Contents

[Introduction 3](#_Toc60665161)

[Habitat Data Exchange Specifications 3](#_Toc60665162)

[Practical Example 13](#_Toc60665163)

[Providing Data to Users 14](#_Toc60665164)

[Conclusion 14](#_Toc60665165)

[References 14](#_Toc60665166)

[Appendix A 16](#_Toc60665167)

# Introduction

Streams are critical to fish, aquatic community structure, and overall watershed health. State, Federal and Tribal entities collect in-stream habitat data to assess the resources' status and trends unique to their management questions.  Climate change, urbanization, and multi-use land management have made it necessary to determine the resources' quality and trends across jurisdictional boundaries using information from multiple collection efforts (Katz et al. 2012). It is not straightforward to combine data from various monitoring programs due to differences in response and survey (spatial and temporal) designs and data management practices. Additionally, data produced by these programs are not always findable, accessible, interoperable, and reusable (FAIR) (Wilkinson et al. 2016).  There is no centralized repository or data dictionary for this type of information.  A standard theory in data science states researchers spend 80% of their time organizing, fixing mistakes, and cleaning data, leaving only 20% of their time to analyze data (Mons 2020).   Well-established rules for integrating and sharing stream habitat data from multiple sources will decrease the time spent finding and organizing data providing accurate and timely information for building indicators, completing analysis, and making decisions.  Over the last few months, PNAMP has facilitated a working group to review the response design, specifically the metrics and metadata from four federal stream habitat monitoring programs, to create a process for combining metric level stream habitat data.

To make stream habitat data available, data exchange specifications are needed,  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). We modeled this data exchange specification for stream habitat data on the Darwin Core principles for exchanging biological data. Darwin Core provides terms and schema for integrating and exchanging biodiversity information (Wieczorek 2012), and it is highly flexible and adapted for many types of information.

# Habitat Data Exchange Specifications

The stream habitat data exchange specifications are structured based on the simple Darwin Core standard. Using data from four federal stream habitat monitoring programs as a model, we outline the elements needed to share or integrate metrics from stream habitat data sets. We included core dataset information, location information, data collection event information, and individual metrics data. We follow the recommend Darwin Core star database schema due to how one location relates to multiple events, and programs produce numerous metrics at a single event (Figure 1). The approach is flexible enough to be adapted later to support the transfer of other data types such as macroinvertebrate counts or riparian vegetation.

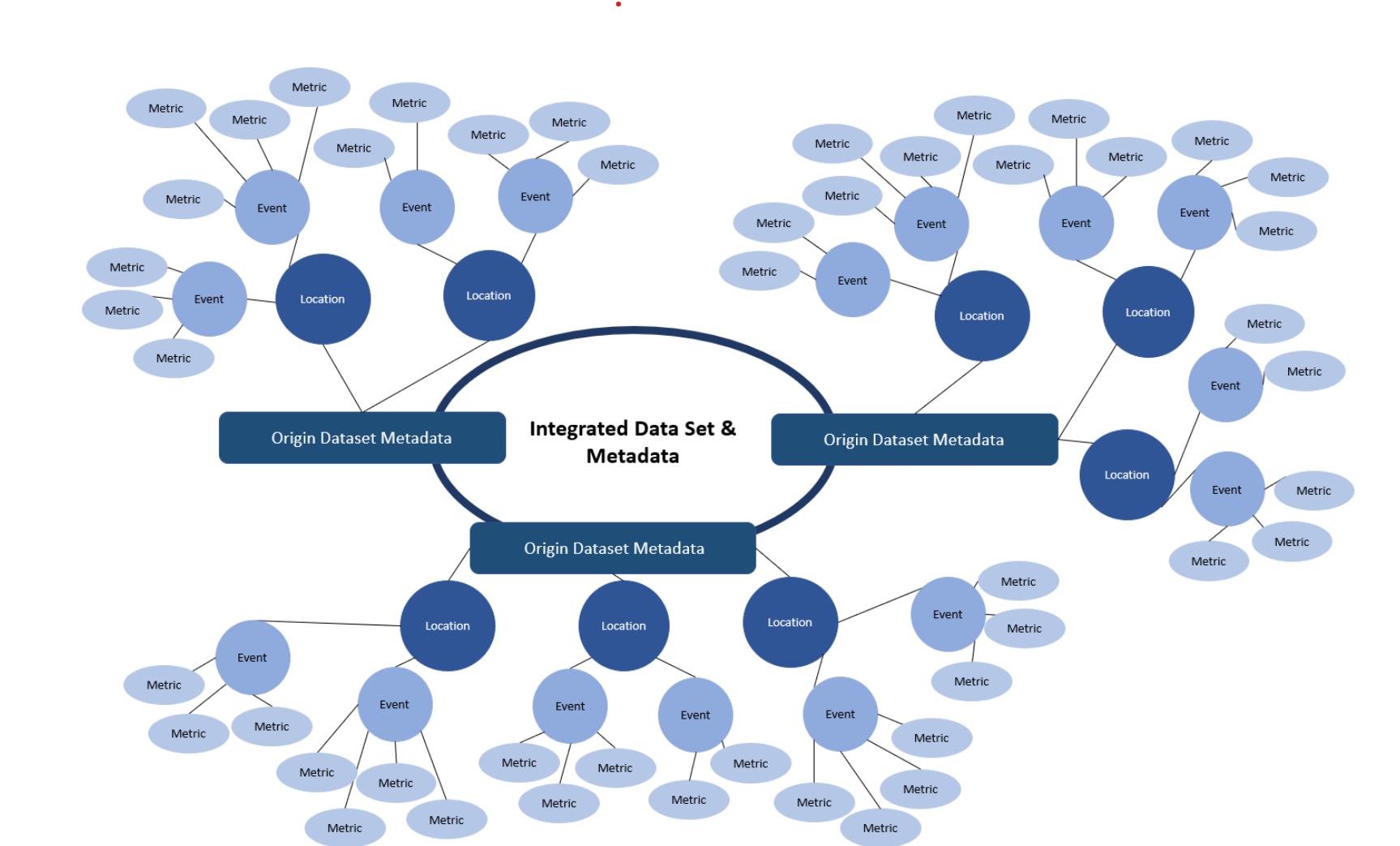


Figure 1 The one to many relationships between datasets, locations, events, and metrics based on the Darwin core principles modified to work with stream habitat data.

We utilize the Darwin Core classes Record-level, Location, Occurrence, and Measurement or Fact. Class in Darwin Core is the title for a group of terms. Record-level class documents information about each dataset and is linked to Location using the DatasetID. Location class document the location and metadata about that location; it is associated with a sampling event using the LocationID. Multiple events can be related to one 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 collet multiple measurements, producing numerous metrics. To promote transparent and consistent metadata, we facilitated a process to describe a standard vocabulary defining the metrics that can be shared using these data exchange specifications.

