Jun 16th, 2:00 PM - 3:20 PM

Enabling Water Science at the CUAHSI Water Data Center

Alva Couch
Tufts University, CUAHSI, alva.couch@tufts.edu

Richard Hooper
CUAHSI

Jon Pollak
CUAHSI

Marie Martin
CUAHSI

Martin Seul
CUAHSI

International Environmental Modelling and Software Society (iEMSs)
7th Int'l Congress on Env. Modelling and Software San Diego, California,
USA Daniel P. Ames, Nigel W.T. Quinn, and Andrea E. Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

Enabling Water Science at the CUAHSI Water Data Center

Alva Couch [ab], **Richard Hooper**[b], **Jon Pollak**[b], **Marie Martin**[b], **and Martin Seul**[b]

[a] *Tufts University, Computer Science, 161 College Avenue, Medford, MA, USA (alva.couch@tufts.edu)*

[b] *CUAHSI, 196 Boston Avenue, Suite 3000, Medford, MA, USA 02155 (acouch@cuahsi.org)*

Abstract: The CUAHSI Water Data Center (WDC) is a community-governed, multi-disciplinary data center focused upon the needs of water-related science in all academic disciplines. The WDC builds upon the successes of the 10-year effort to develop the CUAHSI Hydrologic Information System (HIS), and looks beyond HIS toward providing next-generation water data services. In partnership with the National Science Foundation, the WDC seeks to set the standard for data publication, persistence, and reliability, by providing formal user support services, using cloud-based abstractions and services, building new and accessible user interfaces to data, and establishing and sustaining data curation processes centered around optimizing the user experience. This paper documents the lessons learned in the WDC's first year of operation, and sketches the future of the WDC in the coming years.

Keywords: CUAHSI, water data center, hydrologic information system, user experience, hydro-informatics, cloud computing

1 INTRODUCTION

The CUAHSI Water Data Center (WDC) was founded upon the successes of the CUAHSI Hydrologic Information System (HIS) (Tarboton et al. [2009]; Horsburgh et al. [2009]; Tarboton et al. [2010]) in enabling water science, and with the aim of converting the "prototype" CUAHSI HIS into a "product" that is usable by a large variety of water scientists. CUAHSI HIS consists of three main parts: a data publication server (HydroServer) (Horsburgh et al. [2010]), a data catalog and search engine (HydroCatalog) (Whitenack [2010]), and a data access client (HydroDesktop) (Ames et al. [2012]). The initial aims of the Water Data Center – when it was established one year ago – included addressing a number of known usability and reliability problems in CUAHSI HIS, mostly by taking ownership of the software service stacks for CUAHSI HIS and responsibility for updating those service stacks.

We have come to understand in the first year of WDC operation that assuring usability and utility for CUAHSI HIS requires a broader approach than simply maintaining software. CUAHSI HIS was born because water science graduate students spent roughly 2/3 of their time locating and preparing data for analysis, and only 1/3 of their time actually engaging in scientific inquiry (Maidment [2005]). Accordingly, the original developers of CUAHSI HIS – a majority of whom are academic researchers – placed great emphasis upon increasing academic productivity. The simple fact that academics need to publish has shaped the software as empowering of academic publishing and as a prototype of publishable features. However, this also meant that features that do not seem to have direct and tangible benefits upon academic productivity, and cannot be published in academic venues as academic innovations, remained at low priority.

Before the Water Data Center was even conceived, paper co-author, Computer Science professor, and network management expert Alva Couch was asked by CUAHSI to report on the status of CUAHSI HIS as an external observer. Couch [2012] made several conclusions at that time: (a) Several parts of HydroCatalog needed to be re-engineered for better usability, especially in terms of query performance and harvesting reliability. (b) Having individual scientists run HydroServers for themselves has led to significant and documented sustainability problems for data publishing, especially when funding ends, and cloud options for data publishing should be pursued. (c) The easiest cloud option seemed to be Microsoft Azure, because adopting that platform would require minimum code re-engineering for cloud adoption. (d) The HydroDesktop client should be left to evolve as a tool for expert users, but a simpler data access client should be provided for

less expert and beginning users, who may well adopt HydroDesktop after an initial learning experience. The initial design and proposal for the Water Data Center was based in part upon these recommendations.

