementTerm | NA | A field name for the MeasurementOrFact. Term from the stream habitat controlled vocabulary. | Grad, WetWidth, ReachLen | String |
| Results | measurementValue | DataValue | The value of the measurement, fact, characteristic, or assertion. | 45, “20”, “1”, “14.5”, “UV-light” | NA |
| Results | measurementUnit | NA | The units associated with the measurementValue. Recommended best practice is to use the International System of Units (SI). | mm, “C”, “km”, “ha” | String |
| Results | measurementDeterminedDate | validDateTime | The date on which the MeasurementOrFact was made. Recommended best practice is to use an encoding scheme, such as ISO 8601:2004(E). | 1963-03-08T14:07-0600, “2009-02-20T08:40Z”, “1809-02-12”, “1906-06”, “1971”, “2007-03-01T13:00:00Z/2008-05-11T15:30:00Z”, “2007-11-13/15” | String |
| Results | measurementDeterminedBy | SourceID | A list (concatenated and separated) of names of people, groups, or organizations who determined the value of the MeasurementOrFact. | AREMP, NRSA, PIBO | String |
| Results | measurementMethod | MethodID | A description of or reference to (publication, URI) the method or protocol used to determine the measurement, fact, characteristic, or assertion. | URL of the MonitoringResources.rog field and if avilable calcuation method | String |
| Results | measurementRemarks | NA | Comments or notes accompanying the MeasurementOrFact. | tip of tail missing | String |

### Variable Controlled Vocabularies for Results Table

Critical to implementing the observation data model (ODM) for stream habitat monitoring, is solving semantic heterogeneity, or the differences in languages used to describe observations between datasets (CITE). There needs to be a controlled vocabularies for each data type integrated. The standard language enables the integration of multiple habitat monitoring program metrics into one data set.

We built the variable controlled vocabulary using metadata and metrics from four large scale, long-running federal stream habitat monitoring programs: Environmental Protection Agency (EPA) National Rivers & Streams Assessment (NRSA), Bureau of Land Management (BLM) Aquatic Assessment, Inventory, and Monitoring (AIM), the Forest Service Aquatic and Riparian Effective Monitoring Program (AREMP) and PACFISH/INFISH Biological Opinion (PIBO) Effectiveness Monitoring. Each program has unique objectives, spatial, temporal, response, and inference designs; yet, they produce similar metrics. These four programs collectively produce over 300 metrics but have only a subset of metrics in common across programs. The program leads and data managers from the four programs agreed on a subset of the metrics that can be shared across the programs; these can be found in the first draft of the [controlled vocabulary](Tables/StandardVocabulary.csv).

We focused on defining a set of metric we would integrate. Previous work has been completed compare some of these data sets and field collection procedures. (CITE) For our initial assessment of metric compadability There have been efforts in the past to comparei field processes for these programs, but we wanted to confirm with the project lead that fields were comparable. PNAMP lead an effort to facilitate conversations with data experts from the four programs to define the metrics that can be shared across the programs. The conversations are documented in Appendix A. We agreed that data in the same colume is compadable, even with this agreement we want to provide documentation of the field methods for each of the metrics included in the controlled vocabulary. To make this effective each individual data collection method needs to be documented in a stand alone, machine readable way. This means when the data set is reused the data collection methology can be accessed via APIs. Simplifying the creation of data portals, web maps and other user interfaces. We used MonitoringResources.org an online tool for documenting field data collection or analysis methods to so for each metric and program we document the field protocols (ADD TABLE OF CONTROLLED VOCABULARY AND METRIC IDs).