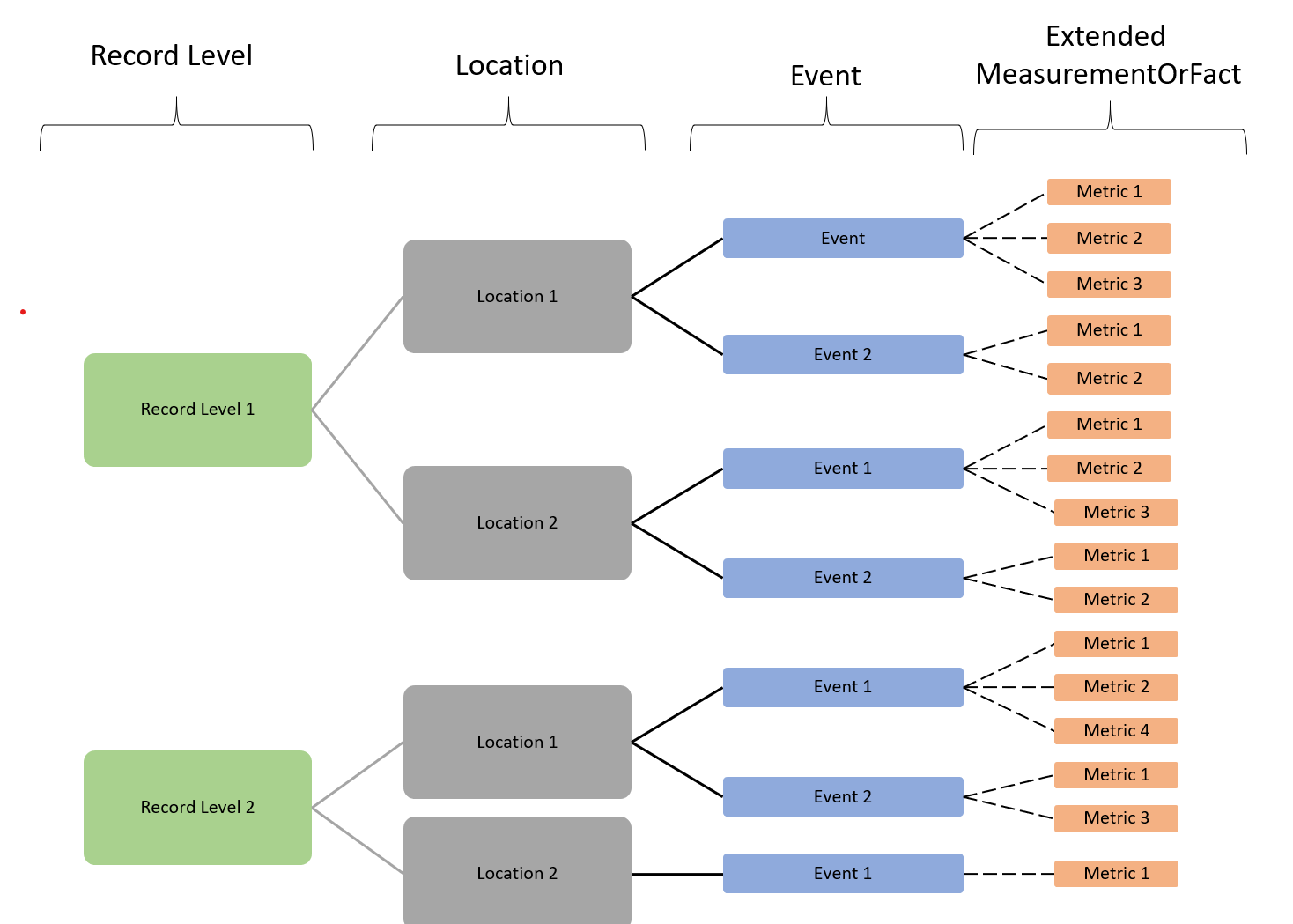


Figure Overview of the stream habitat data exchange specifications format. Record-level elements help us document dataset information and are linked to multiple locations based on the locationID (thin gray line). Locations class stores location information, locations are associated with one or many Events or sample occurrences using the LocaitonID (bold black lines). Metrics are related to events using the EventID field of the MeasurementOrFact Extension (dashed lines).

#### Record Level

The Record Level class documents the core elements of a data set, including information about the origin of the dataset, who collected the data, and how to cite the dataset (Table 1). A data set is a collection of data collection events; for example, a program releases a data set every five years containing all the data collection events occurring in the previous five years. We recommend storing metadata in a trusted online data repository to ensure 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. To can find more information about ScienceBase, go to <https://www.sciencebase.gov/about/>.

Table 1 Describe information about the dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| Record Level | | | |
| Enter information about the dataset, the data provider and the data collector | | | |
| **TERM** | **DESCRIPTION** | **EXAMPLES** | **DATA TYPE** |
| 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. This is the key to link the dataset to the location table. | Integer (need to check this) |
| type | The nature or genre of the resource. | EventMeasurments for all entries.   Must be populated with a value from the DCMI type vocabulary (http://dublincore.org/documents/2010/10/11/dcmi-type-vocabulary/). | 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. |  | 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 |
| basisOfRecord | The specific nature of the data record. | Recommended best practice is to use the standard label of one of the Darwin Core classes. | String |
| informationWithheld | Additional information that exists, but that has not been shared in the given record. | location information not given for endangered species, collector identities withheld | ask about tissue samples | String |

#### Location

Understanding where data are collected is critical to interpreting biological monitoring data. The Location class describes where information is collected (Table 2). There will be multiple locations in each dataset. To join data from various sources to one dataset, latitudes and longitude information must be constant amongst datasets; therefore, all latitude and longitudes are converted to WGS1984. The locationID is the key to link locations to events.

Table 2 This table stores information related to the location of the data collection events.

|  |  |  |  |
| --- | --- | --- | --- |
| Location | | | |
| A spatial region or named place. | | | |
| **TERM** | **DESCRIPTION** | **EXAMPLES** | **DATA TYPE** |
| datasetID\* | An identifier for the set of data. A global unique identifier or an identifier specific to a collection or institution. | The ScienceBase ID number of the dataset. DatasetID is the key to link locations to a specific dataset. | Integer (need to check this) |
| locationID\* | An identifier for the set of location information (data associated with dcterms:Location). May be a global unique identifier or an identifier specific to the data set. | locationID is the location identification for the integrated data set. It is the concatenation of the verbatimlocationID and the institutionCode. Example) 5483AIM, 88963AREMP, WtR563EPA | String |
| verbatimLocationID | Unique number that identifies one sample of a particular site. | In habitat data sets, this is often referred to as the SiteID. This ID identifying the location, while the reachID corresponds to the EventID. We use locationID to corresponds to Darwin Core standard. | 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. |  | Double |
| 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. |  | Double |
| verbatimWaterbody | The name of the water body in which the Location occurs. | The water body name from the original data set. | String |
| verbatimCoordinateSystem | The spatial coordinate system for the verbatimLatitude and verbatimLongitude or the verbatimCoordinates of the location. |  | String |
| stateProvince | The name of the next smaller administrative region than country (state, province, canton, department, region, etc.) in which the location occurs. |  | 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. |  | Double |
| 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. |  | Double |
| geodeticDatum | The ellipsoid, geodetic datum, or spatial reference system (SRS) upon which the geographic coordinates given in decimalLatitude and decimalLongitude as based. |  | String |

#### Event

The Event class describes an action that occurs at a specific time frame (Table 3). To assess the resource trend as a response to management actions, stream habitat monitoring programs often implement rotating panel design, meaning that the project returns to a single location multiple times during the study duration. Therefore a dataset will contain numerous locations, and each location can include numerous events.