The funding of the CUAHSI Water Data Center on April 1, 2013 began a slow and deliberate evolution from our initial vision of "managing and curating software" to the current vision of "managing, maintaining, and curating the user experience". In turn, we realized that the real overarching task of the WDC is to create and manage a software ecosystem that is fundamentally different than the ecosystem in which CUAHSI HIS was born, with the end goal of providing reliable and robust user services rather than just software. This paper documents that evolution and the lessons learned in the first year of WDC operation, and how these lessons inform the future of the Water Data Center.

1.1 Beyond the Triangle Diagram

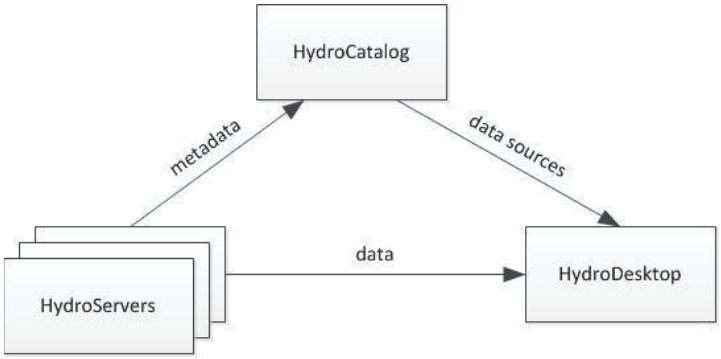

Figure 1. The "triangle diagram" depicting interactions of CUAHSI HIS components HydroCatalog, HydroServer, and HydroDesktop.

A standard representation of CUAHSI HIS is the so-called "triangle diagram" (Figure 1) described in Tarboton et al. [2010]:

1. A data server: HydroServer, analogous to a web server (Horsburgh et al. [2010]).
2. A data catalog: HydroCatalog, analogous to a web search engine (Whitenack [2010]).
3. A data browser: HydroDesktop, analogous to a web browser (Ames et al. [2012]).

In fact, this is somewhat of an over-simplification: there are several other crucial parts to CUAHSI HIS, other than software:

1. A community-edited controlled vocabulary for specifying metadata, including vocabularies for variables, units, measurement quality control levels, and other metadata (Horsburgh et al. [2014]).
2. A shared data model that underpins all subsystems: the *Observations Data Model* (ODM) (Horsburgh et al. [2008]).
3. A shared data representation format: The *Water Markup Language* (WaterML) (Anon. [2014a]).
4. A shared service architecture: *Water One Flow Services* (WOFS) (Tarboton et al. [2009]).

These parts – and the fact that they are shared along with and in addition to the software – contribute substantively to the success of CUAHSI HIS, by *stimulating innovation outside the triangle diagram*. For example:

1. Scientist-developers have built an independent implementation of Hydroserver for linux servers called ”Hydroserver Lite”, that now supplants HydroServer for projects in which server deployment cost is a factor (Conner et al. [2013]).

2. Some scientists have elected to extend ODM and build a new HydroServer around the extended data model, to either represent metadata that ODM does not handle, or to accomplish real-time data logging (e.g., Hersh and Maidment [2013] and Winslow et al. [2014]).

3. Other scientists have built their own data systems around extensions of the ODM concept (e.g., Masona et al. [2013]).

4. Several data publishers have ignored ODM and written their own WOFS compliant data services (e.g., McEnery et al. [2013] interfaced WOFS directly to HydroNexRad, while Anon. [2014b] exposes a variety of database formats as WOFS services). Most recently, this includes the NASA DataRods project to provide time-series representations of satellite data.

5. For several data sources in other formats, HIS team members and others have written WOFS ‘wrappers’ that deliver data originally published in other forms. Examples of this include USGS WQX and EPA STORET.

6. Many data users have written their own access WOFS-compliant data access clients for their own purposes. For this reason, access libraries now exist for languages including python and R (Horsburgh and Reeder [2014]).

7. Several data users have written custom WOFS clients including, e.g., the World Water Online data portal (Anon. [2014c]).

These development efforts are largely independent of the triangle diagram but share common attributes, including the ODM, WOFS, WaterML, and other features of CUAHSI HIS. Those projects that speak WOFS are harvested by the catalog and treated as if they are Hydroservers. Data models based upon the ODM can be published as WOFS through use of Hydroserver or its variants, independent of how data was collected. In this way, multiple, different implementations of CUAHSI HIS tools interoperate with one another through use of a common shared set of interfaces (ODM, WOFS, and WaterML).