Using the metadata from each program we build a comprehensive list of the metrics produced across the programs and then facilitated discussion with experts from each program to define where metrics are comparable across programs (SEE METHOD REPORT). Use this approach we defined a controlled vocabulary of metrics to be integrated. Additional we defined a subset of metadata for data sets, locations, and events to include in the integrated data sets. Using all this information we build a data schema defining a relation data base observation and measurement model based on Darwin core and CAUSI data models. We then created a schema cross walk from the original data sets to the integrate data schema, and wrote code to pull the data from the orgial locations to be intergrated in the final data sets.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| measurementTerm | LongName | Description | Examples | DataType | measurementUnit |
| BFWidth | Average bankfull width from transects | Average bankfull width across transects for the reach. | NA | Numeric | m |
| countTransectsBFWidth | Count of cross sections | Number of cross sections included in the average bankfull width calculation | NA | Numeric | count of measurments |
| Grad | Gradient of stream reach | Mean slope of water surface (%) from the bottom of the reach to the top of the reach. | min: 0, max: 100 | Numeric | % |
| ReachLen | Length of sampling reach | Length of sampling reach measured along the thalweg | NA | Numeric | m |
| AvgBFWDRatio | Bankfull width to depth ratio at transects | Average bankfull width/depth ratio for the reach. | NA | Numeric | m |
| BFHeight | Bankfull Height | Average bankfull height measured from water surface across transects. | NA | Numeric | m |
| WetWidth | Average wetted width from transects | Average wetted width across transects. | NA | Numeric | m |
| MeanThalwegDepth | Average Thalweg Depth | Mean thalweg depth. Metric of how deep water was at the site. | NA | Numeric | m |
| countTransects | Count of Transects | Number of transects in the reach. | NA | NA | NA |
| Sin | Sinuosity of Local Stream Reach | Reach sinuosity (reach length along the thalweg divided by straight line distance between the bottom of the reach and the top of the reach) | NA | Numeric | NA |
| PctDry | Percent of Reach that is Dry | Percent of the reach that was dry | min: 0, max: 100 | Numeric | % |
| Beaver | Beaver Sign at Reach | Beaver value from the provider dataset. | NA | Numeric | NA |
| StreamOrder | Stream Order | Strahler stream order of the site from the provider dataset | NA | Numeric | NA |
| RPD | Residual pool depth | Average of the residual pool depth values for all pools in a reach. Residual pool depth the difference of the pool tail depth from the max depth. | NA | Numeric | m |
| PctPool | Percent pools | Percent of the sample reach (linear extent) classified as pool habitat | min: 0, max: 100 | Numeric | % |
| BankAngle | Bank angle | Measured angle of the stream bank; | min: 0, max: 180 | Numeric | degrees |
| PctStab | Percent stable banks | Percent of 42 banks lacking visible signs of active erosion (e.g., slump, slough, fracture) (units: %, min: 0, max: 100, n= 42) | min: 0, max: 100 | Numeric | % |
| D50 | Diameter of the 50th percentile streambed particle | Median diameter of bed surface particle size corresponding to the areal median (50th percentile of measured particles) . | min: 1, max: 4098 | Numeric | mm |
| PctFines2 | Percent of streambed particles <2mm | Percent of the stream bed area that is comprised of particles with a b-axis < 2 mm | min: 0, max: 100 | Numeric | % |
| PctFines6 | Percent of streambed particles <6mm | Percent of the stream bed area that is composed of particles with a b-axis < 6 mm | min: 0, max: 100 | Numeric | % |
| D16 | Diameter of the 16th percentile streambed particle | Bed surface particle size corresponding to the 16th percentile of measured particles | min: 1, max: 4098 | Numeric | mm |
| D84 | Diameter of the 84th percentile streambed particle | Bed surface particle size corresponding to the 84th percentile of measured particles | min: 1, max: 4098 | Numeric | mm |
| PctBdrk | Percent Bed Surface Bedrock | Percent of the stream bed area that is composed of particles with a b-axis > 4098mm (bedrock) | min: 0, max: 100 | Numeric | % |
| PoolTailFines2 | Percent pool tail fines < 2mm | Average percent fine sediment (< 2mm) on the pool tail | min: 0, max: 100 | Numeric | % |
| PoolTailFines6 | Percent pool tail fines < 6mm | Average percent fine sediment (< 6mm) on the pool tail | min: 0, max: 100 | Numeric | % |
| Temp | Mean annual tempeature | Average of mean daily temperatures during a year | NA | Numeric | Degrees Celsius |
| WinterMean | Mean winter temperature (Dec, Jan, Feb) | Average of mean daily temperatures during December, January, and February | NA | Numeric | Degrees Celsius |
| SpringMean | Mean spring temperature | Average of mean daily temperatures during March, April, and May | NA | Numeric | Degrees Celsius |
| SummerMean | Mean summer temperatures | Average of mean daily temperatures during June, July, and August | NA | Numeric | Degrees Celsius |
