Contents

[Introduction 1](#_Toc41991934)

[Habitat Data Exchange Specifications 3](#_Toc41991935)

[Conclusion 11](#_Toc41991936)

[References 12](#_Toc41991937)

[Appendix A 13](#_Toc41991938)

# Introduction

Streams are critical to fish, water quality, 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. Due to climate change, urbanization, and multi-use land management, there is a need to determine the resources' quality and trends across jurisdictional boundaries by using information from multiple collection efforts. It is not straightforward to combine data from various monitoring programs due to differences in response and spatial design. Additionally, data produced by these programs are not always findable, accessible, interoperable, and reusable (FAIR) (Wilkinson, 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). We establish rules for integrating and sharing stream data from multiple sources to decrease the time spent finding and organizing data to accurate and timely information for building indicators, completing analysis, discovering data, and making decisions. We reviewed the response design, specifically the metrics from four federal stream habitat monitoring programs, to build rules for combining metrics. We design workflows with human and machine-readable data and metadata to make the data FAIR. We know that the spatial design will impact data integration and the data use cases, but at this time, we don’t address that issue.

We base the data exchange specification for metric-level stream habitat data on the Darwin Core principles. Darwin Core provides terms and schema for integrating and exchanging biodiversity information (Wieczorek, 2012). Darwin Core is adaptable, we adapt Darwin Core Location, Event, and Measurements of Fact extensions to exchange and combine stream habitat information. Darwin Core's star schema allows for one to many relationships between elements, consistent with stream habitat data structure, where a single event produces multiple reach level metrics (Figure 1). The data govern the fields included. Fields refer to the column headers in the dataset. Darwin Core recommends the use of standard vocabulary for fields to create homogeneity in the data. (Wieczorek, 2012)

To adapt Darwin Core to stream habitat information, we use data and metadata from four large scale federal 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. We define the data exchange fields, a controlled vocabulary for metrics, and a crosswalk from the program's metadata to the controlled vocabulary. Using this documentation, we define a workflow and write R code for integrating the data from multiple sources.

We create a controlled vocabulary to define a subset of standard metrics across the four programs. 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 a subset of metrics is created across programs. Although the programs report some of the same metrics, values can differ across programs due to variability in sampling protocols and training (Al-Chokhachy and Roper 2010). To establish this standard vocabulary of metrics, we use literature and expert opinion.

We then cross-walked each program's field name from the metadata to the standard vocabulary. Appendix A details the metric combability between programs based on literature and input from the programs' staff. The burden is on the data users to decide if the methods are comparable enough to answer their specific management questions based on metadata.

Robust machine and human-readable documentation of the step-by-step process used to collect they analyzed data is required to assess metrics' compatibility from multiple stream habitat monitoring programs. We document data collection and analysis in MonitoringResources.org. MonitoringResources.org supports the need to describe and document projects throughout the data management lifecycle by providing a structured, publicly accessible resource for recording project metadata. The tools offer access to monitoring project data collection and analysis methods using a Web Browser and application program interface (API).

To create one data set out of

In addition to robust metadata documentation for data collection and analysis, metadata about data sets, such as where the data sets are stored, how often they are updated

datasets need to be in an accessible data repository. Each of the four programs’ release data according to their own set of requirements. We want this intergrated data set to be accessable and well documented so we used ScienceBase to document and share datasets in a human and machine-readable way. ScienceBase is an integrated data-sharing platform managed by the U.S. Geological Survey. It provides access to scientific data and other information using a Web Browser, and application program interface, and representational state transfer (REST)-based Web Services (Chase 2016).

Using these data exchange specifications, MonitoringResources.org, and ScienceBase, we can make instream habitat data findable, accessible, interoperable, and reusable (FAIR). Efficiently delivering stream habitat metrics across jurisdictional boundaries to decision-makers.

# Habitat Data Exchange Specifications

The stream habitat data exchange specifications is based on the simple Darwin Core standard. Using data from four programs, we outline the elements needed to create an integrated dataset, including core dataset information, location information, data collection event information, and individual metrics data. This schema is suited for stream habitat data because at one location, multiple events occur, and programs collect numerous measurements at a single event (Figure 1). The approach is to draft a schema to exchange metric level habitat data (HOW? Via machine-readable automated or something else processes?) that is flexible enough to be adapted later to also support the transfer of other data types such as macroinvertebrate counts or riparian vegetation.