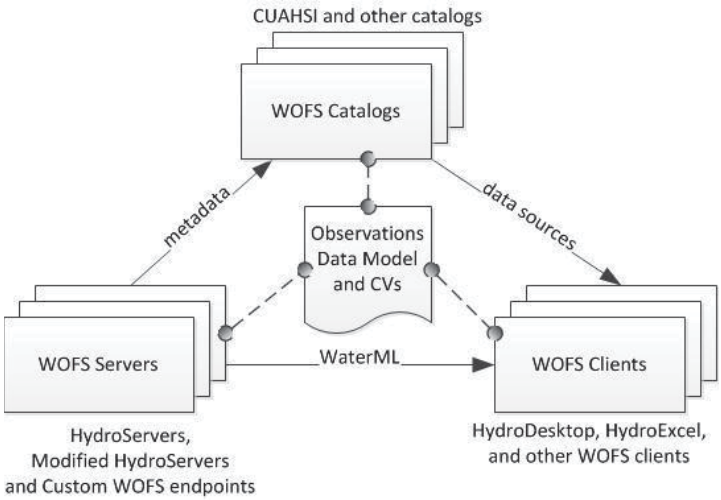

Figure 2. The “revised triangle diagram” depicts interactions of CUAHSI HIS components with standards, including ODM and controlled vocabularies.

In the current catalog, there are twenty registered services out of 100 that are not HydroServers, including seven from the critical zone observatories, Thus a more complete picture of the ”triangle diagram” is depicted in Figure 2, with ODM, WOFS, and WaterML providing the glue whereby users obtain flexibility. Thus, managing CUAHSI HIS is not just a simple matter of managing three software tools, but in fact, requires *managing a large ecosystem of software – not necessarily authored by the CUAHSI HIS team – and glued together by common standards and intents*.

In addition to CUAHSI HIS itself, these third-party deployed services and clients are also important and crucial, and are constrained by other social factors than just software. For example, many WOFS-like services in the WDC catalog are custom-implemented in order to interface with specific vendors' hardware, provide real time data, and to address other similar needs such as data recording outside the scope of ODM. Thus, one challenge is how to adapt these services and keep them responsive to changing needs, including new data models and standards, as these evolve and are adopted.

2 BEYOND MANAGING SOFTWARE

In the beginning, we described the Water Data Center as “managing software” and “doing the necessary engineering to turn CUAHSI HIS into a viable product”. One year into the data center's operations, it is more accurate to describe the Water Data Center as engaging in *user-centered engineering to manage, curate, and improve the user experience* of using the various interoperating parts of CUAHSI HIS. Forms of this curation include maintaining a rigorously vetted catalog, removing and potentially re-hosting failed data services, improving precision of metadata and searches, and improving basic user features including discovery and data publishing. We quickly learned that there are many alternatives to consider in curating the user experience, so we are evolving toward an agile model of software development to replace the more traditional software development models that we initially considered adequate to the task. The focus of development is not simply modifying software, but instead, modifying the software ecosystem by adding desirable components to replace formerly desirable but now outmoded components.

For example, we knew that the “ODM Data Loader” that most users employ to upload data into CUAHSI HIS had usability issues, because our most frequent user help request is for help with data uploading. We were also aware of significant security issues with its mode of operation, especially when users and ODM services are remote from one another on the internet. Also, we knew that there were sustainability issues with encouraging faculty to run their own data services; statistically, when the project is over, there are no resources available to sustain the data services and a data service can disappear permanently without warning.

What we did not initially acknowledge is that repairing these problems requires rethinking the basic overall design of publication and the data loader. The original data loader was optimized for use by expert users in an academic environment where one is running one's own data services. Our requests for help indicated that a fundamentally different kind of uploader was needed to interface with a new kind of data service: suitable for less technically adept users, and targeted at data publishing in a cloud outside an academic laboratory. We are currently in the process of testing and releasing the first version of this new cloud-based data loader and data service. Among other features, the cloud-based data loader provides much more user guidance on the process of uploading, and generates detailed error messages when uploads fail. The design and features of this uploader have been critiqued and revised by the CUAHSI User's Committee, giving the community direct input on the form and substance of this new service.