| AugustMean | Mean august temperature | Average of mean daily temperatures during August | NA | Numeric | Degrees Celsius |
| MeanFall | Mean fall temperature | Average of mean daily temperatures during September, October, and November | NA | Numeric | Degrees Celsius |
| LowMean | Minimum daily temperature | Lowest mean daily temperature during a year | NA | Numeric | Degrees Celsius |
| LowSevenDayAverage | Minimum weekly average temperature | Lowest seven-day running average of mean daily temperature during a year | NA | Numeric | Degrees Celsius |
| HighMean | Maximum daily temperature | Highest mean daily temperature during a year | NA | Numeric | Degrees Celsius |
| HighSevenDayAverage | Maximum weekly average temperature | Highest seven-day running average of mean daily temperature during a year | NA | Numeric | Degrees Celsius |
| DegreeDays | annual degree day s | Cumulative total of degree days during a year (1°C for 24 h = 1 degree day) | NA | Numeric | Count of days |
| SDMeanDaily | Annual Standard Deivation | Standard deviation of mean daily temperature during a year | NA | Numeric | NA |
| SDMeanWinterDaily | Winter Standard Deviation | Standard deviation of mean daily temperature during winter months | NA | Numeric | NA |
| SDMeanSpring | Spring Standard Deviation | Standard deviation of mean daily temperature during spring months | NA | Numeric | NA |
| SDMeanSummer | Summer SD | Standard deviation of mean daily temperature during summer months | NA | Numeric | NA |
| SDMeanAugust | August SD | Standard deviation of mean daily temperature during the month of August | NA | Numeric | NA |
| SDMeanFall | Fall SD | Standard deviation of mean daily temperature during fall months | NA | Numeric | NA |
| DiffMinMax | Range in extream daily termperature s | Difference between minimum and maximum mean daily temperatures during a |  |  |  |
| year (M9 minus M7) | NA | Numeric | NA |  |  |
| DifMinMaxWeekly | Range in streme weekly temperatures | Difference between minimum and maximum weekly average temperatures |  |  |  |
| during a year (M10 minus M8) | NA | Numeric | NA |  |  |
| SDAnnual | Interannual standard deviation of mean |  |  |  |  |
| annual | Interannual standard deviation in mean annual temperature | NA | Numeric | NA |  |
| SDMinWeekly | Interannual standard deviation of |  |  |  |  |
| minimum weekly | Interannual standard deviation in minimum weekly average temperature | NA | Numeric | NA |  |
| SDMaxWeekly | Interannual standard deviation of |  |  |  |  |
| maximum weekly | Interannual standard deviation in maximum weekly average temperature | NA | Numeric | NA |  |
| SD5PercentDegreeDay | Interannual standard deviation of 5% |  |  |  |  |
| degree days | Interannual standard deviation in date of 5% of degree days | NA | Numeric | NA |  |
| SD50PercentDegreeDay | Interannual standard deviation of 50% |  |  |  |  |
| degree days | Interannual standard deviation in date of 50% of degree days | NA | Numeric | NA |  |
| NumberMeanGT20 | Frequency of hot days | Number of days with mean daily temperatures >20°C | NA | Numeric | NA |
| NumberMeanLT2 | Frequency of cold days | Number of days with mean daily temperatures <2°C | NA | Numeric | NA |
| NumberDaysDec | Date of 5% degree days | Number of days from December 1 until 5% of degree days are accumulated | NA | Numeric | NA |
| NumberDaysDec25 | Date of 25% degree days | Number of days from December 1 until 25% of degree days are accumulated | NA | Numeric | NA |
| NumberDaysDec50 | Date of 50% degrees days | Number of days from December 1 until 50% of degree days are accumulated | NA | Numeric | NA |
| NumberDaysDec95 | Date of 95% degree days | Number of days from December 1 until 95% of degree days are accumulated | NA | Numeric | NA |
| NumberDaysBetween95and5 | Growing Season length | Number of days between the 95% and 5% of degree days (T5 minus T1) | NA | Numeric | NA |
| LongestDaysGT20 | Duration of hot days | Longest number of consecutive days with mean daily temperatures >20°C | NA | Numeric | NA |
| LongestDaysLT2 | Duration of cold days | Longest number of consecutive days with mean daily temperatures <2°C | NA | Numeric | NA |
| NumberDaysDec75 | Date of 75% degree days | Number of days from December 1 until 75% of degree days are accumulated | NA | Numeric | NA |
| Conductivity | Conductivity | Measure of the concentration of ionized materials in water, or the ability of water to conduct electrical current. Used as a measure of mineral / ionic concentration. | NA | Numeric | ppm |
| TotalNitrogen | Total Nitrogen | Measured total nitrogen value | NA | Numeric | ug/L |
| TotalPhosphorous | Total Phosphorous | Measured total phosphorous value | NA | Numeric | ug/L |
| SpecificConductance | Specific Conductance | Measured specific conductance value. The specific conductance is conductivity standardized to 25 degrees C. | min: 0, max: 65500 | Numeric | uS/cm |
| pH | pH | Measured pH value | min: 0, max: 14 | Numeric | NA |
| Turbidity | Turbidity | Average water clarity as measured by the suspended solids in the water column | NA | Numeric | NTU |

