June 2011 - ClimDB/HydroDB integration into PASTA

June 2011 - ClimDB/HydroDB integration into PASTA

Thu, 06/02/2011 - 5:46pm — dhenshaw

June 2011 - ClimDB/HydroDB integration into PASTA
Videoteleconference 6 Jun, 7 Jun 2011
Notes: Don Henshaw

6 Jun participants: Don Henshaw, James Brunt, Karen Baker, John Campbell, Effie Greathouse, Gastil-Buhl, Margaret O'Brien, Wade Sheldon, Philip Tarrant, Jonathan Walsh, Hope Humphries, Jamie Hollingsworth, Theresa Valentine

7 Jun participants: Don Henshaw, Yang Xia, Corinna Gries, John Porter, Aaron Stephenson, Hap Garritt, Ken Ramsey

Please see the VTC Powerpoint at http://im.lternet.edu/node/887 [1] as well as the LNO planning document "ClimDB/HydroDB/EcoTrends Network Database Integration Path and Request for Comments" (http://intranet2.lternet.edu/content/climdbhydrodbecotrends-network-data... [2]).

The development of ClimDB/HydroDB (ClimHy) within the Network Information System (NIS) and using the PASTA architecture will be referred to

as ClimHy integration. The movement of the ClimHy data harvester, data warehouse, and webpage from the Andrews LTER to the LNO will be referred to as the ClimHy migration. This migration was completed in 2010 and ClimHy is accessible at http://climhy.lternet.edu/ [3]. The participant page now resides at http://climhy.lternet.edu/harvest.html [4].

A committee is being established to consider the general approach for ClimHy integration. Initial members are Don Henshaw, James Brunt, Yang Xia, and Suzanne Remillard. During the VTC both Jonathan Walsh and Ken Ramsey expressed interest in joining.

Discussion: It is noted that the proposed metadata-driven approach for ClimHy harvest is an inverted model from the current harvester, which is largely data-driven. The current process does not require updated metadata and depends only on preparing data. The new process might require EML-ready data sets of the participants. The goal for what we would like to accomplish should be established – are we looking for a more flexible harvest system that could better utilize existing data streams by leveraging PASTA architecture, or are we hoping to develop an attribute ontology to provide standard output for use in scientific workflows, are we hoping to improve metadata to better describe resulting products, or others. Every integration goal will likely allow the IMC to explore this new management framework (PASTA) that we will need to learn to use to do our work in a new way. We would like to improve the existing product without overly taxing the data providers with new requirements. It is also important to preserve existing public webpage enhancements such as user-defined data downloads and easily generated plots of specified measurement variables.

PASTA may be software we leverage to accomplish this integration, but it is noted that a hybrid model that takes advantage of other resources, e.g., external tables, may be necessary. For example, there is extensive "metadata" that is specific to climate and hydrology that describes specific sites, methods, and watershed characteristics and treatments (See descriptors.xls). This metadata does not often change and may be unwieldy to continuously harvest within EML with regular updates. It might make sense to populate SiteDB with site descriptive information for use alongside the PASTA framework. SiteDB could centralize specific station coordinate information that could be shared with other cross-site data development, e.g., StreamChemDB, or be useful in coordinating an export web service into the CUAHSI ODM. Potentially, SiteDB could be extended to accommodate additional research nodes, for example, adding a hydrology node to include the watershed characteristics or other specific metadata required for HydroDB. Another example might be a central methods code table that could cross-link back to the data as a better means of describing methodology changes for the data user – this table approach is similar to that used by the CUAHSI ODM. The technology we use to capture this extended metadata may not be so important if it only needs to be done once or occasionally and doesn't require continual maintenance.