As another example, we knew the catalog's harvester had usability problems, but were unsure how to address them; the harvester would sporadically make mistakes in updating records for servers whose content had drastically changed since last harvest and correcting the problem would require human intervention. To repair the harvester, we had to first understand the information model of the catalog (which had long been buried in a complex data model for performance reasons) and then design a new information model – for the catalog itself – that addressed the problems. At this time, we are happy to report that what we have always claimed that the harvester does is now what it really does; the numeric identifiers of items constitute their identity, and one can change names of items without confusing the harvester. This harvester has been in production for several months as of this writing.

As these examples show, the software is not really the center of the WDC's mission. The WDC is primarily a service organization that:

1. Assists data providers in adapting to changing circumstances, including new standards for data publication.
2. Provides simplified, cost-effective ways to publish data without the assistance of IT staff, thus eliminating the need for researchers to be concerned with IT infrastructure
3. Assists with migrating older installations on aging hardware to modern, sustainable, and scalable IT architecture.
4. Provides a coherent look, feel, and access for data publishing tools, including single sign-on, web interfaces, etc.
5. Defines and provides tools to curate data and to ensure data and metadata integrity.
6. Defines and manages processes to engage and interact with the community and track changing user needs.

Software is simply a vehicle with which we accomplish these missions.

3 CHALLENGES AND OPPORTUNITIES

At this point in time, the WDC faces many challenges that are also opportunities to serve the community. Most of these involve reacting to changing community needs, including:

1. Providing WaterML2 services and supporting WaterML2 as a world standard for information exchange, supplanting existing WaterML1 services.
2. Providing OGC-compliant data services such as the Sensor Observation Service (SOS), Web Feature Service (WFS), and Catalog Services for the Web (CSW), as eventual replacements for WOFS and a first step toward becoming part of a global water data infrastructure.
3. Non-disruptively and gracefully deprecating data services that are no longer as useful as better alternatives.
4. Supporting the new version of ODM – ODM 2.0 – both with a catalog and appropriate data services.
5. Coordinating and managing consistency between data sources, including the formats in which data is available and the software versions utilized to expose it.
6. Reconciling the various metadata ontologies other than the one embodied in CUAHSI HIS, and leveraging domain-specific ontologies for chemistry, biology, and geology.
7. Curating data sources so that metadata is accurate and search results have high precision in locating data of interest.
8. Working with data providers to classify data according to hydrologic features, so that feature-based discovery becomes possible.
9. Working with researchers to codify and expose measures of data quality that determine whether data is suitable for an intended use.

Dealing with all of these challenges requires software changes, but more than this, coordination with and sensitivity to the needs of many human partners, each with unique constraints. Coping with these kinds of changes – on an international scale – requires engineering beyond maintaining a software stack. This requires considering the service lifecycle, and deciding when services will be instituted, revised, and deprecated. In turn, this can only be done with sensitivity and analysis of impacts to the whole ecosystem of interlocking CUAHSI and non-CUAHSI components that now make up CUAHSI HIS. For example, replacing WOFS with SOS is a gargantuan effort, not because it is difficult to do from a software perspective, but because a large number of valuable websites and tools – beyond CUAHSI – require WOFS to operate.

4 BEYOND THE ACADEMIC ECOSYSTEM

This engineering requires much more than simply managing existing software; improving the user experience often requires development of completely new approaches to satisfy user needs that were low priorities in the "academic ecosystem" in which the software originally developed. In turn, the real task of the WDC becomes to create, nurture, and manage a *new and complimentary software ecosystem aimed at usability and sustainability,* – in partnership with and in addition to the academic ecosystem in which CUAHSI HIS was originally developed – and in which a new set of (sometimes unexpected) players contribute to the task of curating the user experience, increasing reliability, portability, and sustainability, and making CUAHSI HIS usable to a broad variety of new kinds of users.

Some of our allies in this task are fundamentally different than the players in the original ecosystem. Some surprising and extremely active contributors include a government official, Silvano Pecora, in Italy's ISPRA (their analogue of the US Environmental Protection Agency), who made major strides in making CUAHSI HIS adaptable to other platforms, including Linux. This – in turn – allows countries who wish to do so to develop their own water catalogues; this was previously much more difficult and costly. Dr. Pecora's other innovations include adding conformance to European data access standards for both the catalog and data servers. This – in turn – brings to the table other potential collaborators who look to CUAHSI HIS as a potential solution to Europe's data management needs for time series.