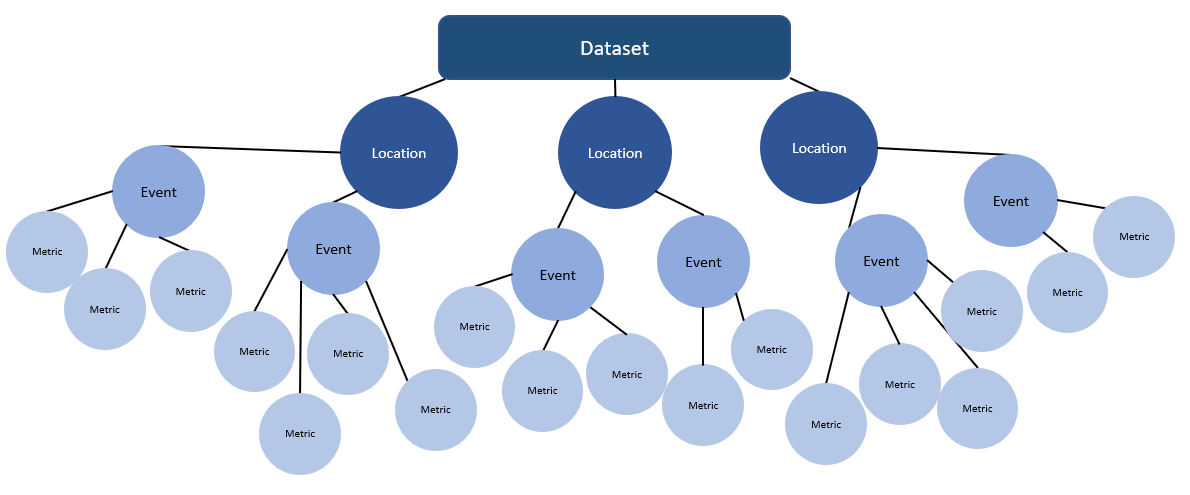


Figure 1 The one to may relationship between datasets, locations, events, and metrics based on the Darwin core principles modified to work with stream habitat data.

#### Core

The core table describes documentation of the core elements of a dataset, including information about the origin of the dataset, who collected the data, and how to cite the dataset (Table 1). A dataset is a collection of data collection events; for example, a program releases release a dataset every five years containing all the data collection events that occurred in the previous five years. For data to be accessible, it should be stored it in a trusted online data repository. 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.

|  |  |  |  |
| --- | --- | --- | --- |
| Core | | | |
| Enter information about the dataset, the data provider and the data collector | | | |
| **Field Name** | **Field Description** | **Data Type** | **Notes, Codes/Conventions** |
| type |  |  | EventMeasurments for all entries |
| modified | The most recent date-time on which the combined dataset was changed. | Date | Recommended best practice is to use a date that conforms to ISO 8601:2004(E). |
| rightsHolder | A person or organization owning or managing rights over the resource. | Text | Example: BLM AIM, EPA NARS, USFS ARMEP |
| bibilographicCititation | A bibliographic reference for the resource as a statement indicating how this record should be cited (attributed) when used. | Text |  |
| InstitutionID | An identifier for the institution having custody of the object(s) or information referred to in the record. | Text | Example) http://biocol.org/urn:lsid:biocol.org:col:34777, http://grbio.org/cool/km06-gtbn |
| CollectionID | An identifier for the collection or dataset from which the record was derived. | Text | Example) https://www.epa.gov/national-aquatic-resource-surveys |
| datasetID | An identifier for the set of data. May be a global unique identifier or an identifier specific to a collection or institution. | Integer (need to check this) | The ScienceBase ID number of the dataset |
| datasetName | The name identifying the data set from which the record was derived. | Text | Example: EPA NARS 2018 dataset |
| institutionCode | The name (or acronym) in use by the institution having custody of the object(s) or information referred to in the record | Text | Example: AIM, NARS, AREMP, PIBO |

