AWRA 2012 SPRING SPECIALTY CONFERENCE
New Orleans, Louisiana
March 26 – 28, 2012

COMMUNITY PRACTICES FOR NAMING AND MANAGING HYDROLOGIC VARIABLES

I. Zaslavsky, D. Valentine, R. Hooper, M. Piasecki, A. Couch, A. Bedig[*]

ABSTRACT: This paper examines naming practices and encoding of hydrologic variables and presents challenges, approaches and technologies for aligning variable descriptions across observation networks. Management of hydrologic variable semantics is one of the key practical issues in the Consortium of Universities for the Advancement of Hydrologic Sciences, Inc., CUAHSI Hydrologic Information System (CUAHSI HIS) project. The project's mission is to encourage information sharing and reuse of hydrologic data from diverse sources. Over 16,000 variables are now harvested into the HIS Central metadata catalog, from several federal and state repositories, and multiple academic data sources. The meaning and construction of hydrologic variables, and the set of characteristics included in variable descriptions (name, measurement procedure, medium, units, offsets, etc.) differ across observation networks. These discrepancies lead to difficulties in the use of data for scientific research, complicating interpretation, variable discovery, and integration of multi-source data. In addressing these challenges, this paper reviews community practices for naming hydrologic variables and identifying data series. Several aspects of formalization and management of hydrologic variable semantics and hydrologic catalogs are considered, in particular standardization of hydrologic time series definitions.

KEY TERMS: Hydrologic Information System; hydrologic observations; cyberinfrastructure; semantics; vocabularies.

INTRODUCTION

Hydrologists collect multiple types of hydrologic indicators, in different subdomains of water, using a variety of measurement and recording methods. One of the key missions of the Consortium for Advancement of Hydrologic Science, Inc.'s Hydrologic Information System (CUAHSI HIS) (Maidment, 2008; Tarboton et al., 2010; Zaslavsky and Maidment, 2011) is to support and promote hydrologic research based upon diverse information sources. Use of diverse data sources requires a researcher to answer two basic questions:

- **Commensurability:** when do two or more distinct information sources represent the exact same measurement?
- **Invariance:** what attributes of a measurement can safely be considered to be invariant over all sources of that measurement?

These answers can be difficult to find, because the way that hydrologic data providers describe data has evolved over time in response to diverse social forces. Data providers commonly use different names for the same measurement, and make differing assumptions about what those names mean. This problem is exacerbated by the scale of available hydrologic data. At the time of writing, the CUAHSI HIS project registers over 80 observation networks from several federal and state repositories, and multiple academic and citizen research groups. Hydrologic observations metadata from each of the networks is harvested into the HIS Central metadata catalog (Whitenack et al., 2010), to enable cross-network discovery based on variable names and other measurement characteristics. The catalog now manages nearly 24 million hydrologic observation series, collected at over 2 million measurement locations. There are 16,000+ unique variable codes (a concatenation of a unique observation network code and a variable code that is unique within a network), and 12,400 unique variable names.

The ideal is that the above issues are addressed by the CUAHSI HIS data discovery services and user interfaces. However, the success of this depends upon the discovery system's ability to characterize a search in commonly used, high-level terms that emphasize commensurability and take into account variations in measurement methods and contexts.

The CUAHSI HIS catalog contents provide much evidence of a fairly low level of standardization and re-use of common hydrologic variable names across data sources. Variable names often contain a complete variable description rather than representing a core measured property. A descriptive variable name can incorporate measurement procedure, sample medium, units, features for which the measured property applies, and other characteristics. These properties can be organized into one or several descriptive elements, whose names may also differ across sources. Variable names that are not

[*] Respectively: San Diego Supercomputer Center, UCSD, zaslavsk@sdsc.edu, valentine@sdsc.edu; CUAHSI, rhooper@cuahsi.org; City University of New York, mpiasecki@ccny.cuny.edu; Tufts University, couch@cs.tufts.edu, Alex.Bedig@tufts.edu

accompanied by units, method or medium information or represent low-level instrument indicators (e.g. “voltage”) are also common. Even the name of the object of our analysis in this paper is elusive, as it is referred to as “parameter” by the United States Geological Survey (USGS, 2011), “characteristic” by the United States Environmental Protection Agency (EPA, 2011a), “element” by the National Climatic Data Center (2005), “variable” by CUAHSI HIS (Horsburgh et al, 2008; CUAHSI, 2011), and “observed property” by the Observations and Measurements specification of the Open Geospatial Consortium (Cox, 2010).