Examples of new features already developed include testing frameworks for the catalog and data servers, cloud-based versions of data servers, a cloud-based data catalog with substantively improved performance, a consistent user sign-on procedure (using OpenID) for data and catalog services, user-friendly data upload for HydroServer, and a more robust harvester for catalog records. Features under current development in the immediate future include automatic harvesting of the new USGS water quality portal, instant and automatic distribution of new controlled vocabulary terms, funding source reporting to meet the needs of the National Science Foundation, a user-friendly web-based data portal for users who have not yet learned HydroDesktop, and interoperability with OGC and GEOSS data standards including the Web Feature Service (WFS) and Web Catalog Service (WCS).

While none of these features are "headline news" from an academic perspective, from a usability and sustainability perspective, we hope that they make the difference between "an interesting and promising prototype" and "a usable and robust product" suitable for worldwide deployment and use.

5 HARNESSING THE CLOUD

One of the defining features of the WDC is an emphasis upon cloud-based solutions rather than physical infrastructure. The CUAHSI Catalog currently runs entirely in the Microsoft Azure cloud, with a 10-fold performance improvement over use of physical machines; approaches yet to be deployed promise another factor of 10 increase. New HydroServers are being deployed inside the cloud, for higher reliability and sustainability. A general mechanism for cloud-based data publication is in testing and will be released soon, with user-friendly data uploading, informative error messages, and (as a next step) tools to aid in migration from physical infrastructure to the cloud.

A key aspect of cloud-based data publishing that is easy to overlook is that once data is in the cloud, the WDC itself manages service updates, bug fixes, and new service availability. Previously, when CUAHSI HIS changed, it took months to years for data providers to download and install the changes. Now that rollout can become instant. For example, once data is in the cloud, the WDC can roll out changes to the data publication software system wide, rather than piecemeal.

One of the tangible outcomes of our commitment to the cloud is that cloud computing makes the cost of data persistence quantifiable and objective. Before the cloud, the cost of data persistence was a jumble of local hardware costs and human labor costs. The cloud makes these costs tangible as a single figure – cents per gigabyte-month – that includes all of those factors. While this might seem expensive to the uninitiated, the cost of persistence includes such contingencies as hardware failure, data restoration after failure, staff time to perform data recovery in contingencies, and other intangibles. We have been paying these costs all along,

but they have never before occupied an explicit line item in our budgets.

Another tangible outcome of our commitment to the cloud has been a new understanding of the limits of a shared platform. Azure SQL is a powerful platform, but has limits based upon the fact that it is shared infrastructure. There is a limit upon the number of record changes in a single transaction that we never encountered in running on non-cloud hardware. This has required substantive software adaptation that we did not initially expect.

Cloud based data publication is not ”free”; we pay a monthly ”rental” charge for each gigabyte of data we store. The initial budget of the WDC includes ”rental” for up to 10 terabytes of data; this is an estimate of the amount of data currently indexed in CUAHSI HIS that is considered to be at risk of disappearing due to funding instability and/or instability of staff support to keep the data online.

6 Conclusions

The CUAHSI Water Data Center is a bold and evolving experiment. In this experiment, almost all of our challenges center around meeting user needs as best we can with limited resources. Many decisions need to be made, and for those decisions, we look to our governance boards: the ”CUAHSI Informatics Standing Committee” approves strategies, while the ”CUAHSI Users Committee” reports on usability issues and tactics. Through use of these two committees, as well as by providing user support, the WDC maintains a close relationship with its user constituencies and attempts to expend effort on initiatives that have most positive impact on our users.

In the short term, there is a large backlog of software development to do, and software development remains a central activity at the WDC. Looking farther ahead, we do not see an end to software development, but also believe that services other than software will become important. For example, active curation of data sources can achieve what software cannot: high precision in data discovery. At the end of the day, we are there to help. Please let us know what we can do for your water-related research.

Acknowledgments

This project was funded through the NSF cooperative agreement NSF-1248152. As well, this project would not have been possible without the wisdom and guidance of the whole CUAHSI HIS team over the last thirteen years. For this paper, we are especially grateful to long-time HIS team member Jeffrey Horsburgh, who provided us with references and some of the history on adoption of parts of the CUAHSI HIS model by others. We of the Water Data Center hope that these people continue to push the limits of water science informatics for years to come, and that we can provide a sustainable home for all of their innovations.

