Earth Sci Inform (2009) 2:157–168
DOI 10.1007/s12145-009-0031-x

RESEARCH ARTICLE

A semantic annotation tool for hydrologic sciences

Michael Piasecki · Bora Beran

Received: 17 March 2009 / Accepted: 29 July 2009 / Published online: 9 August 2009

Abstract Semantic annotations are playing an increasingly important role in the world of metadata, more specifically when dealing with semantic heterogeneities between information systems. The need to bring together disparate data sources (in terms of syntax and semantics) so they can be searched simultaneously from a single search environment has become one of the most challenging tasks in developing information systems that span multiple communities as is common in the geosciences. The key problem lies in the legacy information systems, in which, at the time of development, each system used (and continues to use) its own semantic framework to identify variable codes and names, as well as annotating the collected data with metadata. This lack of a common metadata framework as well the uncoordinated use of descriptors and controlled vocabularies has led to a situation in which synonyms and hyponyms abound. Experience has shown that a centralized system with just one vocabulary for all is not feasible. Rather, in order to overcome these discrepancies it is important to realize that heterogeneity is an inevitable aspect of the scientific data world that needs to be accommodated. This paper describes the development and end use of an application that is designed to connect arbitrary variable names to specific concepts in layered search ontology. We will demonstrate the utility of this application through its deployment for the Consortium for the Advancement of Hydrologic Sciences Inc. (CUAHSI) network of testbeds and report on the issues that emerged carrying out variable and concept tagging. These issues concern specificity of a concept, ancillary information needed when identifying proper ontology locations, and multiple appearances of variables at different locations.

Keywords Concept tagging · Metadata · Ontologies · Semantics

Introduction

The Consortium of Universities for the Advancement of the Hydrologic Sciences Inc. (CUAHSI, http://www.cuahsi.org) and its Hydrologic Information Science Group (HIS, http://his.cuahsi.org) has been developing CyberInfrastructure (CI) for the hydrologic community. This effort has been focusing on developing an information system for the community that would bring together data collected in academia as well as in governmental institutions, such as the US Geological Survey (USGS) and Environmental Protection Agency (EPA) under one umbrella. The initial focus has been on point based time series data for which the group developed a data model (Observations Data Model, Tarboton et al. 2007) that is being implemented in a relational database in addition to a number of peripheral applications that would enable the user to load (ODM Data Loader, Horsburg and Berger 2008a), Streaming Data Loader (Horsburg and Berger 2008b), access and inspect (HydroEXCEL, Whitaker 2008; HydroGet, To and Whitaker 2008) and also query and upload data using a map interface (HydroSeek, Beran and Piasecki 2008). The

Communicated by: H. A. Babaie

M. Piasecki (✉)
Department of Civil, Architectural & Environmental Engrg,
Drexel University,
3141 Chestnut Street,
Philadelphia, PA 19104, USA
e-mail: Michael.Piasecki@drexel.edu
URL: http://www.pages.drexele.edu/~mp29

B. Beran
eScience Research Group, Microsoft Corp.,
455 Market Street, Suite 1690,
San Francisco, CA 94105, USA
e-mail: borabe@microsoft.com

key to these developments has been the creation of a national water information catalogue that can be accessed via service oriented architecture, called WaterOneFlow (Whitaker et al. 2007), that uses SOAP web services (http://www.w3.org/TR/soap12-part1) to publish the information catalogue's contents.

One the most difficult challenges in compiling a national water information catalogue, i.e. a catalogue that stores metadata about water data at numerous and disparate water data bases in a uniform format, has been to overcome syntactic and semantic heterogeneities that exist across these repositories. While the syntactic unification has been achieved through the creation of WaterML (Zaslavsky et al. 2007), an eXtensible Markup Language (XML, W3C 2006) schema that defines a standard format in which water data is being transmitted, semantic mediation is a somewhat harder problem to deal with because the meaning of words and their intentions is subject to interpretation and as such does not provide a framework in which normative statements concerning the correctness or faultiness of definitions and labels can be made easily. In response to this challenge, the HIS group developed an approach to overcome semantic heterogeneities that has led to the development of a map based search engine, HydroSeek, in which users can query the national water data catalogue by using concepts or keywords defined in a search ontology.

