HydroTagger: A Tool for Semantic Mapping of Hydrologic Terms

Michael Piasecki

Dept. of Civil, Architectural & Environmental Engrg., Drexel University, 3141 Chestnut Street. Philadelphia PA 19104
Michael.Piasecki@drexel.edu

Abstract

Semantic heterogeneity is a common problem when trying to unify information from disparate data systems. In order to overcome these heterogeneities it is generally understood that no one system or convention is the best, rather that mechanisms need to be implemented in which researchers can both agree on terms or keywords used and also the establishment of a keyword structure so their relationship to each other are known. This paper introduces an application developed for the hydrologic community that allows viewing and editing of the underlying keyword ontology without the need for specific knowledge of ontology editors like Protégé and also permitting the tagging of individual variable names to its relevant search keywords to support another application (HydroSeek). The underlying keyword ontology is a concept ontology that is used a base for searching for hydrologic data across various data sources.

Introduction

Semantic Heterogeneity is probably the single biggest challenge when trying to develop an environment in which researchers and scientists can query for specific data across national, regional and local data sources [Ref1]. While syntactic differences are important also and certainly access trajectories into the respective data sources (databases holding point oriented time series data), semantic disparities are the hardest to tackle. Semantic differences arise at several levels and concern the existence of hyponyms, synonyms, and polysemes. In linguistics, a hyponym is a word or phrase whose semantic range is included within that of another word, for example “Groundwater Level”, “Stage Height”, and “Reservoir Level” are all hyponyms of “Water Level”. A synonym on the other side refers to the situation where two words have identical or at least similar meanings, for example “Stage Height” and “Gauge”, while a polyseme refers to a word that can have multiple meanings like “stage”.

In response to the strongly voiced need of the hydrologic community (as the result of a survey [Ref2]) to provide “ … an easier access to data …” the Consortium of Universities for the Hydrologic Sciences (CUAHSI) hydrologic information science (HIS) group is developing an information system that has a service oriented architecture at its base called WaterOneFlow [Ref3]. It is, among other components, comprised of a set of 3 core web-services whose signature is identical across data sources and that serves as the backbone of end user applications that make use of these web-services. One such application is HydroSeek, see Figure 1, that allows users to search via a concept keyword across disparate resources without having to know the exact nomenclature of the variable_name or variable_code used by that organization to store its data [Ref4].

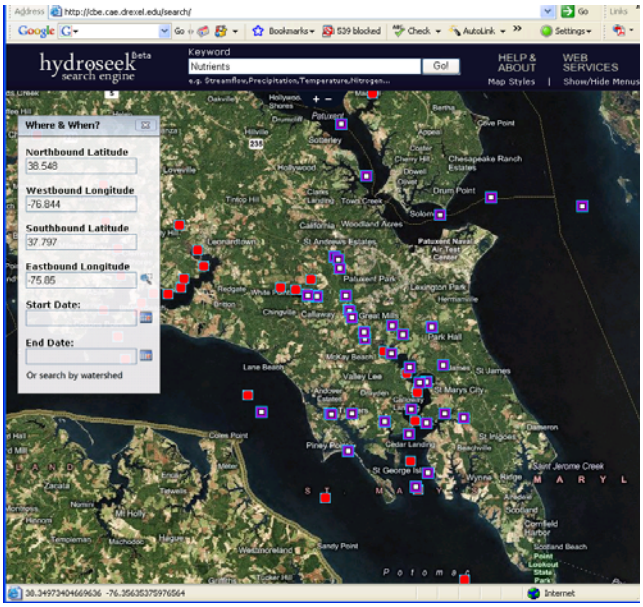

Figure 1 Screenshot of HydroSeek displaying stations that hold Nitrogen for an area in the mid-section of the Chesapeake Bay. Hollow squares are Chesapeake Information Management Systems CIMS and red squares are EPA STORET stations.

The key to this system lies in the realization that semantic heterogeneity is a given fact that cannot be changed and the resulting conclusion that semantic mappings need to be established that mediate the differences in data descriptions. The mediations are carried out within an ontology structure that permits the search via more general concepts like “nutrients” or “macronutrients” rather than specific variable IDs and codes used by each of the data providers to describe specific occurrences of nitrogen or phosphorus.