The established way to manage semantic discrepancies across sources is to map the different representations to a common information model and associated standard vocabularies. This approach has been undertaken in a number of earth science disciplines to enable cross-source data discovery and integration. Some of the key benefits of formal representation of semantic relationships and their use in domain cyberinfrastructure design include:

(1) richer and more flexible data discovery, as formally encoded hierarchical and synonymical relationships between terms are included in various forms of similarity-based search and semantic query rewriting/term expansion;

(2) a more convenient data organization and data management at all stages of data life cycle, as data publishers can group, label and link resources based on semantic relationships between datasets and data items – which in turn would lead to better search and data delivery performance;

(3) more consistent data annotation and interpretation as users converge on a common set of domain vocabularies;

(4) supporting community organization around common conceptual models of the domain and vocabularies/lexicons;

(5) with an appropriate OWL representation of semantic relationships, supporting reasoning about phenomena;

(6) more elaborate data management policies and governance mechanisms that rely on semantic relationships and validation/verification that emphasize semantic consistency.

Many of the benefits of semantic approaches have been already realized in a number of earth science projects. However, a larger-scale implementation success would depend upon community adoption of naming conventions and standards for describing measurements. Significant effort towards standardization of hydrologic variable descriptions has been made within the CUAHSI HIS project, and in several other projects in the hydrology domain (e.g. CSIRO, 2011). In particular, the CUAHSI HIS project developed an information model for observations collected at point locations, and implemented it in a standard relational schema, as the CUAHSI Observations Data Model, or ODM (Hosrburgh et al., 2008), and an XML schema called CUAHSI Water Markup Language, or WaterML (Zaslavsky et al., 2007.) A set of controlled vocabularies for characteristics of hydrologic observations, compliant with the ODM (CUAHSI, 2011) has been also created, to assist data publishers using HydroServer in populating their schemas with values that have consistent interpretation. For variable names, the project also created a hierarchically-organized vocabulary of hydrologic terms, and a procedure for associating variable names from multiple registered datasets with these terms (Piasecki and Beran, 2009). This hydrologic parameter vocabulary has been exposed via a web services interface that enables retrieval of vocabulary concepts and concept-variable mappings and supports term expansion.

While conceptual and technical approaches to semantic management in CUAHSI HIS have been addressed in other references, the scope of the problem -- in particular the established practices and differences in how hydrologic variable descriptions are published by different agencies and research groups -- has not been adequately examined. To fill this gap, this paper summarizes the different ways of naming hydrologic parameters as encountered in the CUAHSI HIS project, as the foundation for developing semantic management techniques for hydrologic data consistent with conceptual models of the domain. In the first section, we explore community practices for naming hydrologic parameters used by different groups. The empirical analysis of parameter names accumulated in the central HIS catalog is the basis for comparing key characteristics of variables and series from different repositories, as shown in the following section. We conclude with a summary and outline of next steps.

VARIABLE AND DATA SERIES DESCRIPTIONS IN HYDROLOGIC REPOSITORIES

Examples from Federal Agencies

The USGS National Water Information System (NWIS) and EPA Storage and Retrieval (STORET) and Water Quality Exchange (WQX) are among the key sources of hydrologic observations used by academic researchers. They provide different representations of “parameters” (in USGS terminology) or “characteristics” (in EPA terminology).