#### Location

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

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

|  |  |  |  |
| --- | --- | --- | --- |
| Location | | | |
| A spatial region or named place. | | | |
| **Field Name** | **Field Description** | **Data Type** | **Notes, Codes/Conventions** |
| verbatimLocation | Unique number that identifies one sample of a particular site. | Text | In habitat datasets 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. |
| 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 | Text |  |
| 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 | Stream Name from the provider dataseet | Text | The water body name from the original data set |
| verbatimCoordinateSystem | The spatial coordinate system for the verbatimLatitude and verbatimLongitude or the verbatimCoordinates of the Location. | Text |  |
| State | The name of the next smaller administrative region than country (state, province, canton, department, region, etc.) in which the Location occurs. | Text |  |
| 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. | Text |  |

#### Event

The Event table describes an action that occurs at a location during a specific time frame (Table 3). The programs used to online these data exchange specifications implement a rotating panel temporal design, meaning that the project returns to a single location multiple times during the duration of the study. Allowing for the calculation of trends in metrics over time in response to changing management actions. Therefore a dataset will contain multiple 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. | |  |  |
| **Field Name** | **Field Description** | **Data Type** | **Notes, Codes/Conventions** |
| 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. | Text | 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. |
| locationID | Number that identifies a unique sampling location. A site is a stream segment with a fixed starting and ending location for sampling | Text |  |
| year | The four-digit year in which the Event occurred, according to the Common Era Calendar. | Interger |  |
| verbatimEventDate | The verbatim original representation of the date and time information for an Event. | Date |  |
| samplingProtocol | The name of, reference to, or description of the method or protocol used during an Event. MonitoringResources.org protocol ID number | Integer |  |
| 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. | Text |  |
| Date | 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. | Date |  |
| Time | The time or interval during which an Event occurred. | Date |  |

#### Measurement of 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. A variety of metrics can be derived 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 both calculated from pool measurements.

It is necessary to accompany each metric value with three pieces of information: a description of how the field measurement is collected, a description of how the field values are analyzed to result in metrics, and additional remarks (Figure 2, Table 4 ). This documentation of the collection and analysis methods was recorded in MonitoringResources.org.

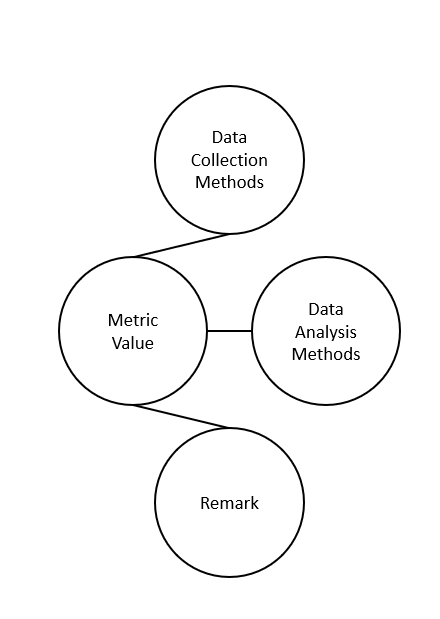


Figure 2. Structure of the Metric level data. Metrics definitions are documented in the controlled vocabulary. The field naming convention is metric title continuation with the field name.

Table 4. Metric Table

|  |  |  |  |
| --- | --- | --- | --- |
| Measurement or Fact | | | |
| A measurement of or fact or metric about a resource | |  |  |
| **Field Name** | **Field Description** | **Data Type** | **Notes, Codes/Conventions** |
| <MetricName> | The value of the measurement, fact, characteristic, or assertion. | Depends on the | From the controlled vocabulary |
| <MetricName>Collection | A description of or reference to (publication, URI) the method or protocol used to determine the measurement, fact, characteristic, or assertion. | Number | MethodID from MonitoringResources.org |
| <MetricName>Analysis | A description of or reference to (publication, URI) the method or protocol used to calculate the measurement, fact, characteristic, or assertion. | Number | MethodID from MonitoringResources.org |
| <MetricName>MetricRemark | A remark related to the data from the data provider. | text |  |

The field (column header) for each metric is a continuation of the metric name and the field value, so for example, for residual pool depth, the field name is RPD. Therefore, the additional fields will be RPDCollection, RPDAnalysis, RPDMetricRemark (Table 5).