Ontology Structure

The concept system consists of an ontology that contains various layers that move from general concepts at the top to more granular or finer concepts at the bottom. The concept tree ends in leafs that are quite specific and against which the variable names in the data sources are mapped. Figure 2 shows an example in which the general realizations of nitrogen are at the bottom and how it is connected to higher level concept classes. The term Nitrogen or Ammonia for instance is hardly ever used by any of the data sources to name a collected parameter because it often depends where it is measured, whether it is filtered or not and so on. Hence, parameter names either tend to be a lot more specific or have synonymous names like "NitConc" instead of "Nitrite Concentration" , or "NO3ConcBot" for NO3 concentrations measured at the bottom. Consequently, "Nitrogen or Ammonia measurements" are common and well understood terms that lend themselves to be used as a search keyword rather specific nitrogen variable names or IDs that are sometimes being very cryptic.

The keyword ontology (we used the Global Change Master Directory [Ref5] as a start point) currently holds about 250 concept leafs to which approximately 750 parameter names are mapped. It starts at the top concept of "HydroSphere"

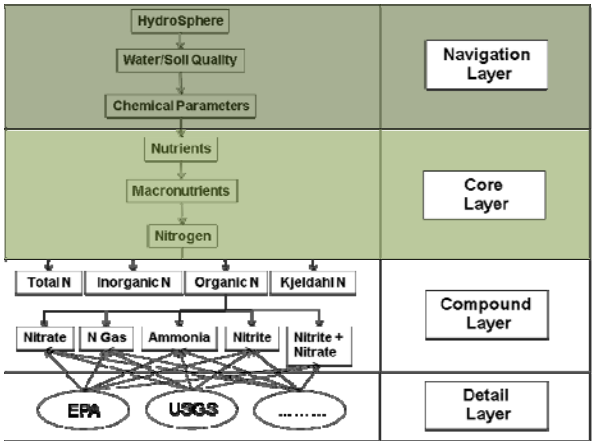

Figure 2 Ontology Layer Structure shown with the example of Nitrogen

and then branches into "Water/Soil Quality", "SubSurface Hydrology", "Surface Hydrology", Atmospheric Hydrology" and "Supplemental". When traversing the ontology a total of four different types of layers appear in the ontology, as shown in Figure 2: a "Navigation Layer", a "Core Layer", a "Compound Layer" and then the "Detail Layer". The top level is called "navigation layer" that provides structure to the ontology but whose class names cannot be used as keywords in the search process because they are considered too broad causing a too many parameters to be returned. Below this is the "core layer" whose class names are represent the first pool of permissible keyword entries for the search. Below that is the compound layer (like "nitrogen" whose classes are more detailed but broad enough to be used a search keyword. Below that is the "detail" which actually holds the variable IDs of each of the data sources. These entries are not permitted in the search either because it would require the knowledge of an overwhelming number of names and codes, the avoidance of which is the primary motivation for building this search engine in the first place. Current limitations of the system are mostly associated with the number of available classes against which variable and parameter can be mapped; EPA's STORET and USGS' National Water Information System, NWIS, alone hold about 19,000 different variable codes. While not all of them are relevant to hydrologists more need to be made accessible through the ontology.

Mapping Application

Because an ever increasing number of data sources is added to the HIS information system more and more data types and heterogeneous variablenames will need to be resolved in addition to covering a wider data spectrum. This poses a challenge in that it requires the involvement of the community because i) the parameter_name ⇔ concept_mapping is often ambiguous and can only be carried out by the researcher himself and ii) the concept ontology needs continued expansion that cannot be supported by a limited number of administrators. In other words, it is preferable to supply a system that permits the users of a network to execute mappings, i.e. providing the linkages between the "Compound Layer" and the "Detail Layer", on their own and also to suggest new concept or class entries into the ontology in case they have parameters that are not well or even not at all represented.

To this end a mapping application was developed, HydroTagger (see Figure 3), that works in close unison with the search mechanism. The HydroTagger, like HydroSeek, is a back end application that makes use of the web-services developed by the CUAHSI HIS group, and a commercial product that allows viewing of ontologies in an hyperbolic StarTree environment. It is an application that is launched at the StarTree viewer home site (in this case the San Diego Supercomputer Center) but corresponds with database entries that reside at Drexel University supporting the HydroSeek application. The HydroTagger controls functionality, i.e. the execution of the actual mappings, the approval of suggested mappings, and the request for new concepts to be added to the keyword ontology. It also includes an administrative interface through which the HydroTagger administrators can approve applications for the user list, and also promote certain users to administrators with limited access rights.