One scenario might have us develop EML for the standard input format now employed by ClimHy, or a somewhat revised "skinny" model that could be highly normalized, simple to describe, and able to accommodate all of the existing measurement values. A more advanced scenario that would involve letting PASTA resample a site data set to prepare data for ClimHy will require greater development. Can a PASTA front-end allow seamless access to these workflows in populating a central database? The workflow tiger team convening this fall may provide considerable feedback to the ClimHy integration effort and help us better understand what is possible. A workflows production workshop scheduled for 2012 may use this integration effort as a means to help sites understand and write workflow scripts.

Of concern is the existence of considerable site climate data, e.g., high-temporal resolution data, that is not currently included in ClimHy. Should we make inclusion of this data a priority? Another concern is that aggregated products developed by ClimHy may not match similar aggregations at the site or in EcoTrends. Consideration of aggregation approaches and data qualifier codes for aggregated data values is needed.

Feedback to the sites on the use of participating data is not available. Any new system should be able to provide a use summary of site data. A recent LNO summary (Yang) of ClimHy data use shows that 38% of the users do not identify the expected use of the data (registration form is optional), 23 % use is research, 14% use is education, 14% use is general, 8% use is site testing, and 3% use is management. The use of this site averages about 25 ClimHy sessions per day (2010-2011) and the number of downloads and data plots are increasing (see Powerpoint slide 4).

The LTER Climate Committee is requesting a status report on currentness of data in ClimHy. An IMC initiative may be required to better populate the existing system to include a richer set of measurement data and to provide necessary ClimHy-specific metadata that is largely incomplete for most sites.

Related Links

- LNO Request for comments on draft integration plan [5]

Attachment

ClimDB_integration_path.pptx [6]
descriptors.xls [7]
ClimDB_integration_path.pdf [8]
SiteDB_schema.pdf [9]
siteDBRevised_2000.jpg [10]
StreamChemDB - 1 - instantaneous & composite chemistry data tables relationships.pdf [11]
StreamChemDB - 2 - monthly derived chemistry data tables relationships.pdf [12]
StreamChemDB - 3 - chemistry methods tables relationships - for monthly and instantaneous & composite da
StreamChemDB - 4 - relationships for additional site-level tables.pdf [14]

Attachment
StreamChemDB - 5 - relationships for additional basin-level tables on watershed disturbances.pdf [15]

- Virtual Updates [16]

Source URL: http://im.lternet.edu/node/887

Links:
[1] http://im.lternet.edu/node/887
[2] http://intranet2.lternet.edu/content/climdbhydrodbecotrends-network-database-integration-path
[3] http://climhy.lternet.edu/
[4] http://climhy.lternet.edu/harvest.html
[5] http://im.lternet.edu/links/goto/887/154/links_related
[6] http://im.lternet.edu/sites/im.lternet.edu/files/ClimDB_integration_path_0.pptx
[7] http://im.lternet.edu/sites/im.lternet.edu/files/descriptors.xls
[8] http://im.lternet.edu/sites/im.lternet.edu/files/ClimDB_integration_path.pdf
[9] http://im.lternet.edu/sites/im.lternet.edu/files/SiteDB_schema.pdf
[10] http://im.lternet.edu/sites/im.lternet.edu/files/siteDBRevised_2000.jpg
[11] http://im.lternet.edu/sites/im.lternet.edu/files/StreamChemDB - 1 - instantaneous & composite chemistry data tables relationships.pdf
[12] http://im.lternet.edu/sites/im.lternet.edu/files/StreamChemDB - 2 - monthly derived chemistry data tables relationships.pdf
[13] http://im.lternet.edu/sites/im.lternet.edu/files/StreamChemDB - 3 - chemistry methods tables relationships - for monthly and instantaneous & composite data.pdf
[14] http://im.lternet.edu/sites/im.lternet.edu/files/StreamChemDB - 4 - relationships for additional site-level tables.pdf
[15] http://im.lternet.edu/sites/im.lternet.edu/files/StreamChemDB - 5 - relationships for additional basin-level tables on watershed disturbances.pdf
[16] http://im.lternet.edu/taxonomy/term/169