The USGS code list (USGS, 2011) has over 18000 terms to describe parameters (as of Aug 2011). Besides hydrologic variables, the list includes biological species, various informational, statistical and error codes, and other ancillary data. Hydrologic parameter descriptions typically include a variable name, a method, and units of the observation, but may also omit one or more of them. By contrast, EPA “characteristics” derive from Substance Registry System (SRS) (EPA, 2011b), which is, in turn, aligned with the Chemical Abstracts Service (CAS) registry (NIST, 2011), as well as with Taxonomic Serial Numbers for biological organisms. SRS includes nearly 10,000 terms, although many terms do not have data associated with them. Table 1 illustrates differences in variable naming between USGS and EPA, using several measures of calcium as an example.

Table 1. Examples of USGS parameter codes and EPA characteristic codes, and corresponding variable descriptions

Agency	Code	Corresponding name
USGS	00910	Calcium, water, unfiltered, milligrams per liter as calcium carbonate
	00915	Calcium, water, filtered, milligrams per liter
	00916	Calcium, water, unfiltered, recoverable, milligrams per liter
	91051	Calcium, water, filtered, micrograms per liter
EPA	75	Calcium
	749	Calcium carbonate
	16957	Calcium hydroxide as CaCO3

Since variable codes and descriptions have been added and evolved as the federal data systems developed, usage of codes and descriptions is not always consistent. For example, the CUAHSI Central metadata catalog contains 54 unique variable codes associated with the term “Discharge, stream”, of which 14 are retrieved from the Daily Values and Unit Values (Real Time) subsystems of NWIS, 6 are retrieved from the NWIS Instantaneous Irregular Data subsystem, 2 are retrieved from the EPA, and the rest come from various academic sources. NWIS contains all of the following variable names: “Discharge, cubic feet per second” (12 unique codes – the most common usage), “Discharge, instantaneous, cubic feet per second” (3 codes), “Discharge” (3 codes), “Discharge, instant” (1 code), “Instant. Discharge” (1 code). USGS real time stations report the “Discharge, cubic feet per second” parameter (and “Streamflow ft^3/sec” when delivered over recently deployed web services) - in this case, the “instantaneous” attribute is not included but implied by the subsystem. This same attribute may be present or not present in other variable descriptions.

Over the last several years, USGS and EPA have collaborated on developing associations between NWIS parameter codes and the SRS names, with the result that the two agencies agree on nearly 3500 codes. In the course of this mapping to SRS concepts, USGS variables have been organized into the root (concept, or characteristic name) part, and other components representing unit codes, sample fraction, as well as temperature, temporal, weight, particle size and statistical bases of the measured result. This is a critical development toward better structured variable information and hence more reliable variable-based discovery in particular through faceted search (Tunkelang, 2009.)

Compared to the above sources, the National Climatic Data Center (NCDC) manages a relatively small number of variables, which are typically encoded as 4-letter abbreviations, e.g. PRCP (“Daily Precipitation”), MXRH (“Maximum Relative Humidity”), or GAHT (“River Gage Height”) (e.g. NCDC, 2005.). At the same time, variable descriptions are comprehensive and include units, encoding details, instrument details, and rules for value assignment in special cases.

CUAHSI HIS and Representation of Variables in Academic Data Repositories

One of the key goals of the CUAHSI HIS project has been development of a comprehensive representation of hydrologic variables that can be used as a model for research groups that collect and publish hydrologic data. The CUAHSI ODM documentation refers to variable names as “…name of the physical, chemical, or biological quantity that the data value represents (e.g. streamflow, precipitation, temperature).” Variables have a number of standard components that are coded separately (Table 2). For consistency and to support data discovery across sources, the HIS project standardized some components of variable descriptions, providing curated controlled vocabularies for variable names, units, mediums, and data types. While conformance with several of these vocabularies was enforced at the database level, it was hard to enforce compliance for variable names. The reasons included a large number and complexity of legacy datasets being registered to the system, a common practice of overloading variable names with auxiliary information, and the need of encoding additional distinctive variable characteristics that did not fit in the set of fields prescribed by ODM, or were not mandated by ODM. Elevating such auxiliary information to variable name is often a way to emphasize that observation series differ in a certain aspect, such as vertical offset, a collection method, or a post-processing method.