This latter feature is a result for the need to permit users to submit their data at a specific CUAHSI network node (which administers its own vocabulary), map it to a concept that is suitable to them (but not necessarily to researchers at other nodes) and also to participate in expanding the concept tree in case the network participants cannot find a concept that suits their needs.

The basic operation of the HydroTagger consists of a number of software installations and applications that work in unison. At the base is an Observations Data Model (ODM) [Ref6] that stores the time series data. We have added to the basic database design a number of tables that establish the linkage of source specific variable name (detail layer) to a specific leaf concept (compound layer), that keep track of what new linkages have been suggested. In addition, the HydroTagger also operates a "sniffer" application that trawls through the network nodes at specified time intervals to interrogate them for the newest

It should be noted that the use of a StarTree viewer is not fully compatible with viewing ontologies because of the possibility that child classes can have multiple parent classes and such does not represent a completely hierarchical system. Since the StarTree viewer cannot digest OWL files either we have developed an application (running in the background) that i) converts the OWL based ontology into a format that the viewer can read, and ii) breaks up the multiple child parent relationships by displaying a certain child concept multiple times in the StarTree, i.e. whenever there is a parent present in the branch system.

Use Scenario

The HydroTagger is an application that can be invoked using the WWW and is accessible to anybody (unrestricted portion) who would like to test and "play" with the system, and then a restricted section that need user registration and

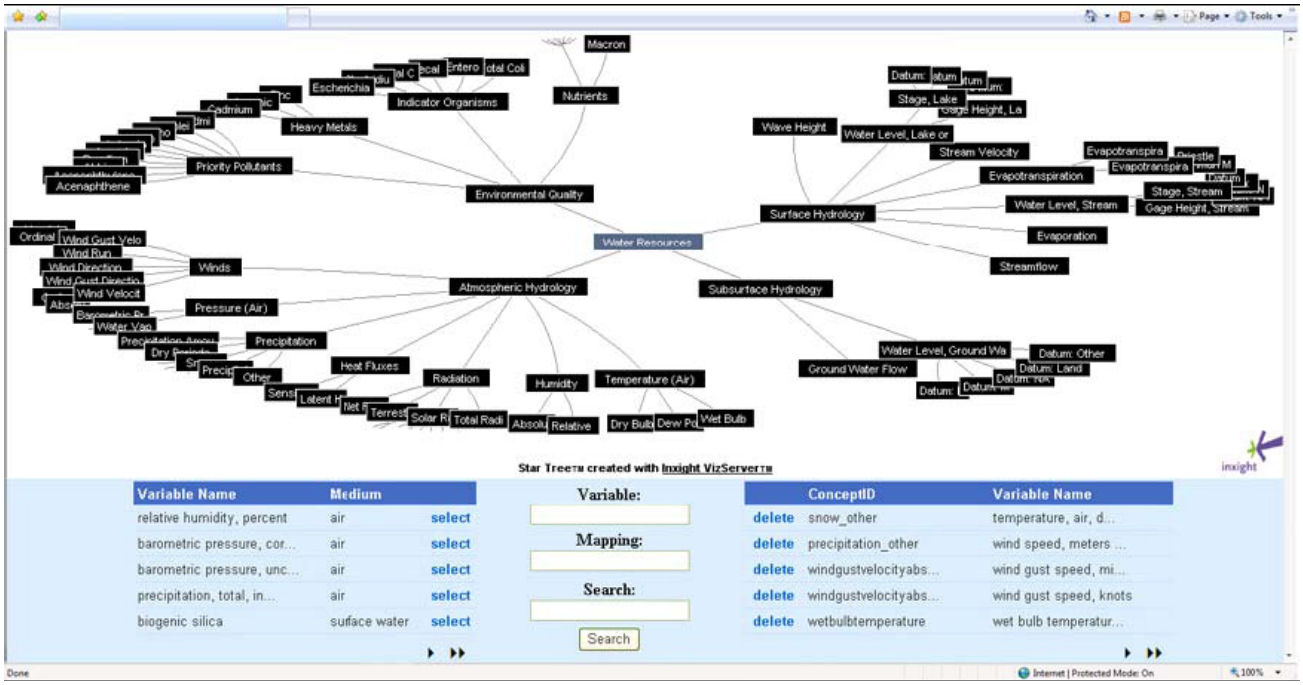

Figure 3 HydroTagger Interface showing the Ontology (upper panel) and variable names, tagging section, and mapped varibalenames <=> concepts in the lower panel