Table 3 Event table describing a single data collection action at a location.

|  |  |  |  |
| --- | --- | --- | --- |
| Event | | | |
| An action that occurs at some location during some time. | | | |
| **TERM** | **DESCRIPTION** | **EXAMPLES** | **DATA TYPE** |
| locationID\* | An identifier for the set of location information (data associated with dcterms:Location). May be a global unique identifier or an identifier specific to the data set. |  | 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. |  | 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", "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" | string |
| 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 |
| eventTime | 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. | string |
| year | The four-digit year in which the Event occurred, according to the Common Era Calendar. | 2008 | integer |
| month | The ordinal month in which the Event occurred. | 1 (=January), "10" (=October) | integer |
| day | The integer day of the month on which the Event occurred. | 9, "28" | integer |
| verbatimEventDate | The verbatim original representation of the date and time information for an Event. | spring 1910, "Marzo 2002", "1999-03-XX", "17IV1934" | string |
| 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 |
| eventRemarks | Comments or notes about the Event. | after the recent rains the river is nearly at flood stage | string |

#### Measurement Or fact (Metrics)

A metric is a value resulting from the reduction or processing of measures 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 measures and produce various metrics from one measure; for example, the metrics percent pools and pool frequency are calculated from pool measurements. Events are associated with measurement by the eventID.

Table 4. Metric Table

|  |  |  |  |
| --- | --- | --- | --- |
| Measurement or Fact | | | |
| A measurement of or fact or metric about a resource | | | |
| **TERM** | **DESCRIPTION** | **EXAMPLES** | **DATA TYPE** |
| 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. |  | string |
| 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. |  | 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" |  |
| 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 |
|  |  |  |  |

We defined a standard vocabulary of metrics to select from for the MeasurmetID and populate the MeasurmentUnit. The standard language enables the integration of multiple habitat monitoring program metrics.

We built the standard 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. Metrics are a value resulting from the reduction or processing of measurements taken at a site at a specific temporal unit one or more times during the study period based on the procedures defined by the response design. You can derive a variety of metrics from the original measurements. These four programs collectively produce over 300 metrics, but have only a subset of metrics across programs. The program leads, and data managers from the four programs agreed on a sub-set of the metrics that can to share across the programs. This subset is the first draft of the controlled vocabulary().

The working group cross-walked each of their program's field names to the standard vocabulary. We also documented detail of the metric combability discussions between the four programs in Appendix A. The burden is on the data users to ultimately decide if the methods are comparable enough to answer their specific management questions based on metadata in MonitoringResources.org.

If partners wish to exchange additional metrics, they must update the controlled vocabulary on GitHub.

Table 6. Controlled vocabulary for habitat metrics. This list is a subset of metrics produced across the four habitat monitoring programs used to build this data exchange standard. Metrics can be added to this list with the proper documentation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LongName** | **Definition** | **DataType** | **NotesCodesConventions** | **Unit** |
| Average bankfull width from transects | Average bankfull width across transects. | Double |  | m |
| Gradient of stream reach | Mean slope of water surface (%) from the bottom of the reach to the top of the reach. | Double | min: 0, max: 100 | % |
| Length of sampling reach | Length of sampling reach measured along the thalweg | Double |  | m |
| Bankfull width to depth ratio at transects | Average Bankfull Width/Depth Ratio across transects. | Double |  | m |
| Average wetted width from transects | Average wetted width across transects. | Double |  | m |
| Wetted width to depth ratio at transects | Mean Wetted Width/Thalweg Depth ratio | Double |  | ratio |
| 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. | Double |  | m |
| Percent pools | Percent of the sample reach (linear extent) classified as pool habitat | Double | min: 0, max: 100 | % |
| Percent of Reach that is Dry | Percent of the reach that was dry | Double | min: 0, max: 100 | % |
| Stream Order | Strahler stream order of the site from the provider dataset | Double |  | NA |
| Bank angle | Measured angle of the stream bank; | Double | min: 0, max: 180 | degrees |
| 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) | Double | min: 0, max: 100 | % |
| Diameter of the 50th percentile streambed particle | Median diamerter of bed surface particle size corresponding to the areal median (50th percentile of measured particles) . | Double | min: 1, max: 4098 | mm |
| Percent of streambed particles <2mm | Percent of the stream bed area that is comprised of pratices with a b-axis < 2 mm | Double | min: 0, max: 100 | % |
| Percent of streambed particles <6mm | Percent of the stream bed area that is coporsed of particles with a b-axis < 6 mm | Double | min: 0, max: 100 | % |
| Percent pool tail fines < 2mm | Average percent fine sediment (< 2mm) on the pool tail | Double | min: 0, max: 100 | % |
| Percent pool tail fines < 6mm | Average percent fine sediment (< 6mm) on the pool tail | Double | min: 0, max: 100 | % |
| Diameter of the 16th percentile streambed particle | Bed surface particle size corresponding to the 16th percentile of measured particles | Double | min: 1, max: 4098 | mm |
| Diameter of the 84th percentile streambed particle | Bed surface partical size corresponding to the 84th percentile of measured particles | Double | min: 1, max: 4098 | mm |
| Percent Bed Surface Bedrock | Percent of the stream bed area that is coporsed of particles with a b-axis > 4098mm (bedrock) | Double | min: 0, max: 100 | % |
| 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. | Double |  | ppm |
| Total Nitrogen | Measured total nitrogen value | Double |  | ug/L |
| Total Phosphorous | Measured total phosphorous value | Double |  | ug/L |
| Specific Conductance | Measured specific conductance value. The specific conductance is conductivity standardized to 25 degrees C. | Double | min: 0, max: 65500 | uS/cm |
| pH | Measured pH value | Double | min: 0, max: 14 | NA |
| Turbidity | Average water clarity as measured by the suspended solids in the water column | Double |  | NTU |
| Mean annual tempeature | Average of mean daily temperatures during a year |  |  | Degrees Celsius |
| Mean winter temperature (Dec, Jan, Feb) | Average of mean daily temperatures during December, January, and February |  |  | Degrees Celsius |
| Mean spring temperature | Average of mean daily temperatures during March, April, and May |  |  | Degrees Celsius |
| Mean summer temperatures | Average of mean daily temperatures during June, July, and August |  |  | Degrees Celsius |
| Mean august temperature | Average of mean daily temperatures during August |  |  | Degrees Celsius |
| Mean fall temperature | Average of mean daily temperatures during September, October, and November |  |  | Degrees Celsius |
| Minimum daily temperature | Lowest mean daily temperature during a year |  |  | Degrees Celsius |
| Minimum weekly average temperature | Lowest seven-day running average of mean daily temperature during a year |  |  | Degrees Celsius |
| Maximum daily temperature | Highest mean daily temperature during a year |  |  | Degrees Celsius |
| Maximum weekly average temperature | Highest seven-day running average of mean daily temperature during a year |  |  | Degrees Celsius |
| annual degree day s | Cumulative total of degree days during a year (1°C for 24 h = 1 degree day) |  |  | Count of days |
| Annual Standard Deivation | Standard deviation of mean daily temperature during a year |  |  |  |
| Winter Standard Deviation | Standard deviation of mean daily temperature during winter months |  |  |  |
| Spring Standard Deviation | Standard deviation of mean daily temperature during spring months |  |  |  |
| Summer SD | Standard deviation of mean daily temperature during summer months |  |  |  |
| August SD | Standard deviation of mean daily temperature during the month of August |  |  |  |
| Fall SD | Standard deviation of mean daily temperature during fall months |  |  |  |
| Range in extream daily termperature s | Difference between minimum and maximum mean daily temperatures during a year (M9 minus M7) |  |  |  |
| Range in streme weekly temperatures | Difference between minimum and maximum weekly average temperatures during a year (M10 minus M8) |  |  |  |
| Interannual standard deviation of mean annual | Interannual standard deviation in mean annual temperature |  |  |  |
| Interannual standard deviation of minimum weekly | Interannual standard deviation in minimum weekly average temperature |  |  |  |
| Interannual standard deviation of maximum weekly | Interannual standard deviation in maximum weekly average temperature |  |  |  |
| Interannual standard deviation of 5% degree days | Interannual standard deviation in date of 5% of degree days |  |  |  |
| Interannual standard deviation of 50% degree days | Interannual standard deviation in date of 50% of degree days |  |  |  |
| Frequency of hot days | Number of days with mean daily temperatures >20°C |  |  |  |
| Frequency of cold days | Number of days with mean daily temperatures <2°C |  |  |  |
| Date of 5% degree days | Number of days from December 1 until 5% of degree days are accumulated |  |  |  |
| Date of 25% degree days | Number of days from December 1 until 25% of degree days are accumulated |  |  |  |
| Date of 50% degrees days | Number of days from December 1 until 50% of degree days are accumulated |  |  |  |
| Date of 75% degree days | Number of days from December 1 until 75% of degree days are accumulated |  |  |  |
| Date of 95% degree days | Number of days from December 1 until 95% of degree days are accumulated |  |  |  |
| Growing Season length | Number of days between the 95% and 5% of degree days (T5 minus T1) |  |  |  |
| Duration of hot days | Longest number of consecutive days with mean daily temperatures >20°C |  |  |  |
| Duration of cold days | Longest number of consecutive days with mean daily temperatures <2°C |  |  |  |