The use of ontologies for semantic mediation and annotation is gaining more and more recognition in the area of earth sciences, and in fact are far too numerous to list here. However, some efforts stand out because of their scope such as the Marine Metadata Initiative (MMI 2009) that seeks to collect and host a number of oceanographic ontologies addressing sensor platform descriptions and controlled vocabularies on term definitions (for example the Climate and Forecast, CF, conventions for the netCDF data format, http://cf-pcmdi.llnl.gov/, and the British Oceanographic Data Center, BODC, definitions) and also features ontological implementations of the International Standard Organization (ISO) metadata frameworks. There are also a number of earth science keyword collections implemented as OWL ontologies, i.e. the Global Change Master Directory (GCMD 2009) and the Semantic Web for Earth and Environmental Terminology ontology (SWEET 2009) providing an upper level representation of keywords of the earth science realm in addition to general components like a units representation. The semantically enabled science data integration (SESDI 2009) and virtual solar terrestrial observatory , VSTO http://vsto.hao.ucar.edu/, (Fox et al. 2008) are efforts in the atmospheric sciences while the SPIRE (http://spire.umbc.edu/us/) project and the ecological ontologies developed at the Information Technology and Systems Center (ISTC 2009) at the University of Alabama at Huntsville are projects in the area of ecoinformatics that deploy ontologies for semantic mediation and annotations to address the descriptive heterogeneities among different data sources. Ontologies have also been used to support data discovery services such as deployed with the GEON grid portal (GEON 2009) and NOESIS (http://noesis.itsc.uah.edu) which is also used by the Linked Environments for Atmospheric Discovery (LEAD 2009) project to provide access to meteorological data.

The HIS ontology, portions of which have been inspired through some of the above mentioned efforts most notably GCMD and SWEET, defines concepts arranged in a tree like structure that starts out with the root concept "HydroSphere" at the top (most general concept) and then traverses across the various branches to more and more specific concepts until the leaf (or core) level is reached. At the leaf level it defines concepts that are just slightly more general in nature than typical parameter names as defined by USGS National Water Information System, NWIS, or EPA STORET (and others). For example, the search ontology contains a leaf concept "Nitrate" to which all nitrate variables collected and defined at the original sources have been associated with or "tagged to". In this example, the concept "nitrate" currently has 26 different nitrate variables tagged to it, which are stored in the central water information catalogue along with the rest of the water metadata. In other words, whenever a user chooses the search keyword (or leaf concept) "nitrate" the global search will be spawned over all registered data sources that contribute to and make up the group of the 26 nitrate variables. It is beyond the scope of this work to further describe the HydroSeek search engine and the reader is referred to Beran and Piasecki (2008) for more details.

For a system such as HydroSeek to function, however, a number of auxiliary applications need to be in place to support the underlying search framework. In other words, there is the need to update the underlying information database (the national water metadata catalogue) in addition to having a system that permits data managers of participating data sources to assist HIS efforts in appropriately tagging their variables to corresponding concepts in the keyword ontology (tagging here means: create a variable name ⇔ concept pairing). The experience of the HIS group has been that this it is not a straight forward task because definitions and interpretation of word meaning remains subject to a great deal of subjectivity for which it is difficult to find and define a commonly accepted concept.

When developing the semantic mediation approach it became clear that the effort needed an interface that would allow active participation of registered data sources so they could tag variables they had collected to concepts presented to them in the keyword ontology. This led to the development of the HydroTagger. The purpose of this application is to allow the graphically supported tagging of variables defined at participating data sources to keyword concepts in the keyword ontology, i.e. the creation of "concept ⇔ variable" pairs. These pairs are stored in a lookup table which in turn is used by discovery tools like HydroSeek to find the variables the user is looking for. While the operational version of this application is part of the HIS Central Registration (http://hiscentral.cuahsi.org) and as such subject to access restrictions, a public version exists at http://www.hydrotagger.org where the application is made freely accessible. This paper will outline and report on the design needs for this application and highlight the underlying concepts that form the foundation of this tagging tool.

Ontology layers

Before outlining the tagging strategy and expanding on the HydroTagger application, it is helpful to outline some of the ontology aspects which are important when examining the tagging approach.