Table 5 Example of how to name fields for metrics in the exchanged data.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **RPD** | **RPDCollection** | **RPDAnalysis** | **RPDMetricRemark** | **PctPool** | **PctPoolCollection** | **PctPoolAnalysis** | **PctPoolMetricRemark** |
| **16.75** | **1542** | **755** |  | **100** | **1542** | **1027** | **Reach on beaver pool** |
| **0.11676175** | **1542** | **755** |  | **1.33333333** | **1542** | **1027** |  |

To promote homogeneity across the data, we built a data dictionary to enable standraization of multiple program's metrics. We survived the existing metadata from the four habiat and incoperated already existing metrics.

standard set of metrics for exchange based on the datasets from the four habitat monitoring programs, BLM AIM, EPA NARS, USFS PIBO, and USFS ARMEP (Table 6). The controlled vocabulary contains a subset of the metrics calculated among the three programs. In Appendix A we document analysis of metric combability between programs based on literature and input from the programs' staff. Ultimately the data will decide on metric compatibility to answer their management question based on the metadata from MonitoringResources.org and Appendix A.

If partners wish to exchange additional metrics, updated the controlled vocabulary needs to be with the name, description, short machine-readable name, data type, and unit. We use GitHub to manage version control of the habitat metric vocabulary. To use the controlled vocabulary, a new program must cross-walk the fields in their datasets fields to the dictionary and document their data collection and analysis methods in MonitoringResources.org. The pilot program's field cross-walk and metrics IDs are shared in ScienceBase OR GITHUB I DON'T KNOW WHERE YET??

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.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Category** | **LongName** | **Field** | **Definition** | **DataType** | **Notes, Codes/Conventions** |
| Channel dimensions | Average bankfull width from transects | BFWidth | Average bankfull width across transects. | Double |  |
| Channel dimensions | Gradient of stream reach | Grad | Mean Slope of water surface (%) from the bottom of the reach to the top of the reach. | Double |  |
| Channel dimensions | Length of sampling reach | RchLen | Length of sampling reach measured along the thalweg | Double |  |
| Channel dimensions | Bankfull width-to-depth ratio at transects | BFWDRatio | Average Bankfull Width/Depth Ratio across transects. | Double |  |
| Channel dimensions | Bankfull Height | BFHeight | Average bankfull height measured from water surface across transects. | Double |  |
| Channel dimensions | Average wetted width from transects | WetWidth | Average wetted width across transects. | Double |  |
| Channel dimensions | Wetted width-to-depth ratio at transects | WetWidthToDepth | Mean Wetted Width/Thalweg Depth ratio | Double |  |
| Pools | Residual pool depth | RPD | 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 |  |
| Pools | Percent pools | PctPool | Percent of the sample reach (linear extent) classified as pool habitat | Double |  |
| Channel Characteristics | Sinuosity of Local Stream Reach | Sin | Reach sinuosity (reach length along the thalweg divided by straight line distance between the bottom of the reach and the top of the reach) | Double |  |
| Channel Characteristics | Percent of Reach that is Dry | PctDry | Percent of the reach that was dry | Double |  |
| Channel Characteristics | Beaver Sign at Reach | Beaver | Beaver value from the provider dataset. | Text |  |
| Channel Characteristics | Stream Order | StreamOrder | Strahler stream order of the site from the provider dataset | Double |  |
| Substrate | Diameter of the 50th percentile streambed particle | D50 | Particle size corresponding to the 50th percentile of measured particles | Double |  |
| Substrate | Percent pool tail fines < 2mm | PoolTailFines2 | Average percent fine sediment (< 2mm) on the pool tail | Double |  |
| Substrate | Percent of streambed particles <2mm | PctFines2 | Percent of particles counted with a b-axis < 2 mm | Double |  |
| Substrate | Percent of streambed particles <6mm | PctFines6 | Percent of particles with a b-axis < 6 mm | Double |  |
| Substrate | Percent pool tail fines < 6mm | PoolTailFines6 | Average percent fine sediment (< 6mm) on the pool tail | Double |  |
| Wood | Large wood frequency | LWDFreq | Frequency of large woody debris within the bankfull channel of the reach. | Double |  |
| Wood | Large wood volume | LWDVol | Volume of LWD within the bankfull channel of the reach | Double |  |
| Macroinvertebrates | O/E Macroinvertebrate Index | OEratio | Biological condition was assessed using an observed/expected (O/E) index. | Double |  |
| Macroinvertebrates | Multimetric Macroinvertebrate Index | MMI | Biological condition was assessed using the MMI (MultimetricIndex) | Double |  |
| Channel Characteristics | Beaver Present | BeaverPresent | Indication if beaver's have disturbed the flow of the stream in the reach. | Categorical |  |
|  |  |  |  |  |  |