# Practical Example

We use this data exchange standard to build an integrated data set. We have a paper in progress describing the combined data set, but to better explain this data structure, below, we include a practical example of data from the EPA NARS data.

First, we create the Record-Level table by extracting the data set information from ScienceBase.

ADD Practical EXAMPLE

Next, we can construct the Location table. This table would contain all the locations where sampling occurred. The first column of the table references the Record-level table.

ADD TABLE

Next, we build the Event table. The event table includes all the

Finally, there is the MeasuremtOrFact extension table, which has the metrics from the data collection events.

# Providing Data to Users

We model this data schema on the Darwin Core star schema. The star schema is an efficient format for updating and storing data, but not for quick and efficient data analysis. Therefore we will build tools for managers and analysts to download the data in an analysis-ready format based on the Tidy Data principles of each observation is a row; each variable is a column (Wickham 2014).  We chose this structure to simplify loading the data into R and Python for analysis. Each data collect event is a row, and the multiple reach level metrics produced and metadata from an event are fields in the data set. Fields refer to the column headers in the data set, and fields are defined by these data exchange specifications.

To simplify this data's reuse, we will also link the data to other data sets to provide environmental context to each data collection location, such as elevation, land ownership, management unit, percent forested, etc.

# 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 supported in project budgets or for a centralized data manager to help implement and updated the necessary documentation and code to share data.

Outlining a data exchange specification is only the first step towards delivering timely stream habitat data across jurisdictional boundaries. We need additional statistical support to answer the question of how to integrate sites selected using different site selection methods.

# References

Al-Chokhachy, R., & Roper, B. B. (2010). Different approaches to habitat surveys can impact fisheries management and conservation decisions. *Fisheries*, *35*(10), 476-488.

Chase, K. J., Bock, A. R., & Sando, R. (2016). *Sharing our data—An overview of current (2016) USGS policies and practices for publishing data on ScienceBase and an example interactive mapping application* (No. 2016-1202, pp. 1-10). US Geological Survey.

"How the Exchange Network works:" EPA. Environmental Protection Agency, 14 Feb. 2013. Web. 23 July 2014. <http://www.epa.gov/exchangenetwork/info/>.

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

Puls, A., Dunn, A., Hudson, G. (2014). Evaluation and Prioritization of Stream Habitat Monitoring in the Lower Columbia Salmon and Steelhead Recovery Domain as related to the Habitat Monitoring Needs of ESA Recovery Plans. PNAMP Series 2014-003. URL. <http://www.pnamp.org/document/4769>

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., ... & Vieglais, D. (2012). Darwin Core: an evolving community-developed biodiversity data standard. *PloS one*, *7*(1).

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

# Appendix A

Metrics are a value resulting from the reduction or processing of measurements taken at a site and temporal unit at one or more times during the study period based on the procedures defined by the response design. A variety of metrics can be derived from the original measurements.

Variability in data collection between sampling protocols can affect the interpretation of habitat quality and quality (Al-Chokhachy and Roper 2010). Therefore to integrate metrics from multiple streams habitat monitoring programs, robust machine and human-readable documentation of the step by step process used to collect and analyzed data are required.

Channel Dimensions -Bankfull Width

**Discussion:**All four programs produce average bankfull width using comparable field and analysis methods.

**Conclusion:**

* Include average bankfull width

**For Data Exchange Specifications & Data Discovery Tools:**

* Include average bankfull width
* Update methods in MonitoringResources.org

**Controlled Vocabulary**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LongName** | **Definition** | **DataType** | **NotesCodesConventions** | **Unit** | **Field** |
| Average bankfull width from transects | Average bankfull width across transects. | Double |  | m | BFWidth |

**Crosswalk**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Field** | **USFS\_AREMP\_Field** | **USFS\_AREMP\_Method\_Comment** | **BLM\_AIM\_Field** | **BLM\_AIM\_Method\_Comment** | **EPA\_NARS\_Field** | **EPA\_NARS\_Method\_Comment** | **USFS\_PIBO\_Field** | **USFS\_PIBO\_Method\_Comment** |
| BFWidth | average\_bfwidth |  | BNKFLL\_WT |  | XBKF\_W |  | Bf |  |

Chanel Dimensions - Gradient

**Discussion:**Gradient is an essential metric of in-stream habitat monitoring because many of the other metrics, such as pools and wood, are related to the localized stream gradient. Calculate gradient by dividing the elevation change along the reach by the thalweg's total reach length (Roper et al. 2010).

Gradient has a low signal to noise, meaning it is measured consistently across crews within programs (Roper et al. 2010).  Of the four programs in this exercise, BLM AIM, USFS AREMP, and PIBO have similar field methodologies and are comparable. EPA NARS implements different mythology for measuring gradient depending on the site based on type.

For EPA NARS the variable “*XSLOPE*” is the (best) slope value for all data uses.  It is in all EPA physical habitat data sets (EMAP, REMAP, WSA, NRSA).  Starting in 2008 with NRSA 2008-2009, we additionally include the variables “*XSLOPE\_FIELD”, “XSLOPE\_MAP*”, and “*PCTCLINOMETER*” in more detailed data sets that are also publicly available.  Certain reach attributes trigger using a different methodology, and those values are flagged in a comments field. There can be 2-3 variables for slope/gradient. It’s inappropriate to combine low gradient clinometer gradients with more precise laser level gradients. 

* "*XSLOPE*"                           "Mean Slope of water surface (%)" *This is the “best” slope value and is equal to either field or map slope.*
* "*XSLOPE\_FIELD*"               "Reach mean channel slope according to field-based measurements"
* "*XSLOPE\_MAP*"                 "Reach mean channel slope according to map-based calculations"
* “*PCTCLINOMETER*”           "Percent of slope values based on hand held clinometer" *There are 10 field measurements of slope, measured over the distances between each of the 11 transects --- this tells percentage of the measurements that are questionable if the slope is low (e.g., <2%)*

The programs have shown that the gradient of a reach does not change a lot over time. Each program deals with this differently:

* AREMP assigns one gradient to the reach once they get two values that are the same.
* PIBO has an average gradient. They take the measurement each time at every site.
* AIM makes sure measurements are within a certain precision and continue to take measurements until within that precision.
* EPA NARS takes the measurements every time visiting a site, even within the same year. Whether they are averaged or treated individually depends on the question being asked.

We need to think about this when we discuss how to update the integrated data set over time.  