As a result, organizations adopting ODM resorted to either recording a more complete descriptive name in the variable name field instead of using a concise term offered in the variable name controlled vocabulary, or – more often - provided a single-word term for the core concept as a variable name, leaving out other non-mandatory components of variable description. Continuing the “Discharge, stream” concept example: there were 11 academic observation networks that used a single term “Discharge” for the variable name, 8 other networks that used its synonym, “Streamflow”, while other networks used such terms as “Volume” or “Flow”. While synonyms are handled by the CUAHSI semantic system, incomplete representation of variable names would lead to potentially incorrect association between variables and concepts, and erroneous search results. However, we conclude that this is not a major concern for data published via CUAHSI

HydroServers. Of the variables named “Discharge” or “Streamflow”, only four variable codes have not been accompanied by unit information, and two codes have not included sample medium. This indicates a fairly high level of compliance of academic datasets with the data representation model advocated by the project.

Table 2. ODM variable components

Component	**Description**
Variable Name	The name of the variable
Variable Code	The unique code within a data source to identify a variable
Speciation	Refers to specific form of an element, with respect to isotopic composition, electronic or oxidation state, molecular structure, etc.
Sample Medium	Medium in which measurements are taken (surface water, groundwater, soil, etc.)
Value Type	The type of information (raw data, field observation, simulation, etc.)
Data Type	Statistic (Mean, Minimum, Maximum) and characteristic of the time series (Continuous, Sporadic, Incremental)
Units	According to controlled vocabularies for units
Time	This includes if an item is regularly sampled, and the spacing between samples.

CF Standard Names

The NetCDF Climate and Forecast (CF) standard name table specifies the possible values of the standard name attribute defined by the NetCDF-CF standard (CF Metadata Group 2011). The list presently includes over 2000 entries. The entries are defined in a specific pattern:

[surface] [component] standard_name [at surface] [in medium] [due to process] [assuming condition]

Examples of the terms are shown in Table 3. CF standard names are available as an xml document, which details the variable names, descriptions, units, and aliases, but not the information contained in the above pattern; it is expected that the patterns will be parsed for automated processing (Gregory, 2010). These patterns, and their mapping to structured representation of variables adopted by CUAHSI HIS, is of critical interest for hydrologists as CF-compliant NetCDF data are increasingly being used in hydrologic analysis and modeling.

Table 3. Examples from the CF Standard Names

air_temperature
air_temperature_at_cloud_top
runoff_amount
runoff_amount_excluding_baseflow
mole_concentration_of_water_vapor_in_air

COMPONENT NAMES IN KEY HYDROLOGIC REPOSITORIES: A COMPARISON

Table 4 summarizes key characteristics of variables and observation series as they are represented in USGS and EPA systems and shows how they are related to terminology and information model developed by the CUAHSI HIS project. While significant naming differences exist, it appears that one-to-one mappings are possible for most elements of the representations. These mappings are the basis of web service wrappers developed by the CUAHSI HIS project to provide access to USGS and EPA data from clients of CUAHSI water data services. This in turn provides the basis for unified variable discovery across government and academic data sources. Currently, this mapping is performed in the course of ingesting USGS and EPA catalog fragments into the CUAHSI metadata catalog from catalog or database dumps periodically provided by the agencies.

The following overall assessment is derived from the catalog ingestion and harmonization work performed in 2010. Of the total number of parameter codes included in the cross-walk table between USGS parameters and SRS characteristics (9178 parameter codes), 4339 parameter codes had publicly accessible associated data, and therefore could be found in the USGS catalog dump ingested in the HIS Central catalog. Of these, we successfully associated 2106 parameters with SRS concepts (SRS concepts are “root names” of SRS characteristics, there are 1141 concepts). A similar number of parameters (2046) had associated data but were not mapped to SRS concepts, as they mostly represented variables measured in sample mediums other than water or suspended sediment, or reflected atmospheric deposition measures, counting errors, uncertainties, or site characteristics. As SRS is further developed as a cross-walk mechanism between USGS and EPA parameters, we expect these numbers to change, and the mapping to become more accurate and complete.

