Read Me Project Team Word

Becca Scully

2/25/2021

## 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). The Stream Monitoring Habitat Data Exchange Specifications are a standard for exchanging metric-level habitat data based on 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>. The Stream Habitat Metric Data Integration working group facilitated by the Pacific Northwest Aquatic Monitoring Partnership (<https://www.pnamp.org/project/habitat-metric-data-integration>) and the USGS adapted the Darwin Core standard for stream habitat metrics, and as a use case, integrate stream habitat metrics from three federal stream habitat monitoring programs in a separate Git Hub Repository: <https://github.com/rascully/Integrating-Stream-Monitoring-Data-From-Multiple-Programs>.

# Structure

We utilize the Darwin Core classes: Record-level, Location, Event, and Measurement or Fact [Data Structure](https://github.com/rascully/Stream-Monitoring-Data-Exchange-Specifications/blob/master/Figures/StructureOfDarwinCoreForHabitatMetrics.png). Class in Darwin Core is the title for a group of terms (Wieczorek et al. 2012). Record-level Class documents information about each data set and links to Location using the DatasetID. Location Class documents the location and metadata about a specific location; it is associated with a sampling event using the LocationID. Multiple events can be related to a single location. The Event Class documents the data collection event and metadata about the sampling event. The event is linked to the specific metric using the EventID. The Measurement or Fact Class documents the metrics and metadata about each metric. At each event, programs collect multiple measurements, producing numerous metrics. To promote transparent and consistent metadata, we facilitated a process to describe a controlled vocabulary defining the metrics that can be shared using these data exchange specifications. This type of data is suited to a star data schema due to the one to many relationships between locations and events, and events and metrics. We adapted the stream habitat metrics to the [star schema](https://github.com/rascully/Stream-Monitoring-Data-Exchange-Specifications/blob/master/Figures/HabiatDataSharingSchema.png)

## 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/>.

Record Level Class

Term

Description

Examples

DataType

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

String

type

The nature or genre of the resource.

This will be Stream Habitat for all values

String

modified

The most recent date-time on which the combined dataset was changed.

Recommended best practice is to use a date that conforms to ISO 8601:2004(E).

Date

rightsHolder

A person or organization owning or managing rights over the resource.

Example: BLM AIM, EPA NARS, USFS ARMEP

String

bibilographicCititation

A bibliographic reference for the resource as a statement indicating how this record should be cited (attributed) when used.

NA

String

InstitutionID

An identifier for the institution having custody of the object(s) or information referred to in the record.

Example) <http://biocol.org/urn:lsid:biocol.org:col:34777>, <http://grbio.org/cool/km06-gtbn>

String

CollectionID

An identifier for the collection or dataset from which the record was derived.

Example) <https://www.epa.gov/national-aquatic-resource-surveys>

String

datasetName

The name identifying the data set from which the record was derived.

Example: EPA NARS 2018 dataset

String

institutionCode

The name (or acronym) in use by the institution having custody of the object(s) or information referred to in the record

Example: AIM, NARS, AREMP, PIBO

String

## Location Class

Understanding where data are collected is critical to interpreting biological monitoring data. The Location class describes where information are collected, see the list of terms in the [Location table](Tables/Location_table.csv). There will be multiple locations in each data set. The locationID 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.

Term

Description

Examples

DataType

locationID

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

NA

String

verbatimlocationID

Number that identifies a unique sampling location. A site is a stream segment with a fixed starting and ending location for sampling

NA

String

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 as the location.

Numeric

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 as the location.

Numeric

verbatimWaterbody

The water body name from the original data set.

For this data set this field is often refered to as Stream Name.

String

verbatimCoordinateSystem

The spatial coordinate system for the verbatimLatitude and verbatimLongitude or the verbatimCoordinates of the Location.

NA

String

StateProvince

The name of the next smaller administrative region than country (state, province, canton, department, region, etc.) in which the Location occurs.

NA

String

decimalLatitude

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

decimalLongitude

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

geodeticDatum

The ellipsoid, geodetic datum, or spatial reference system (SRS) upon which the geographic coordinates given in decimalLatitude and decimalLongitude as based.