**Conclusion:**

* Gradient is critical for analysis and must be included in an integrated data set.
* Metadata is necessary to know how data are collected.
* Don’t share incomparable data in the same column of an integrated data set.
* Gradient is critical to many analyses, so need to find a way to include EPA NARS gradient values and indicate the differences in methodology.

**For Data Exchange Specifications & Data Discovery Tools:**

* Each column represents one metric, meaning all values in a column of the combined data set are comparable across the programs.
* BLM AIM, USFS AREMP and PIBO are comparable, EPA NARS uses multiple methods to collect the data and may or may not be comparable.
* We need to think about how each program updates its gradient measures as we start talking about automating the updates of the integrated data set over time.

**Controlled Vocabulary**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LongName** | **Definition** | **DataType** | **NotesCodesConventions** | **Unit** | **Field** |
| Gradient of stream reach | Mean slope of water surface (%) from the bottom of the reach to the top of the reach. | Double | min: 0, max: 100 | % | Grad |

**Crosswalk**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Field** | **USFS\_AREMP\_Field** | **USFS\_AREMP\_Method\_Comment** | **BLM\_AIM\_Field** | **BLM\_AIM\_Method\_Comment** | **EPA\_NARS\_Field** | **EPA\_NARS\_Method\_Comment** | **USFS\_PIBO\_Field** | **USFS\_PIBO\_Method\_Comment** |
| Grad | gradient |  | SLPE |  | XSLOPE |  | Grad |  |

Channel dimension – Reach Length

**Discussion:**No Discussion

**Conclusion:** Comparable across all four programs.

**For Data Exchange Specifications & Data Discovery Tools:**

* Include reach length

**Controlled Vocabulary**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LongName** | **Definition** | **DataType** | **NotesCodesConventions** | **Unit** | **Field** |
| Length of sampling reach | Length of sampling reach measured along the thalweg | Double |  | m | RchLen |

**Crosswalk**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Field** | **USFS\_AREMP\_Field** | **USFS\_AREMP\_Method\_Comment** | **BLM\_AIM\_Field** | **BLM\_AIM\_Method\_Comment** | **EPA\_NARS\_Field** | **EPA\_NARS\_Method\_Comment** | **USFS\_PIBO\_Field** | **USFS\_PIBO\_Method\_Comment** |
| RchLen | REACH\_LENGTH |  | TOT\_RCH\_LEN |  | REACHLEN |  | RchLen |  |

Channel Dimensions - Bankfull width to depth ratio

**Discussion:** Bankfull width to depth ratio is not in all four data sets. AREMP does not share this metric in the public-facing data se AIM could compute this metric, but they would need to assume a rectangular channel, because they only collect thalweg depth and depth across the cross-section.  Additionally, we need to understand how the metric is calculated. Do the programs calculate BFWidth divided by thalweg depth or the cross-section width divided by the average of the cross-section profiles (Rosgen method)?

**Questions:**

* How does each program calculate bankfull width to depth ratio? Sum of the averages or the average of the sums?
* Does assuming a rectangular channel skew values compared to the USFS AREMP and PIBO, and EPA NARS method? Can we call the AIM “comparable”?
* How do we get the AREMP metric in public data sets?

**Conclusion:** We would like to include this metric in the integrated data set, but we need to figure out how the data are analyzed to understand if it is comparable across all four data sets.

**For Data Exchange Specifications & Data Discovery Tools:**

* TBD

**Controlled Vocabulary**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LongName** | **Definition** | **DataType** | **NotesCodesConventions** | **Unit** | **Field** |
| Bankfull width to depth ratio at transects | Average Bankfull Width/Depth Ratio across transects. | Double |  | m | BFWDRatio |

**Crosswalk**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Field** | **USFS\_AREMP\_Field** | **USFS\_AREMP\_Method\_Comment** | **BLM\_AIM\_Field** | **BLM\_AIM\_Method\_Comment** | **EPA\_NARS\_Field** | **EPA\_NARS\_Method\_Comment** | **USFS\_PIBO\_Field** | **USFS\_PIBO\_Method\_Comment** |
| BFWDRatio | ave\_widthDepth\_ratio | We do have bankfull depth to the wetted edge, which would be comparable to the controlled vocab definition, but it would need to be extracted. |  |  | BFWD\_RAT |  | WDTrans |  |

Channel Dimensions – Bankfull height

**Discussion:**Bankfull height might not be useful by itself but combined with other information, it could be interesting. AREMP and PIBO don't calculate and share bankfull height.  BLM AIM suggests calculating different values such as floodplain connectivity, which might be more useful in a combined data set.

**Question:**

* Are there floodplain metrics that could be calculated by all four programs, released, and used in the integrated data set?

**Conclusion:**

* Don’t include in the initial controlled vocabulary. Continue the discussion on producing flood plain metrics?

**For Data Exchange Specifications & Data Discovery Tools:**

* Omit BFHeight in the first draft of the controlled vocabulary and crosswalk.

Channel Dimensions - Average wetted width from transects & Wetted width to depth ratio

**Discussion:**Average wetted width from transects is a crucial variable, and the methodology is the same between all the programs, so it is low hanging fruit to include. USFS PIBO and AREMP collect this data but don’t share it as part of their standard data set. AREMP included an analysis of the average wetted width in their 25-year report.

**Questions:**

* Can PIBO and AREMP share there metrics as part of their standard data set?
* Should we include wetted width to depth ratio in the first draft of the controlled vocabulary and crosswalk? BLM AIM would have to assume a trapezoid channel because they only take thalweg depth at cross-sections. We also need to assess if the analysis methods are the same, average of the sums or the sum of the averages.

**Conclusion:**

* Average wetted width data collation and analysis is the same across all four programs, include this metric
* Wetted with to depth ratio, we need to understand the analysis before we can say it is comparable

**For Data Exchange Specifications & Data Discovery Tools:**

* Include average wetted width in the first draft

**Controlled Vocabulary**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LongName** | **Definition** | **DataType** | **NotesCodesConventions** | **Unit** | **Field** |
| Average wetted width from transects | Average wetted width across transects. | Double |  | m | WetWidth |
| Wetted width to depth ratio at transects | Mean Wetted Width/Thalweg Depth ratio | Double |  | ratio | WetWidthToDepth |

**Crosswalk**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Field** | **USFS\_AREMP\_Field** | **USFS\_AREMP\_Method\_Comment** | **BLM\_AIM\_Field** | **BLM\_AIM\_Method\_Comment** | **EPA\_NARS\_Field** | **EPA\_NARS\_Method\_Comment** | **USFS\_PIBO\_Field** | **USFS\_PIBO\_Method\_Comment** |
| WetWidth | wettedWidth\_ave | This isn't in our public-facing data, but we do have it. | WTTD\_WT |  | XWIDTH |  |  |  |
| WetWidthToDepth |  | we have the source data to calculate this, but it would take a little effort |  |  | XWD\_RAT |  | WDWetTrans |  |

Pools- Residual pool depth (RPD) & Percent Pools

**Discussion:**All four programs have residual pool dept (RPD) and percent pool values. AREMP does not share them as part of their standard data set.  BLM AIM, USFS AREMP and PIBO all use the same data collection method for pools, the Lisle 1987 method.

The methods for defining pools:

* BLM AIM: <https://www.monitoringresources.org/Document/Method/Details/6859>
* USFS AREMP: <https://www.monitoringresources.org/Document/Method/Details/6844>
* USFS PIBO: <https://www.monitoringresources.org/Document/Method/Details/6872>

Although field methods are compatible, Roper et al. 2010 showed training and crews impact pool measurements, so we should expect to see some variability within and across programs. EPA NARS does not use the same collection method for polls. Therefore, the EPA NARS is likely not comparable, but BLM AIM collects pool data using the Lisel 1987 and the EPA NARS methods, so we could analyze how the data compares (put this analysis on hold due to BLM AIM data restructuring).