Table 4. Key components of encoding hydrologic variables and series information

Component	**CUAHSI**		**US Geological Survey**	**US Environmental Protection Agency**	
	Observations Data Model	WaterML 1	NWIS (NWISWeb)	STORET (Web)	STORET WQX Services
Organization	Organization	Organization	Agency	Agency	OrganizationIdentifier
Site	Site	Site	Site	Station	MonitoringLocation
Location	Site>Lat-Long	Site>GeoLocation	Site>Lat-Long	Record>Lat-Long	GeographicMonitoringLocation
Variable	Variable	Variable	Parameter	Characteristic	CharacteristicName
Method	Method	Method	Associated with Parameter Name	Method	Method
Medium	Medium	Medium	Associated with Parameter Name	Media	ActivityMediaName
Series	Series	Series	Period Of Record		~Activity or Project
Sample	Sample	Sample	Sample	Sample	Result Lab Information
Precision	ValueAccuracy	ValueAccuracy	Reporting level	PrecisionValue	PrecisionValue
Bias	ValueAccuracy	ValueAccuracy	Reporting level	BiasValue	BiasValue
Confidence Interval	ValueAccuracy	ValueAccuracy	Reporting level		ConfidenceIntervalValue
Result Value	DataValue	DataValue	Column contains values	Value ValueText	ResultMeasureValue ResultMeasureText
Result Value Qualifier	Qualifier	Qualifier	Header Includes Multiple Qualifiers. Column contains values	Comment	ResultStatusIdentifier
Result Unit	Variable Units	VariableUnits	Associated with Parameter	MeasureUnit	MeasureUnitCode

CONCLUSION

This paper presented an empirical analysis and comparison of naming practices and conventions for hydrologic variables, across multiple sources of hydrologic data assembled in the CUAHSI HIS project. The differences across hydrologic repositories in terms of how variable names and other series information is encoded, lead to difficulties in the use of variables for discovery, data annotation and integration, especially where structured representation of variable information is expected. Common semantic data management techniques, such as those implemented in the CUAHSI HIS project, in particular a collection of controlled vocabularies for different components of hydrologic series descriptions, and mapping variable names to a community-driven concept hierarchy, provide a partial solution, which has been quite successful for academic data publishers following ODM conventions. However, additional effort is needed to disambiguate variable names – especially if key information is implicit or presented elsewhere in observation network description – and to harmonize variable descriptions across key legacy or non-compliant observation data repositories used by hydrologists. Potential steps in this direction would include crowd-sourcing of variable and series tagging and evolving the system to compliance with a common conceptual model and data encoding standards, with linked vocabularies. A key standard that the CUAHSI HIS project is transitioning to is Water Markup Language 2.0. WaterML 2.0 is being developed by a group of international experts organized under the Hydrology Domain Working Group of the Open Geospatial Consortium and the World Meteorological Organization, and its first part, focused on time series encoding, is nearing adoption as an international standard at the time of writing (OGC, 2012).

ACKNOWLEDGMENTS

We would like to express our appreciation to members of the CUAHSI HIS development team: assembling the central catalog which is examined here for variable name consistency is an effort of many people. NSF funding (awards EAR 0622374 and EAR 0949196) is gratefully acknowledged.

REFERENCES

CF Metadata Group. 2011. NetCDF Climate and Forecast (CF) Metadata Convention-CF Standard Names. Version 18, 22 July 2011 http://cf-pcmdi.llnl.gov/documents/cf-standard-names/about. Accessed January 2012.

Cox, S. 2010. Geographic Information: Observations and Measurements OGC Abstract Specification Topic 20, v2.0.0, OGC 10-004r3, Open Geospatial Consortium, Inc., 49 p., http://portal.opengeospatial.org/files/?artifact_id=41579. Accessed January 2012.