additions both in terms of data expanse (new data from a certain censor has been accumulated for the specified time interval) and new data variables added (a certain parameter is no also being reported on the network node, which needs to be added to the ODM tables and then mapped to a concept in the keyword ontology). The trawling action (carried out automatically through the scheduler service of MicroSoft XP) is a necessary support action i) to keep the search engine (HydroSeek) updated, and ii) to provide the HydroTagger with new variable names that require mappings.

approval. Modifications, suggestions, and mappings carried out in the latter section are migrated to the underlying databases of the two Hydro applications. A typical user scenario would look like the following sequence: 1) the "sniffer" trawls around the registered sites to see if new variables have been added in any of the sites (making use of the WaterOneFlow web-services). 2) Newly found variables (codes) are listed in a PendingMappingsTable awaiting the mappings. 3) The data administrator of the respective network node then logs onto the HydroTagger and will see only those variable

codes that the "sniffer" detected on his network. 4) The administrator then looks for a leaf class (concept) to which he would like to map the variable code in question. 5) If he cannot find an appropriate leaf class, he can then map it to a "wild card" class and is then asked to provide a class code (plus a name) and at the specific location he requested the wild card class. 6) This new class (concept) suggestion is then passed on to the ontology administrator who then approves or rejects it. Specific network administrators are in control of adding new users to their user group, i.e. individuals that are permitted to map variables for these networks and also suggesting new classes and concepts.

Future Outlook

The HydroTagger currently supports two functions; on the one side it can be used to execute mappings, and on the other side it can be used to expand the ontology at the same time providing an extremely efficient viewing feature. The latter functionality currently is somewhat rudimentary and needs to be expanded in order to make it a more suitable application for adding concept classes at all levels and also making sure to establish all parent relationships. Particularly the latter is not straight forward as one needs to work backward from a truly hierarchical viewing system. We are also planning on making this application available to researchers and scientists (possibly during dedicated workshops) to further build up the keyword ontology so that it can encompass a larger variety of data sets possibly from different communities other than just hydrology and environmental engineering. Given the fact that this system presently has only about 250 concepts represented for a total of about 750 registered variable names (or codes) it is clear that that this ontology is in need to continuously grow at a rate that is determined by participating researchers and scientists.

Acknowledgements

The work presented in this paper was sponsored by the National Science Foundation through grant numbers 0609832 and 0412904. The author would also like to acknowledge Bora Beran for the substantial amount of development work invested into this project, Ilya Zaslavsky at the San Diego Supercomputer Center for making available their serves and the StarTree environment, and also the Consortium for the Advancement of the Hydrologic Sciences, Inc. (CUAHSI) for their support and contribution in developing the ideas in this manuscript.

References

[Ref1] Marine Metadata Initiative MMI, Marine Metadata Interoperability, accessed February 2008, http://marinemetadata.org

[Ref2] Consortium of Universities for the Advancement of Hydrologic Science, Inc, CUAHSI, Technical Report #2, CUAHSI Hydologic Information Systems, August 2002, http://www.cuahsi.org/publications/cuahsi_tech_rpt_2.pdf

[Ref3] Whitaker, T., Tarboton, D., Beran, B., Valentine, D., To, E., Min, T., (2007), "CUAHSI WaterOneFlow Workbook V. 1.0), accessed October 2007 http://www.cuahsi.org/his/documentation.html

[Ref4]Beran B. (2007), HYDROSEEK: An Ontology-Aided Data Discovery System for Hydrologic Sciences, Ph.D. Thesis. 155 pp., Drexel University, Philadelphia, 5 September.

[Ref5] GCMD (2006), Global Change Master Directory, a directory to Earth Science data and services, NASA Goddard Space Flight Center, accessed February 2006, http://gcmd.gsfc.nasa.gov/index.html

[Ref6] Horsburgh, J. S., D. G. Tarboton and D. R. Maidment, (2005), "A Community Data Model for Hydrologic Observations, Chapter 6," in Hydrologic Information System Status Report, Version 1, Ed. by D. R. Maidment, p.102-135, http://www.cuahsi.org/docs/HISStatusSept15.pdf.