Geophysical Research Abstracts,
Vol. 11, EGU2009-11115-1, 2009
EGU General Assembly 2009

A Constrained and Versioned Data Model for TEAM Data

S. Andelman (1), C. Baru (2), S. Chandra (2), E. Fegraus (1), and K. Lin (2)

(1) TEAM Network, Center for Applied Biodiversity Sciences, Conservation International, (2) San Diego Supercomputer Center, UCSD, Science R&D, La Jolla, United States (chandras@sdsc.edu)

The objective of the Tropical Ecology Assessment and Monitoring Network (www.teamnetwork.org) is "To generate real time data for monitoring long-term trends in tropical biodiversity through a global network of TEAM sites (i.e. field stations in tropical forests), providing an early warning system on the status of biodiversity to effectively guide conservation action". To achieve this, the TEAM Network operates by collecting data via standardized protocols at TEAM Sites. The standardized TEAM protocols include the Climate, Vegetation and Terrestrial Vertebrate Protocols. Some sites also implement additional protocols. There are currently 7 TEAM Sites with plans to grow the network to 15 by June 30, 2009 and 50 TEAM Sites by the end of 2010.

At each TEAM Site, data is gathered as defined by the protocols and according to a predefined sampling schedule. The TEAM data is organized and stored in a database based on the TEAM spatio-temporal data model. This data model is at the core of the TEAM Information System - it consumes and executes spatio-temporal queries, and analytical functions that are performed on TEAM data, and defines the object data types, relationships and operations that maintain database integrity.

The TEAM data model contains object types including types for observation objects (e.g. bird, butterfly and trees), sampling unit, person, role, protocol, site and the relationship of these object types. Each observation data record is a set of attribute values of an observation object and is always associated with a sampling unit, an observation timestamp or time interval, a versioned protocol and data collectors.

The operations on the TEAM data model can be classified as read operations, insert operations and update operations. Following are some typical operations:

- The operation *get(site, protocol, [sampling unit block, sampling unit,] start time, end time)* returns all data records using the specified protocol and collected at the specified site, block, sampling unit and time range.
- The operation *insertSamplingUnit(sampling unit, site, protocol)* saves a new sampling unit into the data model and links it with the site and protocol.
- The operation *updateSampligUnit(sampling_unit_id, attribute, value)* changes the attribute (e.g. latitude or longitude) of the sampling unit to the specified value.
- The operation *insertData(observation record, site, protocol, sampling unit, timestamps, data collectors)* saves a new observation record into the database and associates it with specified objects.
- The operation *updateData(protocol, data_id, attribute, value)* modifies the attribute of an existing observation record to the specified value.

All the insert or update operations require: 1) authorization to ensure the user has necessary privileges to perform the operation; 2) timestamp validation to ensure the observation timestamps are in the designated time range specified in the sampling schedule; 3) data validation to check that the data records use correct taxonomy terms and data values. No authorization is performed for get operations, but under some specific condition, a username may be required for the purpose of authentication.

Along with the validations above, the TEAM data model also supports human based data validation on observed data through the Data Review subsystem to ensure data quality. The data review is implemented by adding two

attributes *review_tag* and *review_comment* to each observation data record. The attribute *review_tag* is used by a reviewer to specify the quality of data, and the attribute *review_comment* is for reviewers to give more information when a problem is identified. The review_tag attribute can be populated by either the system conducting QA/QC tests or by pre-specified scientific experts. The following is the review operation, which is actually a special case of the operation *updateData*:

- The operation *updateReview(protocol, data_id, judgment, comment)* sets the attribute *review_tag* and review_*comment* to the specified values.

By systematically tracking every step, The TEAM data model can roll back to any previous state. This is achieved by introducing a historical data container for each editable object type. When the operation *updateData* is applied to an object to modify its attribute, the object will be tagged with the current timestamp and the name of the user who conducts the operation, the tagged object will then be moved into the historical data container, and finally a new object will be created with the new value for the specified attribute.

The diagram illustrates the architecture of the TEAM data management system.

A data collector can use the Data Ingestion subsystem to load new data records into the TEAM data model. The system establishes a first level of review (i.e. meets minimum data standards via QA/QC tests). Further review is done via experts and they can verify and provide their comments on data records through the Data Review subsystem. The data editor can then address data records based on the reviewer's comments. Users can use the Data Query and Download application to find data by sites, protocols and time ranges. The Data Query and Download system packages selected data with the data license and important metadata information into a single package and delivers it to the user.