ENABLING THE FREE FLOW OF WATER DATA

Daniel P. Ames and David G. Tarboton

In April 2011, I (Daniel Ames), attended the first OpenWater Symposium at the UNESCO-IHE Institute for Water Education in The Netherlands to present the MapWindow GIS software and some water-related plugins. Although other meetings, including AWRA specialty conferences, had previously held sessions on open source software tools in the water domain, the UNESCO-IHE meeting was the first international gathering completely dedicated to the topic of free and open source models, data management tools, and systems for working with water data.

It was an exciting meeting with presentations and workshops on several innovative open software tools. At the end of the meeting, one of my European colleagues pulled me aside and said something along these lines: "You Americans don't know how lucky you are. You have so much freely available data. Here in Europe, we pay for data three times. First we pay the government through our taxes to collect the data. Then we pay the government again to get a copy of the data to use in our research. Finally, we pay again if we want to reproduce someone else's research results." This conversation made me appreciate the immense amount of freely available water data we have in the United States (U.S.). At the same time, sheer volumes of data are meaningless if they can't be searched, retrieved, and used efficiently.

To their credit, the Europeans are making progress on open data sharing through policies such as the European Union Water Framework Directive. In the U.S., the Open Water Data Initiative (http://acwi.gov/spatial/owdi/) continues our tradition of supporting freely available water information. As these open data policies advance, here and abroad, we can expect a "flood" of new climate and water data to become available for use in water research and management. This increased data availability, in turn, creates a greater need for international standards for passing data between models, databases, and other software tools. Without such standards, water data tools risk becoming stove-piped into specific "software stacks" that are able to communicate with each other but not with other programs that exist outside the stack (consider the Android versus iPhone software ecosystems).

This need for interoperability between machines, is anticipated in the U.S. presidential executive order on open data, entitled "Making Open and Machine Readable the New Default for Government Information" (Obama, 2013). Making data "machine readable" will ensure that the coming water data deluge can be met with equally powerful software tools. In the remainder of this article, we present three important developments in open water standards and software that will help us rise to the water data challenge.

WATERML2

It is not uncommon for time series water data to be stored and passed between agencies, researchers, and practitioners in homemade text files, Microsoft Excel files, vendor-specific data logger files, or simply as graphs and charts published in articles such as this one. The inherent inconsistency in both data content and organization in these methods makes it extremely difficult for programmers to work with the information contained in the files. For the sake of illustration, imagine if every Internet server used a different means for encoding web pages. The Internet requires agreement between data providers (web servers) and data consumers (web browsers) regarding how information should be encoded. Hypertext markup language (HTML) and its more generic cousin, XML, are the core standards that make the Internet work.

Many international organizations have tried to address the water data inconsistency problem by proposing standard file formats. Notable efforts include the Water Data Transfer Format (WDTF) in Australia, WaterML 1.0 by the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) in the U.S., the xHydro standard in Germany, and the EA XML standard in the United Kingdom. The creation of all of these standards raised a new problem, how to standardize the standards? Beginning in 2009 with the formation of the Open Geospatial Consortium (OGC) Hydrology Domain Working Group (HDWG) a harmonization effort was undertaken, resulting in a new protocol that inherited the best features of each of the other standards. The new protocol, WaterML2, was presented at the OpenWater Symposium in The Netherlands, and, one year later, was adopted by OGC as an international standard – and a potential replacement for each of its progenitors (Taylor *et al.*, 2013).

A WaterML2 file – in the most practical sense – is a formatted text file that includes metadata about the data collection location together with details about observed variables, data quality, and the time and value of specific measurements. The advent WaterML2 is helping foster a favorable open software climate and supports fundamental interoperability between both open and closed source water models and water management systems. Two of the largest commercial water data management software companies, KISTERS and Aquatic Informatics, have committed to supporting WaterML2 as a means for water data exchange. Several major federal agencies are also adopting the standard.

The WaterML2 standard builds upon a number of other OGC standards for time and space data encoding including Observations and Measurements (O&M), Geographic Markup Language (GML), and Keyhole Markup Language (KML). It also compliments standards for web based data retrieval including Sensor Observation Service (SOS), Web Feature Service (WFS), and Web Processing Service (WPS). Other open data encoding protocols, such as NetCDF, HDF5, GeoTIFF, and Esri Shapefile have become de facto standards by virtue of their published formats and wide adoption. Each of these fills different water information needs such as encoding terrain, watershed boundaries, and satellite observations data. Collectively, the emergence of these published means for water information sharing enables the Open Water Data Initiative.

CUAHSI HIS

A landmark data management system in the hydrologic sciences community is the CUAHSI Hydrologic Information System (HIS) (Tarboton *et al.*, 2009). Funded by the National Science Foundation beginning in 2004, the CUAHSI HIS was the first major open source effort of its kind. The CUAHSI HIS is a three-legged system including server tools for data publication, client tools for data download and visualization, and cataloging tools for data search and discovery. Data is organized using the Observations Data Model, searched via the WaterOneFlow (WOF) web service, and transmitted using the WaterML 1.0 file format.