The working group crosswalked each of their program’s field names to the controlled vocabulary. We documented details of the metric compatibility discussions between the four programs in Appendix A of the [Data Exchange Specification](MetricLevelExchangeSpecifications.docx) document.

If partners wish to exchange additional metrics, the controlled vocabulary must be updated and cross-walk. The list of metrics from the four programs not included in the first draft of the standard vocabulary or data exchange specifications is here: [list of metrics not in the controlled vocabulary](Tables/NotInControlledVocabularyOrDES.csv)

# Use Case

We wrote code based on these data exchange specifications to share habitat metrics from three federal habitat monitoring programs: Environmental Protection Agency (EPA) National Rivers & Streams Assessment (NRSA), Bureau of Land Management (BLM) Aquatic Assessment, Inventory, and Monitoring (AIM),and the Forest Service Aquatic and Riparian Effective Monitoring Program (AREMP). The work flow pulls program information from ScienceBase, the exchange specifications and the field crosswalk from this repository, and data collection metrics documented from MonitoringResources.org [work flow diagram](https://github.com/rascully/Stream-Monitoring-Data-Exchange-Specifications/blob/master/Figures/WorkFlow.png). The R code to integrate data sets can be found at <https://github.com/rascully/Integrating-Stream-Monitoring-Data-From-Multiple-Programs> and the data set documentation in ScinceBase at ADD SCIENCEBAES LINK WHEN I CAN

# Conclusion

The data exchange specifications contain the details of what will be share and the format to be shared. We recognize preparing data to be shared requires an investment of time, resources, expertise, and careful documentation of the data collection process and the results. A recent opinion piece in Nature by Barend Mons (2020), the director of a Global Open FAIR office, recommends that ‘5% of research funds be invested in making data reusable’. Projects producing this type of data are already working beyond their capacity, so to integrate data between habitat programs, there needs to be support in project budgets or for a centralized data manager to help implement and updated the necessary documentation and code to share data.

# References

Mons, B. (2020). Invest 5% of research funds in ensuring data are reusable. Nature, 578(7796), 491.

Kulvatunyou, B., Morris, K. C., Ivezic, N., & Frechette, S. (2008). Development life cycle for semantically coherent data exchange specification. Concurrent Engineering, 16(4), 279-290.

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, et al. (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7(1): e29715. <https://doi.org/10.1371/journal.pone.0029715>

Wikipedia contributors. ‘Machine-readable data.’ Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 6 Aug. 2013. Web. 21 Aug. 2014.