NA

string

## Event Class

The Event class describes an action that occurs at a specific time frame see the [Event table](Tables/Event_table.csv) for the terms. 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.

Term

Description

Examples

DataType

verbatimEventID

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

eventID

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

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"](http://dx.doi.org/10.1111/j.1466-8238.2009.00467.x%22);,”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"](http://www.bsc-eoc.org/download/Owl.pdf%22);

Integer

verbatimEventDate

The verbatim original representation of the date and time information for an Event.

spring 1910, “Marzo 2002”, “1999-03-XX”, “17IV1934”

Date

EventDate

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).

1963-03-08T14:07-0600 is 8 Mar 1963 2:07pm in the time zone six hours earlier than UTC, “2009-02-20T08:40Z” is 20 Feb 2009 8:40am UTC, “1809-02-12” is 12 Feb 1809, “1906-06” is Jun 1906, “1971” is just that year, “2007-03-01T13:00:00Z/2008-05-11T15:30:00Z” is the interval between 1 Mar 2007 1pm UTC and 11 May 2008 3:30pm UTC, “2007-11-13/15” is the interval between 13 Nov 2007 and 15 Nov 2007.

Date

verbatimEventTime

The time or interval during which an Event occurred. Recommended best practice is to use an encoding scheme, such as ISO 8601:2004(E).

14:07-0600 is 2:07pm in the time zone six hours earlier than UTC, “08:40:21Z” is 8:40:21am UTC, “13:00:00Z/15:30:00Z” is the interval between 1pm UTC and 3:30pm UTC.

Date

day

The integer day of the month on which the Event occurred.

9, “28”

Date

month

The ordinal month in which the Event occurred.

1 (=January), “10” (=October)

Date

year

The four-digit year in which the Event occurred, according to the Common Era Calendar.

2008

Integer

fieldNumber

An identifier given to the event in the field. Often serves as a link between field notes and the Event.

RV Sol 87-03-08

String

fieldNotes

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

eventRemark

Comments or notes about the Event.

after the recent rains the river is nearly at flood stage

String

## Measurement Or Fact (Metrics) Class

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.

Term

Description

Examples

DataType

measurementID

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

String

measurementType

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

measurementValue

The value of the measurement, fact, characteristic, or assertion.

45, “20”, “1”, “14.5”, “UV-light”

NA

measurementAccuracy

The description of the potential error associated with the measurementValue.

0.01, “normal distribution with variation of 2 m”

String

measurementUnit

The units associated with the measurementValue. Recommended best practice is to use the International System of Units (SI).

mm, “C”, “km”, “ha”

String

measurementDeterminedDate

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

measurementDeterminedBy

A list (concatenated and separated) of names of people, groups, or organizations who determined the value of the MeasurementOrFact.

Javier de la Torre, “Julie Woodruff; Eileen Lacey”

String

measurementMethod

A description of or reference to (publication, URI) the method or protocol used to determine the measurement, fact, characteristic, or assertion.

minimum convex polygon around burrow entrances for a home range area, “barometric altimeter” for an elevation

String

measurementRemarks

Comments or notes accompanying the MeasurementOrFact.

tip of tail missing

String

### Controlled Vocabulary

We defined a controlled vocabulary of metrics to select from for the MeasurmetID and populate the MeasurmentUnit in the Measurement or Fact Class. The standard language enables the integration of multiple habitat monitoring program metrics into one data set.

We built the 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).

Term

LongName

Description

Examples

DataType

Unit

BFWidth

Average bankfull width from transects

Average bankfull width across transects.

NA

Numeric

m

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

%

RchLen

Length of sampling reach

Length of sampling reach measured along the thalweg

NA

Numeric

m

BFWDRatio

Bankfull width to depth ratio at transects

Average Bankfull Width/Depth Ratio across transects.

NA

Numeric

m

WetWidth

Average wetted width from transects

Average wetted width across transects.

NA

Numeric

m

WetWidthToDepth

Wetted width to depth ratio at transects

Mean Wetted Width/Thalweg Depth ratio

NA

Numeric

ratio

countTransects

Count of Transects

Number of transects in the reach.

NA

NA

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 combability 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.