CSIRO. 2011. Formalization of a vocabulary - SKOS and RDF https://www.seegrid.csiro.au/wiki/Siss/VocabularyFormalization Accessed January 2012.

CUAHSI. 2011. Master Controlled Vocabulary Registry for ODM 1.1 http://his.cuahsi.org/mastercvreg/cv11.aspx. Accessed January 2012.

EPA. 2011a. WQX Domain Value Services and Downloads. http://www.epa.gov/storet/wqx/wqx_getdomainvalueswebservice.html. Accessed January 2012.

EPA, 2011b. Substance Registry System web services. http://iaspub.epa.gov/sor_internet/registry/substreg/home/overview/home.do. Accessed January 2012.

Gregory, J. 2010. Parsing CF standard names. http://www.met.reading.ac.uk/~jonathan/CF_metadata/14.1/. Accessed January 2012.

Horsburgh, J.S., Tarboton, D.G., Maidment, D.R., Zaslavsky, I. 2008. A relational model for environmental and water resources data, Water Resources Research, 44, W05406, doi:10.1029/2007WR006392.

Maidment, D.R. 2008. Bringing water data together, Journal of Water Resources Planning and Management, 134(2), 95-96, http://link.aip.org/link/?QWR/134/95/1. Accessed January 2012.

National Climatic Data Center, 2005. Data Documentation for Data Set 3210 (DSI-3210). Summary of the Day – First Order, May 4, 2005. http://dss.ucar.edu/datasets/ds509.0/docs/2005may.td3210.pdf. Accessed January 2012

Open Geospatial Consortium. 2012., WaterML 2.0 Standards Working Group, http://www.opengeospatial.org/projects/groups/waterml2.0swg. Accessed January 2012.

NIST. 2011. Chemical Name Service. http://webbook.nist.gov/chemistry/name-ser.html. Accessed January 2012.

Piasecki, M., Beran, B. 2009. A semantic annotation tool for hydrologic sciences, Earth Science Informatics, 2(3), 157-168, doi:10.1007/s12145-009-0031-x.

Tarboton, D.G., Maidment, D., Zaslavsky, I., Ames, D.P., Goodall, J., Horsburgh, J.S. 2010. CUAHSI Hydrologic Information System 2010 Status Report, Consortium of Universities for the Advancement of Hydrologic Science, Inc., Washington, D.C., http://his.cuahsi.org/documents/CUAHSIHIS2010StatusReport.pdf. Accessed January 2012.

Tunkelang, D. 2009. Faceted Search. Synthesis Lectures on Information Concepts, Retrieval, and Services, 1(1), 1–80, 2009. DOI: 10.2200/S00190ED1V01Y200904ICR005.

USGS. 2011. Technical Documentation of USGS Water-Quality Web Services http://qwwebservices.usgs.gov/technical-documentation.html. Accessed January 2012.

Whitenack, T., Valentine, D., Zaslavsky, I., Piasecki, M., Tarboton, D., Horsburgh, J., Whiteaker, T., Ames, D., Maidment, D.R. 2010. Hydrologic metadata catalog and semantic search services in CUAHSI HIS, Francisco Olivera (Ed.), 2010 AWRA Spring Specialty Conference: Geographic Information Systems (GIS) and Water Resources VI, American Water Resources Association, TPS-10-1, ISBN 1-882132-82-3.

Zaslavsky, I., Maidment, D.R. 2011. Service orientation in the design of a community hydrologic information system, In: Geoinformatics: Cyberinfrastructure for the Solid Earth Sciences, Edited by G. R. Keller and C. Baru, Cambridge University Press, p.193-209.

Zaslavsky, I., Valentine, D., Whiteaker, T. (Eds.). 2007. CUAHSI WaterML v0.3.0. OGC07-041r1. Open Geospatial Consortium, Inc. 76 pp. http://portal.opengeospatial.org/files/?artifact_id=21743S. Accessed January 2012.