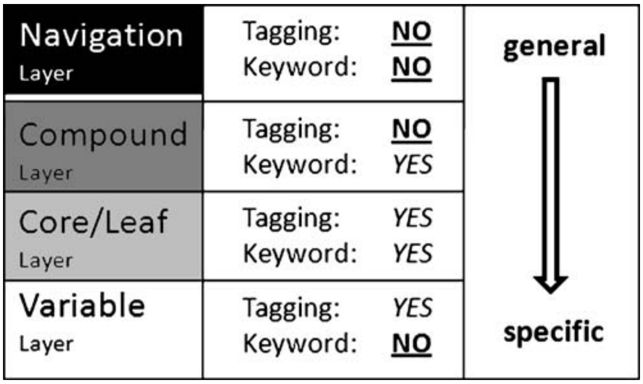
Ontology structure

The ontology was designed with the single purpose of discovery of data in mind. This meant an adoption of a keyword structure that would organize data variables along thematic classifications going from general concepts to finer concepts. The initial structure was a conglomeration of classification approaches as used in the USGS National Water Information System, NWIS, (USGS 2008), the STORET system (EPA 2009), and also NASA's Global Change Master Directory, GCMD, (NASA 2009) all of which contributed a subset of concepts to the existing HydroSeek ontology. The SWEET ontology (Raskin 2009) was also examined as a potential start point but because its hydrology section contained only a limited number of concepts relevant to the effort at the time in addition to providing a mixed assembly of concepts describing processes, locations, data types, and also some more general areas of data collections the decision was made to start anew.

A key aspect of the design was to avoid the so-called low-precision high-recall (too much returned) or high-precision low-recall (to little retuned) problems prompting the idea to find a reasonable middle ground in terms of "specificity vs general enough". The basic idea behind the solution to this problem was to only permit a subset of the keyword collection for search purposes by defining an internal boundary that would separate concepts considered too general higher up in the ontology form those further down that are considered permissible. The resulting ontology (version 1.0, to be viewed at http://hiscentral.cuahsi.org/startree.html) is thus organized in four layers, as shown in Fig. 1, each of which serves a different purpose. The differences lie in their ability to be used as the layer to which variable names can be tagged layer (leaf layer only) and whether or not the layer's respective concepts are permitted as a search keyword (leaf and compound layers). The top level (most general and called "Navigation" layer) is the backbone layer that besides hosting the root concept "HydroSphere", provides a first classification along divisions of mostly 'where' (land, atmosphere, surface, groundwater, etc) and one 'what' (water/soil quality), as shown in Fig. 2. This layer is not accessible to the user in any way, i.e. it cannot be used as a source for search keywords nor are the concepts open to tagging.

The next layer is the "Compound" layer which has a greater emphasis on 'what' rather than 'where'. The concepts in this layer are used to further break down more general concepts such that a user can traverse the concept structure en route to the specific type of data the user is interested in. This layer at its upper bound, i.e. connections to the top layer, provides the most general entries from the pool of permissible keywords that are offered up during the search. However, the user is not permitted to use any of these keywords as tagging concept, i.e. a concept to register a data set with. This layer while traversing several branches, is not uniform in its 'thickness' as there may be just one branch in some places (for example the concept "nobelGases", which is a subclass of the Navigation Layer concept "chemicalParameters") and up to 4 in others (for example the chain "nutrients" => "macronutrients" => "nitrogen" => "ammoniaNitrogen", which is also a subclass of the "chemicalParameters" class in the Navigation Layer).

The ontology has been extended since its inception in late 2006 and now hosts about 400 concepts (total) of which 323 are leaf concepts and of which 128 have been used to register variables (the delta is largely due to many leaf concepts that



Navigation Layer	Tagging: **NO** Keyword: **NO**	general
Compound Layer	Tagging: **NO** Keyword: *YES*	⇩
Core/Leaf Layer	Tagging: *YES* Keyword: *YES*	
Variable Layer	Tagging: *YES* Keyword: **NO**	specific

Fig. 1 The 4-layer ontology structure stacked from top to bottom with increasing specificity. Each layer has 2 attributes: whether or not it can be used as search keyword, and whether or not it can be tagged to

