ReadMeForProjectTeam

Becca Scully

2/25/2021

R Markdown

Stream Monitoring Habitat Data Exchange Specifications

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

Structure

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

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. 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

Term	Description
datasetID	An identifier for the set of data. May be a global unique identifier or an identif
type	The nature or genre of the resource.
modified	The most recent date-time on which the combined dataset was changed.
rightsHolder	A person or organization owning or managing rights over the resource.
bibilographicCititation	A bibliographic reference for the resource as a statement indicating how this re
InstitutionID	An identifier for the institution having custody of the object(s) or information r
CollectionID	An identifier for the collection or dataset from which the record was derived.
datasetName	The name identifying the data set from which the record was derived.
institutionCode	The name (or acronym) in use by the institution having custody of the object(s

Term	Description	Examples	DataType
locationID	This is the location identification for the integrated data set the value is the concatenation of the verbatimlocationID and the institutionCode. Example) 5483AIM, 88963AREMP, WtR563EPA	NA	String
verbatimlocationID	Number that identifies a unique sampling location. A site is a stream segment with a fixed starting and ending location for sampling	NA	String
verbatimLatitude	The verbatim original latitude of the Location. The coordinate ellipsoid, geodeticDatum, or full Spatial Reference System (SRS) for these coordinates should be stored in verbatimSRS and the coordinate system should be stored in verbatimCoordinateSystem.	For this dataset we use the botton of the reach as the location.	Numeric
verbatimLongitude	The verbatim original longitude of the Location. The coordinate ellipsoid, geodeticDatum, or full Spatial Reference System (SRS) for these coordinates should be stored in verbatimSRS and the coordinate system should be stored in verbatimCoordinateSystem.	For this dataset we use the botton of the reach as the location.	Numeric
verbatimWaterbody	The water body name from the original data set.	For this data set this field is often refered to as Stream Name.	String
verbatimCoordinateSystem	The spatial coordinate system for the verbatimLatitude and verbatimLongitude or the verbatimCoordinates of the Location.	NA	String
StateProvince	The name of the next smaller administrative region than country (state, province, canton, department, region, etc.) in which the Location occurs.	NA	String
decimalLatitude	The geographic latitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are north of the Equator, negative values are south of it. Legal values lie between -90 and 90, inclusive.	NA	Numeric
decimalLongitude	The geographic longitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic center of a Location. Positive values are east of the Greenwich Meridian, negative values are west of it. Legal values lie between -180 and 180, inclusive.	NA	Numeric
geodeticDatum	The ellipsoid, geodetic datum, or spatial reference system (SRS) upon which the geographic coordinates given in decimalLatitude and decimalLongitude as based.	NA	string

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. 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 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 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.

Term	Description
verbatimEventID	Unique number that identifies one sample of a particular site.
eventID	An identifier for the set of information associated with an Event (something that
samplingProtocol	The name of, reference to, or description of the method or protocol used during ar
verbatimEventDate	The verbatim original representation of the date and time information for an Even
EventDate	The date-time or interval during which an Event occurred. For occurrences, this is
verbatimEventTime	The time or interval during which an Event occurred. Recommended best practice
day	The integer day of the month on which the Event occurred.
month	The ordinal month in which the Event occurred.
year	The four-digit year in which the Event occurred, according to the Common Era C
fieldNumber	An identifier given to the event in the field. Often serves as a link between field no
fieldNotes	One of a) an indicator of the existence of, b) a reference to (publication, URI), or
eventRemark	Comments or notes about the Event.

Term	Description
measurementID	An identifier for the MeasurementOrFact (information pertaining to mea
measurementType	The nature of the measurement, fact, characteristic, or assertion. Recom
measurementValue	The value of the measurement, fact, characteristic, or assertion.
measurementAccuracy	The description of the potential error associated with the measurementV
measurementUnit	The units associated with the measurementValue. Recommended best p
measurementDeterminedDate	The date on which the MeasurementOrFact was made. Recommended b
measurementDeterminedBy	A list (concatenated and separated) of names of people, groups, or orgar
measurementMethod	A description of or reference to (publication, URI) the method or protoc
measurementRemarks	Comments or notes accompanying the MeasurementOrFact.

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.

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 document.

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

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. 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.

Term	LongName	Description
BFWidth	Average bankfull width from transects	Average bankfull wi
Grad	Gradient of stream reach	Mean slope of wate
RchLen	Length of sampling reach	Length of sampling
BFWDRatio	Bankfull width to depth ratio at transects	Average Bankfull W
WetWidth	Average wetted width from transects	Average wetted wid
WetWidthToDepth	Wetted width to depth ratio at transects	Mean Wetted Widt
countTransects	Count of Transects	Number of transect
PctDry	Percent of Reach that is Dry	Percent of the reach
Beaver	Beaver Sign at Reach	Beaver value from t
StreamOrder	Stream Order	Strahler stream ord
RPD	Residual pool depth	Average of the resid
PctPool	Percent pools	Percent of the samp
BankAngle	Bank angle	Measured angle of t
PctStab	Percent stable banks	Percent of 42 banks
D50	Diameter of the 50th percentile streambed particle	Median diameter of
PctFines2	Percent of streambed particles <2mm	Percent of the strea
PctFines6	Percent of streambed particles <6mm	Percent of the strea
D16	Diameter of the 16th percentile streambed particle	Bed surface particle
D84	Diameter of the 84th percentile streambed particle	Bed surface particle
PctBdrk	Percent Bed Surface Bedrock	Percent of the strea
PoolTailFines2	Percent pool tail fines < 2mm	Average percent fin
PoolTailFines6	Percent pool tail fines < 6mm	Average percent fin
Temp	Mean annual tempeature	Average of mean da
WinterMean	Mean winter temperature (Dec, Jan, Feb)	Average of mean da
SpringMean	Mean spring temperature	Average of mean da
SummerMean	Mean summer temperatures	Average of mean da
AugustMean	Mean august temperature	Average of mean da
MeanFall	Mean fall temperature	Average of mean da
LowMean	Minimum daily temperature	Lowest mean daily
LowSevenDayAverage	Minimum weekly average temperature	Lowest seven-day ru
HighMean	Maximum daily temperature	Highest mean daily
HighSevenDayAverage	Maximum weekly average temperature	Highest seven-day r
DegreeDays	annual degree day s	Cumulative total of
SDMeanDaily	Annual Standard Deivation	Standard deviation
SDMeanWinterDaily	Winter Standard Deviation	Standard deviation
SDMeanSpring	Spring Standard Deviation	Standard deviation
SDMeanSummer	Summer SD	Standard deviation
SDMeanAugust	August SD	Standard deviation
SDMeanFall	Fall SD	Standard deviation
DiffMinMax	Range in extream daily termperature s	Difference between
DifMinMaxWeekly	Range in streme weekly temperatures	Difference between
SDAnnual	Interannual standard deviation of mean annual	Interannual standar
SDMinWeekly	Interannual standard deviation of minimum weekly	Interannual standar
SDMaxWeekly	Interannual standard deviation of maximum weekly	Interannual standar
SD5PercentDegreeDay	Interannual standard deviation of 5% degree days	Interannual standar
SD50PercentDegreeDay	Interannual standard deviation of 50% degree days	Interannual standar
NumberMeanGT20	Frequency of hot days	Number of days wit
NumberMeanLT2	Frequency of cold days	Number of days wit
NumberDaysDec	Date of 5% degree days	Number of days fro
NumberDaysDec25	Date of 25% degree days	Number of days f