Originally envisioned as a federated data sharing system with identical HydroServer "appliances" deployed at several partner universities, the CUAHSI HIS has evolved to include both distributed servers and centralized data services at the CUAHSI Water Data Center (WDC). This means you can host your data at the WDC or on your own web server. The original Microsoft SQL Server based HydroServer has been re-implemented in a lightweight Linux package called HydroServer Lite and as a Python package called WOFpy. Each of these software packages is available for free download, modification, and use.

Several free client tools for downloading and viewing data stored on a HydroServer have also been developed. Microsoft Windows users can install and use the GIS-based application, HydroDesktop, available at http://www.hydrodesktop.org. Users of the R statistical software can retrieve data using the WaterML R package, which can be found the R CRAN library. A web-based search tool, the CUAHSI Water Data Client, can also be used to search and download data (see http://data.cuahsi.org). Each of these tools provides a means for searching the billions of data observations in the CUAHSI HIS and downloading and viewing selected time series datasets.

The search capability of CUAHSI HIS client tools is facilitated by a data catalog at the Water Data Center that regularly scans data servers and stores searchable metadata – not entirely unlike the approach used by Internet search engines to catalog web sites. Using this catalog, data publishers can register a new HydroServer and make their datasets discoverable via the catalog's exposed web services. You can learn more about registering your data server with in the CUAHSI catalog at http://wdc.cuahsi.org. Future enhancement of the Water Data Center catalog will include support for newer data sharing standards and a new web-based R scripting platform.

HYDROSHARE

The new era of Big Data has been associated with a so-called "fourth paradigm" of scientific progress where discoveries are produced by exploration and data intensive analysis of massive datasets (Hey *et al.,* 2009). Use of large datasets requires machine readability (via common data standards) and scalable tools (e.g. through cloud computing.) The computer programs and software applications that move data around the Internet and allow us to view and analyze it are an integral part of the cyberinfrastructure needed to fulfill the promise of open water data. A new and exciting development in this broader area of water cyberinfrastructure is HydroShare, a collaborative website for the sharing of data and models (Tarboton *et al.,* 2014). HydroShare is under active development with version 1.5 presently deployed at http://www.hydroshare. org.

When introducing HydroShare at technical meetings we often refer to it as "YouTube for water data." If you have ever uploaded a video to YouTube (or simply watched a funny cat video that someone else uploaded) then the analogy should make perfect sense. A user uploads a data "resource" to HydroShare, tagged with filterable keywords and metadata. Data resources can include references (such as a link to an HIS HydroServer dataset) or actual files of a nearly any file type. Once a user has uploaded a resource, it can be marked as public, which allows it to be discoverable by other HydroShare users. And, as with a YouTube video, a HydroShare resource can receive comments and "likes" and thus grow in value through this crowd sourced "social" metadata.

Just as YouTube simplifies uploading and sharing videos in any video format, we are creating HydroShare to simplify the sharing of hydrologic data in any file format. In HydroShare we call data files "resources" and the most basic resource type is a "generic resource," which can literally be any file on your computer – even a cat video (as long as it is water related!). If a user flags their data as one of several "known" types then HydroShare can perform extra functions such as showing raster or feature dataset on a map, plotting a time series graph, or even launching a particular model from a set of model input files. We are actively adding new resource types and visualization tools that will be available in the coming year.

Using the Tethys Platform (Jones *et al.,* 2014), future versions of HydroShare will be able to run water simulation models directly in the cloud, presenting results to users for further visualization and decision-making. The collaborative elements of HydroShare are expected to encourage live discussions among groups of engineers, water managers and water scientists regarding data and model results.

SUMMARY

In summary, it is an exciting time to be a water data geek. New and emerging standards for water data encoding such as WaterML2 are creating opportunities for research and engineering applications that were never possible even just a few years ago. Distributed data sharing systems like the CUAHSI HIS are leading the vanguard in a global effort to lower the data discovery bar. HydroShare and systems like it are going to change the way we do science with and manage the massive quantities of data that are rapidly filling our virtual clouds. Indeed, the tools are here to realize the vision of the Open Water Data Initiative.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under collaborative grants ACI 1148453 and ACI 1148090 for the development of HydroShare (http://www.hydroshare.org) and grant EAR 1343785. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Hey, A.J.G., S. Tansley, and K. M. Tolle, 2009, The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmond, Washington.

Jones, N., J. Nelson, N. Swain, S. Christensen, D. Tarboton, and P. Dash, 2014. Tethys: A Software Framework for Web-Based Modeling and Decision Support Applications, *In:* 7th International Congress on Environmental Modelling and Software, San Diego, California.