References

Ames, D. P., Horsburgh, J. S., Cao, Y., Kadlec, J., Whiteaker, T., and Valentine, D. (2012). Hydrodesktop: Web services-based software for hydrologic data discovery, download, visualization, and analysis. *Environmental Modelling & Software*, 37(0):146 – 156.

Anon. (2014a). OGC WaterML. http://www.opengeospatial.org/standards/waterml. Accessed: 2014-05-15.

Anon. (2014b). WOFpy – a python wrapper for WaterOneFlow services. https://pythonhosted.org/WOFpy/. Accessed: 2014-05-15.

Anon. (2014c). World water online. http://www.worldwateronline.org. Accessed: 2014-05-15.

Conner, L. G., Ames, D. P., and Gill, R. A. (2013). Hydroserver lite as an open source solution for archiving and sharing environmental data for independent university labs. *Ecological Informatics*, 18(0):171 – 177.

Couch, A. (2012). A CUAHSI datacenter for hydroinformatics: Draft specifications. Technical report, Consortium of Universities for the Advancement of Hydrologic Science, Inc.

Hersh, E. S. and Maidment, D. R. (2013). Extending hydrologic information systems to accommodate arctic marine observations data. *Deep Sea Research Part II: Topical Studies in Oceanography*.

Horsburgh, J. S. and Reeder, S. L. (2014). Data visualization and analysis within a hydrologic information system: Integrating with the r statistical computing environment. *Environmental Modelling & Software*, 52(0):51 – 61.

Horsburgh, J. S., Tarboton, D. G., Hooper, R. P., and Zaslavsky, I. (2014). Managing a community shared vocabulary for hydrologic observations. *Environmental Modelling & Software*, 52(0):62 – 73.

Horsburgh, J. S., Tarboton, D. G., Maidment, D. R., and Zaslavsky, I. (2008). A relational model for environmental and water resources data. *Water Resources Research*, 44.

Horsburgh, J. S., Tarboton, D. G., Piasecki, M., Maidment, D. R., Zaslavsky, I., Valentine, D., and Whitenack, T. (2009). An integrated system for publishing environmental observations data. *Environmental Modelling & Software*, 24(8):879 – 888.

Horsburgh, J. S., Tarboton, D. G., Schreuders, K. A. T., Maidment, D. R., Zaslavsky, I., and Valentine, D. (2010). Hydroserver: a platform for publishing space-time hydrologic datasets. In *AWRA 2010 Spring Specialty Conference*.

Maidment, D. R. (2005). Hydrologic information system status report. Technical report, Consortium of Universities for the Advancement of Hydrologic Science, Inc.

Masona, S. J., Cleveland, S. B., Llovet, P., Izurietaa, C., and Poole, G. C. (2013). A centralized tool for managing, archiving, and serving point-in-time data in ecological research laboratories. *Environmental Modelling and Software*.

McEnery, J. A., McKee, P. W., Shelton, G. P., and Ramsey, R. W. (2013). Hydrologic information server for benchmark precipitation dataset. *Computers & Geosciences*, 50(0):145 – 153. Benchmark problems, datasets and methodologies for the computational geosciences.

Tarboton, D. G., Horsburgh, J. S., Maidment, D. R., Whiteaker, T., Zaslavsky, I., Piasecki, M., Goodall, J., Valentine, D., and Whitenack, T. (2009). Development of a community hydrologic information system. In Anderssen, R. S., Braddock, R. D., and Newham, L., editors, *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*, pages 988–994.

Tarboton, D. G., Maidment, D., Zaslavsky, I., Ames, D. P., Goodall, J., and Horsburgh, J. S. (2010). CUAHSI hydrologic information system: 2010 status report. Technical report, Consortium of Universities for the Advancement of Hydrologic Science, Inc.

Whitenack, T. (2010). CUAHSI HIS central 1.2. Technical report, Consortium of Universities for the Advancement of Hydrologic Science, Inc.

Winslow, L. A., Benson, B. J., Chiu, K. E., Hanson, P. C., and Kratz, T. K. (2014). Vega: A flexible data model for environmental time series data. http://wordpress.gleon.org/wp-content/uploads/2012/03/Winslow_vega.pdf. Accessed: 2014-05-15.