

There's a Hole in My Bucket — and in the Data as Well!

SDSC Researchers Use Cyberinfrastructure to Standardize Water Data Collections

May 5, 2008

Jan Zverina

Like the popular children's song "There's a Hole in My Bucket," in which Liza and Henry try to patch a leaking pail, researchers with the San Diego Supercomputer Center at UC San Diego are plugging a hole in the data management process by creating a universally accepted cyberinfrastructure to study our most valuable natural resource - water.

The initiative, called the Hydrologic Information System (HIS), is supported by a 5-year grant from the National Science Foundation (NSF) to a team of researchers and software developers from five universities. The HIS project is being developed in close collaboration with the Consortium of Universities for the Advancement of Hydrologic Science, Inc., or CUAHSI (Pronounced 'quasi'), it is a joint effort among more than 100 universities and funded by NSF to advance research in hydrology, or the science of water, its properties, distribution and circulation on and below the earth's surface and in the atmosphere.

Ilya Zaslavsky, director of SDSC's Spatial Information Systems Laboratory and a key architect of HIS, points to the flood of data on water quality and quantity that's collected daily via thousands of sensor stations through a multitude of agencies including the Environmental Protection Agency (EPA), U.S. Geological Survey (USGS), U.S. Department of Agriculture (USDA), and the National Oceanic and Atmospheric Administration (NOAA).

"We're drowning in data, but the problem is that most, if not all, of these databases are incompatible with each other," said Zaslavsky. "Despite water being such a precious commodity and its conservation being such an important issue these days, researchers still don't have an accurate assessment of just how much water we have as a nation."

Developed by Zaslavsky and a team of researchers from around the country, HIS is currently in the first phases of forming a web-based cyberinfrastructure, or the interrelation of computing power, data services and academic expertise. SDSC is the technical partner in HIS, with the national supercomputer center contributing its expertise in web services, online serving of geospatial data, and development of cyberinfrastructure nodes. SDSC houses comprehensive observations catalogs referencing water data collections, and is also responsible for hosting project data and related services as well as the deployment of HIS applications.

HIS is designed to serve several functions. It facilitates broad and uniform user access to comprehensive distributed collections of water data from federal, state and local repositories, and lets users publish new observation datasets. HIS also provides a common information model and relational schema for storing hydrologic observations data, water data exchange protocols and web services, and a range of hydrologic controlled vocabularies.

Additionally, HIS is intended to better enable cross-scale analysis of hydrologic cycles and processes on either a regional or continental scale by linking with a range of climate models and integrating data from neighboring disciplines.

This summer, HIS researchers will release "Version 1.1" of the HIS server software stack to eleven NSF hydrologic observatory test bed sites, after several months of collecting feedback from users and enhancing the overall system. Late last year, SDSC researchers installed the first version of the HIS server software - including databases, tools for web publishing of observations data, front-end applications and a comprehensive web-based data discovery and retrieval system - on dedicated servers before shipping them to the test bed sites, including one at UC Merced. The other NSF test bed sites are in Florida, Iowa, New York, North Carolina, Maryland, Minnesota, Montana, Texas, Utah, and Virginia.

At the core of the HIS system is WaterOneFlow web services, a set of web services for finding and retrieving hydrologic observations data in WaterML format. Under development by HIS researchers, WaterML is an Extensible Markup Language (XML) specification for exchanging water observations that is now being widely accepted throughout the hydrologic community. WaterOneFlow services provide access to large repositories of hydrologic observations maintained at federal agencies such as the EPA, USDA, USGS and NOAA, as well as numerous academic data collections developed in the course of university projects all over the country.

The ability to access this catalog and retrieve observations data from distributed repositories made this approach attractive to many developers and analysts. Environmental agencies in several states, including Florida, Texas and Idaho, are already working with the HIS team on incorporating their data repositories into the overall system. These agencies have plans to either install the HIS server software stack on their computers, or work with local universities on jointly managing access to their data collections

"We have had application interest from Arizona to Australia," said Zaslavsky, adding that the HIS team at SDSC is offering server deployment and maintenance services to organizations interested in online serving and integration of hydrologic observations, including universities, local governments, community groups, and environmental consultants.

In addition, the USGS recently agreed to adopt the web services application programming interface developed under the HIS program, while the National Climatic Data Center (NCDC) began using CUAHSI's WaterML specification for its Automated Surface Observing System (ASOS) last year. CUAHSI researchers are also working with the EPA to harmonize WaterML with the EPA's WQX web services.

"We are extremely encouraged that the USGS and NCDC have chosen to adopt specifications developed within the HIS project," said Zaslavsky. "Quite simply, the advancement of water science is directly dependent on the integration of all this data into a single representation as we seek the answers to key questions about our water supply."

David Maidment, a world-renowned hydrologist from University of Texas at Austin, leads the overall project. Other key members of the CUAHSI HIS project include SDSC researcher and distinguished scientist Chaitan Baru; David Tarboton of Utah State University; Michael Piasecki of Drexel University; and Jon Goodall of University of South Carolina.

Media Contacts: Jan Zverina, SDSC Communications, 858-534-5111 Warren R. Froelich, SDSC Communications, 858-822-3622