# Stream Monitoring Habitat Data Exchange Specifications

Data exchange specifications are a set of guidelines and rules for using and combining information. Rigorous data exchange specifications support reuse, promote interoperability, and reduce data integration costs (Morris and Frechette 2008, Hamm 2019). The [Stream Monitoring Habitat Data Exchange Specifications](MetricLevelExchangeSpecifications.docx) are a standard for exchanging metric-level habitat data based on the Darwin Core principles as outlined by Wieczorek et al. in 2012. The Darwin Core standard is maintained at the Git Hub repository <https://github.com/tdwg/dwc>. The Stream Habitat Metric Data Integration working group facilitated by the Pacific Northwest Aquatic Monitoring Partnership (<https://www.pnamp.org/project/habitat-metric-data-integration>) and the USGS adapted the Darwin Core standard for stream habitat metrics, and as a use case, integrate stream habitat metrics from three federal stream habitat monitoring programs in a separate Git Hub Repository: <https://github.com/rascully/Integrating-Stream-Monitoring-Data-From-Multiple-Programs>.

# Structure

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

## Record Level Class

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

## Location Class

Understanding where data are collected is critical to interpreting biological monitoring data. The Location class describes where information are collected, see the list of terms in the [Location table](Tables/Location_table.csv). There will be multiple locations in each data set. The locationID is the key to link locations to events. To view and analysis data from various sources, latitudes and longitude information must be consistent among data sets; therefore, for this data all latitude and longitudes are converted to WGS1984.

## Event Class

The Event class describes an action that occurs at a specific time frame see the [Event table](Tables/Event_table.csv) for the terms. To assess the status and trend of a resource as a response to management actions, stream habitat monitoring programs often implement a rotating panel design, meaning that the project returns to a single location multiple times during the study duration. Therefore, a data set will contain numerous locations, and each location can include numerous events.

## Measurement Or Fact (Metrics) Class

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

### Controlled Vocabulary

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

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

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

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

# Use Case

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

# Conclusion

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

# References

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

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

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

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