# Conclusion

The data exchange specifications contains the details of what will be share and the format to be shared. To share information data managers need to:

1. make datasets publicly accessible in their repository or a persistent data repository such as ScienceBase,
2. document collection and analysis methods in MonitoringResources.org,
3. map datasets fields to the DCS,
4. write and publish code convert the dataset to the DCS format and the correct spatial format
5. THEN THERE IS A STEP TO INTERGEATRE THE DATA EVEN AFTER IT IS IN THE CORRECT FORMAT I DON" T KNOW HOW THAT IS DONE?

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 is required.

Reach Characteristics :

**Reach length (m) -** Length of sampling reach measured along the thalweg

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | AIM | AREMP | NARS | PIBO |
| Collection Method Link |  |  |  |  |

**Slope (%)-** Gradient is calculated by dividing the elevation change along the reach by the total reach length.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | AIM | AREMP | NARS | PIBO |
| Collection Method Link |  |  |  |  |

Gradient is measured consistency within a protocol (Roper et al. 2010). AIM, AREMP, and PIBO measure slope using the same method, but with different in tool used to measures. AIM and PIBO use an auto level, while AREMP users a laser and prism. NARS measures slope between transects, but is ?? Between the four programs the slope data is considered comparable between the four programs and has low error within programs between crews (AIM PROTOCOL 2017, Roper et al. 2010 )

Channel Charistics

Mean Bank-full width (m)- Bank-full width is the average width of the reach

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | AIM | AREMP | NARS | PIBO |
| Collection Method Link |  |  |  |  |

AIM, AREMP and PIBO all use the same method and definition for bank-full.

Substrate

Median particle size –

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | AIM | AREMP | NARS | PIBO |
| Collection Method Link |  |  |  |  |

There is a large variability in D50 within protocol (CITITATION). Therefore, it can be assumed the variability in D50 will increases as the field methods differ amongst programs. All programs use some variety on the Women pebble count yet there are differences in the area particles are collected AIM collects in the active channel, NARS collects in the wetted width, while PIBO collects within the bank-full channel (bank-full definition is describes the same between the programs, but high variable based on training). Even with the differences in collection methods and the variability in the data it has been shown that this metrics is suffeciet to distinguish broad differences in D50 and critical threshelds for find sediment (Roper et al 2010)

Cross-Sections (m)

Average bank-full width.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Substrate | Percent pool tail fines < 2mm | PoolTailFines2 | Average percent fine sediment (< 2mm) on the pool tail | Double |
| Substrate | Percent pool tail fines < 6mm | PoolTailFines6 | Average percent fine sediment (< 6mm) on the pool tail | Double |

**Residual Pool Depth (RPD) – (m)**

The average residual pool depth.

**Percent Pools (# pools/km)**

Collection protocol creates a substantial difference between estimates of percent pools.

**Large Wood Frequency (# pieces/100m)**

Large Wood Frequency is the pieces of large wood per 100m of stream reach. AIM, AREMP and PIBO all user similar procedures collection within the bank-full channel, while NARS counts wood in the channel and extending over the bank-full channel. AIM and PIBO define programs the definition the minimum size of the wood to count as AIM >0.1 m in diameter, >1m in length, AREMP >0.3 m in diameter, >3 m in length and NARS ???

Large Wood Volume

Volume of large wood debris within the