EPA NARS residual depth metric RP100 is not directly comparable to the RPD from the other programs.  RP100 is mean residual depth in cm, which is the VERTICAL (like sagittal) cross-section area of all residual pools in the sample reach divided by the length of the sample reach (see Figure 4 below from Kaufmann et al. 1999).  The residual pool variables for the EPA surveys are all calculated from the thalweg depth profile (combined with reach gradient), which allows a great deal of flexibility in the derivation of pool metrics to fit different objectives.  The closest metric EPA NARS has to the definition in the controlled vocabulary is RPGT10X, the mean depth of all residual pools in the sample reach which have residual depth greater than 10cm.

Again, pool identification is highly variable based on training and crew, so BLM AIM, USFS PIBO and AREMP have the same field mythology, yet we expect this metric to be highly variable.

**Questions:**

* Are the EPA NARS values comparable to the other three methods? BLM AIM does the data collection using the EPA NARS method and the Lidel method. What is the comparison? (On hold while BLM AIM restructures data system)
* EPA NARS suggests we could investigate other reach metrics like RPGT75, which is the number of pools in the sample reach with residual depth 75cm or counts of other pool depths like RPGT50, RPGT20, etc. To compare EPA NARS RPD values to other RPD values, additional analysis would need to be analyzed to normalize by reach length. EPA NARS also calculate metrics from the maximum residual depth of the deepest pool in the reach RPMDEP and we could think of ways to calculate this value of the other programs. In the future would this group be interested in producing and sharing any of there metrics?
* What other metrics are essential for users to understand the variability in pools? Percent of the reach that was dry? Gradient? Beaver present? Count of beaver pools?

**Conclusion:**

* RPD and Percent Pool for BLM AIM, USFS AREMP and PIBO are comparable metrics and should be included. Still, we also need to have a way to indicate the metrics are subjective, variable on training, crew, and flow stage.
* The status of pools is important, but due to pool metrics' subjectivity, it would be beneficial to find a way to share pool metrics' trends over time instead of just the status of RPD and Percent Pool

**For Data Exchange Specifications & Data Discovery Tools:**

* Need a way to indicate to the user that although the data collection methodology is compatible between BLM AIM, USFS AREMP and PIBO, results might vary given differences in training, flow stage, stream type.
* For metrics with high signal to noise, it might be more useful if we focus on building an analysis on top of the data to share trends and not just status.

**Controlled Vocabulary**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LongName** | **Definition** | **DataType** | **NotesCodesConventions** | **Unit** | **Field** |
| 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. | Double |  | m | RPD |
| Percent pools | Percent of the sample reach (linear extent) classified as pool habitat | Double | min: 0, max: 100 | % | PctPool |

**Crosswalk**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Field** | **USFS\_AREMP\_Field** | **USFS\_AREMP\_Method\_Comment** | **BLM\_AIM\_Field** | **BLM\_AIM\_Method\_Comment** | **EPA\_NARS\_Field** | **EPA\_NARS\_Method\_Comment** | **USFS\_PIBO\_Field** | **USFS\_PIBO\_Method\_Comment** |
| RPD | ave\_residual\_depth | This isn't in our public-facing data, but we do have it. | RES\_PL\_DEP |  |  | Residual Pools:  I believe that RP100 is the best all-around pool metric. However, it is not directly comparable to the residual pool metrics from the other programs.  The spreadsheet refers to EPA’s residual depth metric RP100 as the Average of the residual pool depth values for all pools in a reach. Residual pool depth [is} the difference of the pool tail depth from the max depth.”  This is not correct for RP100, is mean residual depth in cm, which is the VERTICAL (like sagittal) cross-section area of all residual pools  in the sample reach divided by the length of the sample reach  (see Figure 4 below from Kaufmann et al. 1999)  --- Perhaps we might talk on phone about other residual pool variables.  The residual pool variables for the EPA surveys are all calculated from the thalweg depth profile (combined with reach gradient), which allows a great deal of flexibility in the derivation of pool metrics to fit different objectives.  The closest to the definition in the spreadsheet would be RPGT10X, the mean depth of all residual pools in the sample reach which have residual depth greater than 10cm.  RPGT10X is one of a series of mean residual pool depth variables (RPXDEP, RPGT05X, RPGT10X, RPGT20X), which are the mean residual depths for progressively larger depth classes of residual pools ranging from all residual pool features up to all with maximum residual depth >20cm.  You may want to look at others, like RPGT75 which is the number of pools in the sample reach with residual depth 75cm or counts of other pool depths like RPGT50, RPGT20, etc. You would need to divide these numbers by REACHLEN to and multiply by to express them in per 100 meter basis.  We also calculate metrics form the maximum residual depth of the deepest pool in the reach RPMDEP. | PoolDp |  |
| PctPool | PctReachInPools | This isn't in our public-facing data, but we do have it. | PCT\_PL |  |  | PCT\_POOL it the tilte of this metric, but given the | PoolPct |  |

Substrate

**Discussion:**All four programs collect partials using similar mythology with subtle but important differences. 

* BLM AIM - 10 at each of the 20 transact, a total of 200 particles, scour line to scour line
* EPA NARS - 5 at each of the 21 transects, a total of 105 particles, in the wetted channel
* PIBO - 5 at each of the 21 transects, a total of 105 particles, bankfull to bankfull, but in the analysis, particles above scour line are excluded
* AREMP - 5 at each of the 21 transects, a total of 105 particles, bankfull to bankfull

Bunte et al. 2009 demonstrated results differ based on where channel particles are collected, bankfull to bankfull vs. collection in the wetted width. We wanted to understand if there is a difference between particles collected scour to scour line vs. contained within the wetted channel. The BLM AIM labels particles according to channel location, above or below the waterline. They volunteered to run an analysis to see the impact of where the particles are collected wetted vs active channel. Based on the BLM AIM analysis, the group decided it was appropriate to summarize the particles from the active and wetted channel into one metric. The group also indicated that the crew effect might result in a larger deviation than the outliers identified in the BLM AIM analysis of active vs. wetted channels (Roper et al. 2010).

In addition to the metrics proposed in the initial version of the controlled vocabulary and crosswalk (<2mm, <6mm and D50), the group decided to include bed surface D16, D84, and percent of stream bed area bedrock (>4098). The field and analysis methods to produce these metrics are the same between BLM AIM, USFS AREMP and PIBO.

**Questions:**

* USFS AREMP, are the particles above the scour line included in the analysis? Or is it just scour line to scour line?
* Is there is a difference in particles collected in the wetted vs. the active channel?
* Answer: Based on the quick analysis and the experts opinion in the large majority of cases, we can combine the active and wetted channel data. BLM AIM collects data in the active channel, labeling the particles that are collected in the wetted channel. BLM AIM completed a quick analysis and found there is no difference in results between   particles collected in the active vs. wetted channel. Expertise indicated that there may be a bigger deviation in results due to crew effect then other outliers in the charts below.



**Conclusion:**

* BLM AIM and USFS PIBO can combine data for these metrics. If USFS AREMP summarized bed particles from scour line to scour line include their metric.