Obama, B., 2013. Executive Order – Making Open and Machine Readable the New Default for Government Information. Available at https://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government.

Tarboton, D.G., J.S. Horsburgh, D.R. Maidment, T. Whiteaker, I. Zaslavsky, M. Piasecki. J. Goodall, D. Valentine, and T. Whitenack. 2009. Development of a Community Hydrologic Information System. *In:* 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation, pp. 988-994.

Tarboton, David G., Ray Idaszak, Jeffrey S. Horsburgh, Jeff Heard, Daniel P. Ames, Jonathan L. Goodall, Larry Band, Venkatesh Merwade, Alva Couch, Jennifer Arrigo, Richard Hooper, David Valentine, David Maidment, 2014. HydroShare: Advancing Collaboration Through Hydrologic Data and Model Sharing. *In:* Proceedings of the 7th International Congress on Environmental Modelling and Software, San Diego, California.

Taylor, Peter, Simon Cox, Gavin Walker, David Valentine, and Paul Sheahan, 2013. WaterMl2.0: Development of an Open Standard for Hydrological Time-Series Data Exchange. Journal of Hydroinformatics, doi.10.2166/hydro.2013.174, http://www.iwaponline.com/jh/up/jh2013174.htm.

AUTHOR LINK

Daniel P. Ames
Water Resources Engineering and
Geospatial Technologies
242J Clyde Bldg.
Brigham Young University
Provo, UT, 84664
(801) 422-3620

David G. Tarboton
Utah Water Research Laboratory
Civil and Environmental Engineering
4110 Old Main Hill
Utah State University
Logan, UT 84322-4110
(435) 797-3172

E-Mail dan.ames@byu.edu
dtarb@usu.edu

Dr. Daniel P. Ames is an Associate Professor in Civil and Environmental Engineering at Brigham Young University in Provo, Utah. His teaching and research interests include hydroinformatics, water science Big Data, and geographic information systems. Dr. Ames leads several open source software projects including MapWindow, DotSpatial, and HydroDesktop. Dr. Ames is a member of the AWRA Technology Committee.

Dr. David G. Tarboton is a professor of Civil and Environmental Engineering, Utah Water Research Laboratory, Utah State University. His research focuses on advancing the capability for hydrologic prediction by developing models that take advantage of new information and process understanding enabled by new technology. He has developed a number of models and software packages including the TauDEM hydrologic terrain analysis and channel network extraction package and Utah Energy Balance snowmelt model. He is lead on the National Science Foundation HydroShare project for the development of a collaborative environment for sharing hydrologic data and models. He teaches Hydrology and Geographic Information Systems in Water Resources.

❖ ❖ ❖

▲ HIGHLIGHTS OF JAWRA TECHNICAL PAPERS
OCTOBER 2015 • VOL. 51 • NO. 5

EDITORIAL: WIGINGTON DISCUSSES LARGE-SCALE WATER RESOURCE ISSUES ADDRESSED BY RECENT JAWRA JOURNAL ARTICLES.

TECHNICAL PAPERS

- **Perrone *et al.*,** present a comprehensive literature review of national and regional United States water-use estimates and projections.
- **Keum** and **Kaluarachchi** conduct uncertainty analysis using the SPARROW model to estimate dissolved-solids transport in the Upper Colorado River basin.
- **Lam *et al.*,** present a cost-effective laser scanning method for stream channel geometry and roughness.
- **Kharel** and **Kirilenko** use SWAT to evaluate the influence of climate change on long-term flood risks of Devils Lake in North Dakota.
- **Moorhead *et al.*,** assess the accuracy NOAA gridded daily reference evapotranspiration data for the Texas High Plains.
- **Sangwan** and **Merwade** develop a fast, economical approach to floodplain mapping using soil information.
- **King *et al.*,** develop an improved weather generator algorithm for multisite simulation of precipitation and temperature.
- **Johnson *et al.*,** use SWAT to model the sensitivity of streamflow and water quality to climate change and urban development in large U.S. watersheds.
- **Vineyard *et al.*,** compare green and gray infrastructure using life cycle cost and environmental impact for a rain garden case study in Cincinnati, Ohio.
- **Stets *et al.*,** use long-term water quality monitoring data to discern regional and temporal nitrate trends across the U.S.
- **Jessup** and **Pappani** develop an assessment approach for Idaho streams that combines biological and habitat indices.
- **Reiter *et al.*,** examine a long-term data set from forested watersheds in the Pacific Northwest to evaluate the combined effects of hydro-climatic patterns and forest management on stream temperature.
- **Evans *et al.*,** review literature dealing with the hydrologic effects of surface coal mining in the Appalachian Mountains.

A full Table of Contents may be viewed at
http://www.onlinelibrary.wiley.com/doi/10.1111/jawr.2015.51.issue-5/issuetoc
JAWRA ~ Journal of the American Water Resources Association