**For Data Exchange Specifications & Data Discovery Tools:**

* Include <2 mm, <6mm, D16, D50, D84, and % stream bed area bedrock

**Controlled Vocabulary**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LongName** | **Definition** | **DataType** | **NotesCodesConventions** | **Unit** | **Field** |
| Diameter of the 50th percentile streambed particle | Bed surface particle size corresponding to the areal median (50th percentile of measured particles) | Double | min: 1, max: 4098 | mm | D50 |
| Diameter of the 16th percentile streambed particle | Bed surface particle size corresponding to the 16th percentile of measured particles | Double | min: 1, max: 4098 | mm | D16 |
| Diameter of the 84th percentile streambed particle | Bed surface particle size corresponding to the 84th percentile of measured particles | Double | min: 1, max: 4098 | mm | D84 |
| Percent Bed Surface Bedrock | Percent of the stream bed area that is composed of particles with a b-axis > 4098mm (bedrock) | Double | min: 0, max: 100 | % | PctBdrk |

**Crosswalk**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Field** | **USFS\_AREMP\_Field** | **USFS\_AREMP\_Method\_Comment** | **BLM\_AIM\_Field** | **BLM\_AIM\_Method\_Comment** | **EPA\_NARS\_Field** | **EPA\_NARS\_Method\_Comment** | **USFS\_PIBO\_Field** | **USFS\_PIBO\_Method\_Comment** |
| D50 | D50 | This isn't in our public-facing data, but we do have it. | D50 |  | Sub\_dmm | Bed particle D50.  It should specify that this is median diameter of Bed Surface particles, and it refers to an areal median.  That said, a D50 based on 105 particles is not a very stable statistic.  The best variable to match the D50 of the other projects is Sub\_dmm, the geometric mean surficial bed particle diameter (mm), or its Log10 transformation (LSub\_dmm).  Faustini and Kaufmann (2007) compare the accuracy and precision of various metrics from field survey protocols. Full citation is Faustini, J.M. and P.R. Kaufmann. 2007. Adequacy of Visually Classified Particle Count Statistics from Regional Stream Habitat Surveys. J. Am. Water Resourc. Assoc. 43(5):1293-1315. | D50 |  |
| D16 |  | AREMP can calculate and share | D16 |  |  |  | D16 |  |
| D84 |  | AREMP can calculate and share | D84 |  |  |  | D84 |  |
| PctBdrk | bedrock\_pcnt | AREMP can calculate and share |  |  | PCT\_BDRK |  |  |  |

Substrate- Pool Tail Fines

**Discussion:**While discussing substrates, the group decided to add pool tail fines measurements to the controlled vocabulary and crosswalk. The BLM AIM and USFS PIBO use the same method to define pools and collect pool tail fines (<https://www.monitoringresources.org/Document/Method/Details/6862>), AREMP describes pools the same but counts pool tail fines with a slightly different approach: <https://www.monitoringresources.org/Document/Method/Details/6798>. USFS AREMP counts pool fines at all pools in the reach and only counts <2mm particles; in contrast, BLM AIM and USFS PIBO count at the first ten scour and plunge pools < 2mm and <6 mm. So, for three of the four programs, the <2mm is comparable, and BLM AIM and USFS PIBO <6 mm is equal for pool tail fines.

**Conclusion:**

* We can include pool tail fines

**For Data Exchange Specifications & Data Discovery Tools:**

* For <2mm particles in pool tail fines USFS AREMP, PIBO and BLM AIM are comparable
* For <6mm particles in pool tail fines BLM AIM and USFS PIBO are comparable

**Controlled Vocabulary**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LongName** | **Definition** | **DataType** | **NotesCodesConventions** | **Unit** | **Field** |
| Percent of streambed particles <2mm | Percent of the stream bed area that is comprised of practices with a b-axis < 2 mm | Double | min: 0, max: 100 | % | PctFines2 |
| Percent of streambed particles <6mm | Percent of the stream bed area that is composed of particles with a b-axis < 6 mm | Double | min: 0, max: 100 | % | PctFines6 |

**Crosswalk**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Field** | **USFS\_AREMP\_Field** | **USFS\_AREMP\_Method\_Comment** | **BLM\_AIM\_Field** | **BLM\_AIM\_Method\_Comment** | **EPA\_NARS\_Field** | **EPA\_NARS\_Method\_Comment** | **USFS\_PIBO\_Field** | **USFS\_PIBO\_Method\_Comment** |
| PctFines2 | pcnt\_fines\_tran2 | This isn't in our public-facing data, but we do have it. | PCT\_FN |  | PCT\_SAFN |  |  |  |
| PctFines6 | pect\_fines\_tran6 | This isn't in our public-facing data, but we do have it. | PCT\_FN6 |  | PCT\_FN |  |  |  |

Wood

**Discussion:** Each program collects measurements related to wood and produces metrics. The methodologies for measuring wood differ based on the program and relate to the environmental context and underlying biological question each program is designed to answer. For example, in the Pacific Northwest Coast Range, where forest cover is abundant, the minimum size of large wood collected by monitoring programs is more significant than in the Upper Columbia River region. As a group, we agree that wood is an important metric to answer multiple types of questions as a covariate and trend analysis. We were interested in finding a way to share the data. We agreed we cannot combine metrics across the four programs or create a new analysis method of increasing the minimum wood size to match the largest minimum wood size of the four programs. A new metric calculated across all four programs would not provide the scale of information necessary to answer specific biological questions in regions with low wood recruitment potential.

**Question:**

* How can we direct users to origin data sets to get wood metrics with enough metadata and caveats to understand that the values are not comparable across programs, environmental gradients, or biological questions?

**Conclusion:**

* Wood is an important value for data users. The amount of available wood depends on the geographic and physiographic variables within the region, i.e., for areas with denser forest cover, you expect to see more instream wood (and different size classes of wood) and for areas with less forest cover, you'd expect to see less instream wood.
* We could not compare the metric values even if you created a new, unifying metric across these monitoring programs' environmental spectrum.
* Give the above conclusion our goal for this work is to identify where large wood data are available from these programs, specific caveats, and provide a sufficient amount of metadata for users to understand the context in which the large wood data was collected, the collection methods,  and the value produced.

**For Data Exchange Specifications & Data Discovery Tools:**

* We cannot aggregate large wood values from the four programs into one column in the data set.
* Associate large wood values with enough metadata describing the environmental context and the collection process.
* We need to identify a way to indicate to users that data exists, but it is not comparable across the programs.
* We need to find a way to allow users to access the origin data sets.

Macroinvertebrates – O/E & MMI

**Discussion:** Macroinvertebrate data comparability, like all other metrics, discussed, depend on the field collection methodology and lab identification methods. We have identified that the EPA NARS, BLM AIM, USFS AREMP and PIBO have similar field methods to combine macroinvertebrate data. They all sample using an area method, not an specified lent of time. The controlled vocabulary for this metric will only be valid for data collected using the area method.  Our initial suggestion was to share O/E and MMI index, but we realized we could not combine the data into one column of a data set because each program calculates O/E and MMI based on program-specific models. We discussed using the BLM AIM west-wide comprehensive model for calculating O/E and MMI. BLM AIM has already completed most of this work and volunteered to run the model on others’ data sets. Our data-sharing effort would then shift focus to sharing the 10-12 metrics needed in the west-wide model. In a combined data set, we could share the 10-12 metrics without the west-wide or program O/E and MMI values and create an analysis package to let data users compute the indicator values. This approach would require significant investment in building an R or Python package and adequate metadata to accomplish that work. Additionally, it would make the data less approachable by certain groups of users, and one of our key goals was to promote data discovery and reuse of these robust long-term data sets. We also discussed sharing the "raw data".

**Question:**

* Macroinvertebrates inform Clean Water Act reporting, by sharing these data raw, metric, or indicators, are we potentially opening the agencies up to legal issues?

**Conclusion:**

* Sharing raw macroinvertebrate data, metrics, or indicators depends on the field method and lab methods, including sub-sampling the sample and taxonomic list and resolution used to identify specimens.
* We will only share data that have used collection methods that sample a specified area and not methods that sample for specified length of time
* Sharing O/E and MMI as produced by each program could potentially present issues because, for the same or similar locations, indicators values could contradict each due to differences in the underlying model. Adding the west-wide model could create additional complications and contradictions.
* Macroinvertebrate data are used in work related to the Clean Water Act and misinterpretation could open agencies to lawsuits, so anything we share must contain sufficient metadata to be defensible
* Sharing "raw data" as a species list could be problematic because we are publishing a list that may contain rare species, again opening the agencies to legal issues and potential data-sharing issues related to revealing specific locations of rare species

**For Data Exchange Specifications & Data Discovery Tools:**

* We will start by sharing a sub-set of macroinvertebrates metrics standardized to a species list. BLM AIM, USFS PIBO and AREMP samples are all processed at the Utah State University (USU) Bug Lab. Meaning the samples are sub-set the same way and processed to the same taxonomic level. The lab might also have the EPA NARS data.
* We will then do outreach and user testing to understand the use cases for this type of data. This will help us understand if there is a need to use the data in regulatory actions, and whether we should share the modeled data.
* No matter what we end up sharing or not sharing, sufficient metadata are needed. If we share O/E and MMI we need to identify the models used and provide documentation on why indicator values in close spatial proximity or at the same location might differ based on the model used.
* Data analysis process matters! We need to understand the taxonomy and random subsampling procedure used on all biological samples to create standardized data.

**Controlled Vocabulary**

In process

**Crosswalk**

In process

Temperature

**Discussion:**BLM AIM and EPA NARS collect instantaneous temperature data, USFS PIBO and AREMP collect continuous summer temperature data. USFS AREMP only takes one temperature record at one site in a watershed. USFS PIBO is moving away from summer temperature data and towards year-round temperatures measurements. BLM AIM extracts average August stream temperature data from the NorWeST data set for each data collection point. USFS PIBO compared collected summer temperature values to the metrics from the NorWeST data set, found the differences between measured and predicted to be low and now planning on using the NorWeST temperature model for summer temperatures. Given this information and the program’s shift away from collecting summer temperatures, we will use the NorWeST temperature model to attribute each data collection location with temperature data. Brett recommends research metrics in Isaak et al. 2019 paper on thermal regimes.

**Conclusion:**

* Temperature is important, and the model values will provide a value for all data collection locations.

**For Data Exchange Specifications & Data Discovery Tools:**

* Use the NorWeST temperature model to attribute each data collection location with temperature metrics
* Review the Isaak et al. 2019 paper to figure out which metrics to include in the integrated data set

Water Quality

**Discussion:**USFS AREMP, BLM AIM and EPA NARS take water quality grab samples. BLM AIM and USFS AREMP process the samples in the same lab at USU and produce the same metrics. NEED TO UNDERSTAND MORE ABOUT EPA NARS PROCESSING. USFS PIBO collects conductivity (total dissolved solids), but no grab samples.

Infrequent grab samples for water quality data are not enough to make Clean Water Act decisions, but they do provide valuable information. For BLM AIM and USFS AREMP, the protocols are simple and the same and analysis is done in the same lab at USU therefore we decided to report total nitrogen, total phosphorous, specific conductance, pH and turbidity. NEED TO TALK TO EPA about process.

In addition to totals BLM AIM produces predicated values for reference nitrogen, phosphorus and specific conductance concentrations at each site based on Olson and Hawkins 2012.  Additional analysis work would need to be done to run the USFS AREMP data through the predicted models. Jennifer sent the papers along and we can discuss if we want to calculate those values for the USFS AREMP data. PIBO measures total dissolved solids, so that can also be included.

**Conclusion**

* Grab samples have limitations, but we should share this data
* As we start thinking about the way we deliver the integrated data set, we can think about bringing in water quality data from the WQX to add context to these data collection locations and provide more data values for different types of users
* In the future we can have discussions about the value of calculated the predicted values for nitrogen, phosphorus and specific conductance concentrations.

**For Data Exchange Specifications & Data Discovery Tools:**

* Share the USFS AREMP and BLM AIM water quality values in the integrated data set

**Controlled Vocabulary**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LongName** | **Definition** | **DataType** | **NotesCodesConventions** | **Unit** | **Field** |
| 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. | Double |  | ppm | Conductivity |
| Total Nitrogen | Measured total nitrogen value | Double |  | ug/L | TotalNitrogen |
| Total Phosphorous | Measured total phosphorous value | Double |  | ug/L | TotalPhosphorous |
| Specific Conductance | Measured specific conductance value. The specific conductance is conductivity standardized to 25 degrees C. | Double | min: 0, max: 65500 | uS/cm | SpecificConductance |
| Turbidity | Average water clarity as measured by the suspended solids in the water column | Double |  | NTU | Turbidity |

**Crosswalk**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Field** | **USFS\_AREMP\_Field** | **USFS\_AREMP\_Method\_Comment** | **BLM\_AIM\_Field** | **BLM\_AIM\_Method\_Comment** | **EPA\_NARS\_Field** | **EPA\_NARS\_Method\_Comment** | **USFS\_PIBO\_Field** | **USFS\_PIBO\_Method\_Comment** |
| Conductivity |  |  |  |  | TSS |  | Cond |  |
| TotalNitrogen |  |  | TotalNitrogen |  | NTL |  |  |  |
| TotalPhosphorous |  |  | TotalPhosphorous |  | PTL |  |  |  |
| SpecificConductance |  |  | SpecificConductance |  | COND |  |  |  |
| Turbidity |  |  | Turbidity |  | TURB |  |  |  |

Bank Measurements

**Discussion:**For bank measurements, BLM AIM and USFS PIBO have comparable collection methods, EPA NARS methods are very different, and USFS AREMP does not collect bank measurements because there is no grazing in their study area. BLM AIM and USFS PIBO calculate bank stability slightly differently, so we need to discuss the metric produced. BLM AIM also derives cover values, but they plan to stop producing those values and releasing the raw data from those cover measurements, so there is no comparable cover metric at this time.

**Conclusion**

* We will include bank angle and, where we can, bank stably in the controlled vocabulary and crosswalk.

**For Data Exchange Specifications & Data Discovery Tools:**

* Bank angle will be included for USFS PIBO and BLM AIM
* BLM AIM and USFS PIBO need to have a discussion on how to they calculate the bank stability

**Controlled Vocabulary**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LongName** | **Definition** | **DataType** | **NotesCodesConventions** | **Unit** | **Field** |
| Bank angle | Measured angle of the stream bank; | Double | min: 0, max: 180 | degrees | BankAngle |
| 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) | Double | min: 0, max: 100 | % | PctStab |

**Crosswalk**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Field** | **USFS\_AREMP\_Field** | **USFS\_AREMP\_Method\_Comment** | **BLM\_AIM\_Field** | **BLM\_AIM\_Method\_Comment** | **EPA\_NARS\_Field** | **EPA\_NARS\_Method\_Comment** | **USFS\_PIBO\_Field** | **USFS\_PIBO\_Method\_Comment** |
| BankAngle |  |  | BNK\_AN |  | XBKA |  | BankAngle |  |
| PctStab |  |  | BankStability |  |  |  | Stab |  |