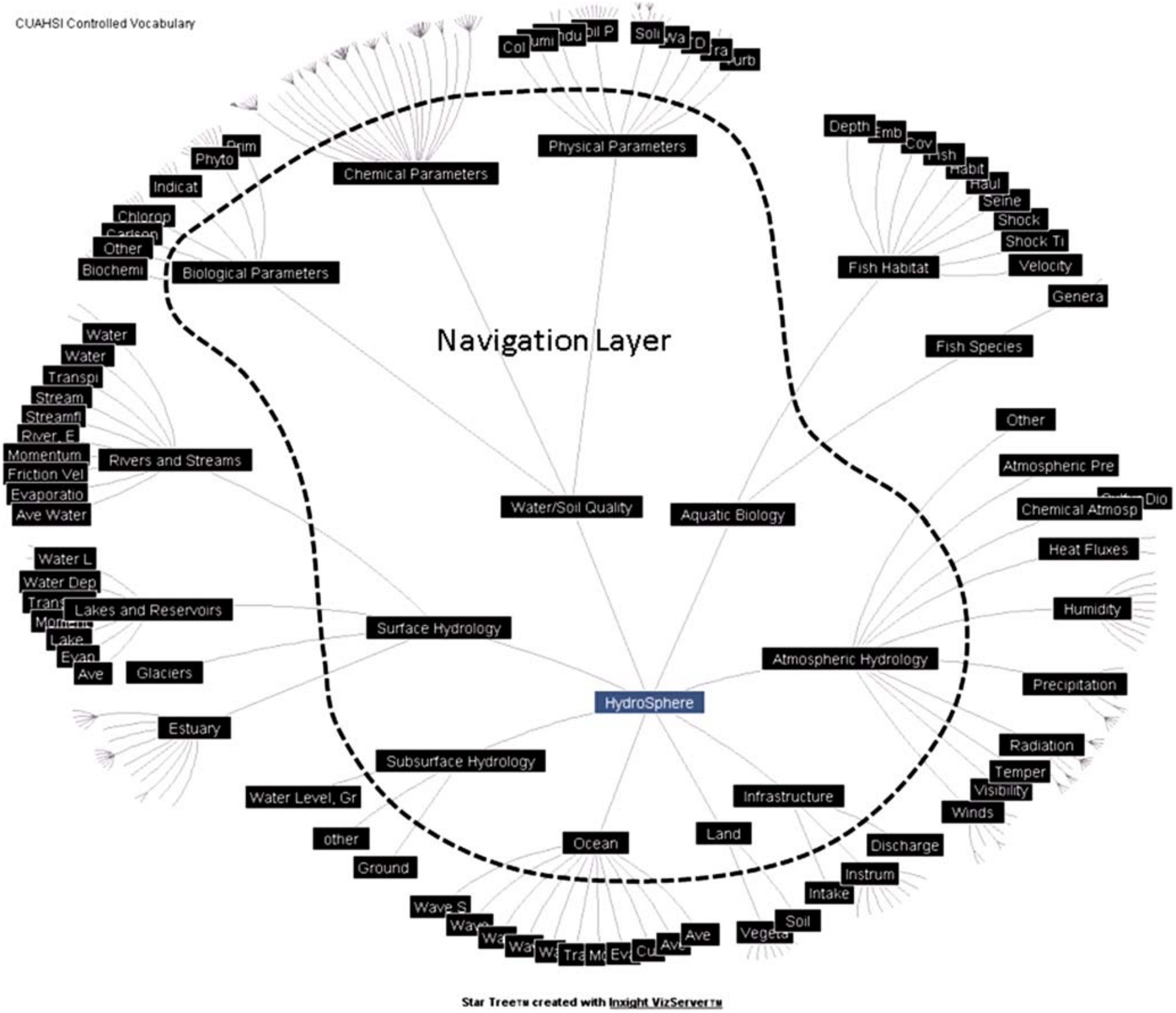


Fig. 2 Hyperbolic StarTree visualization of the top (*Navigation*) layer of the current search ontology. It provides the backbone for the entire search keyword structure

were introduced to accommodate the variableName controlled vocabulary of the Observations Data Model DataBase and as such are not all in use). About 30 leaf concepts have multiple parents, and for 50 variables a list of synonyms has been created that can be used alternatively to conduct the keyword search. The ontology also covers the top 500 parameters of the USGS National Water Information System parameter list (total of about 9,000) and with this addresses about 90% of the data holdings (records volume) of the NWIS database. The ontology is realized as a multi-file OWL (Web Ontology Language, W3C 2007) collection in which the 3 top layers (Navigation, Compound, and Core/Leaf) are stored in separate files for each major branch as defined in the "Navigation" layer thus providing a horizontal (referring to a layer) and vertical (referring to each major branch) structure making it easier to access and edit specific sections in the ontology construct. The ontology is freely accessible and can be downloaded from http://www.hydroseek.net/ontology/.

Tagging concepts

Leaf layer

The last layer of the ontology is the "Core" or "Leaf" layer which consists of all concepts at the end of any branch in the ontology. It has by this definition only one concept for each end of a branch, and represents a concept that is quite specific, i.e. specific enough so that it can be used to tag (or register) a variable name with it. As shown in Figs. 2 and 3

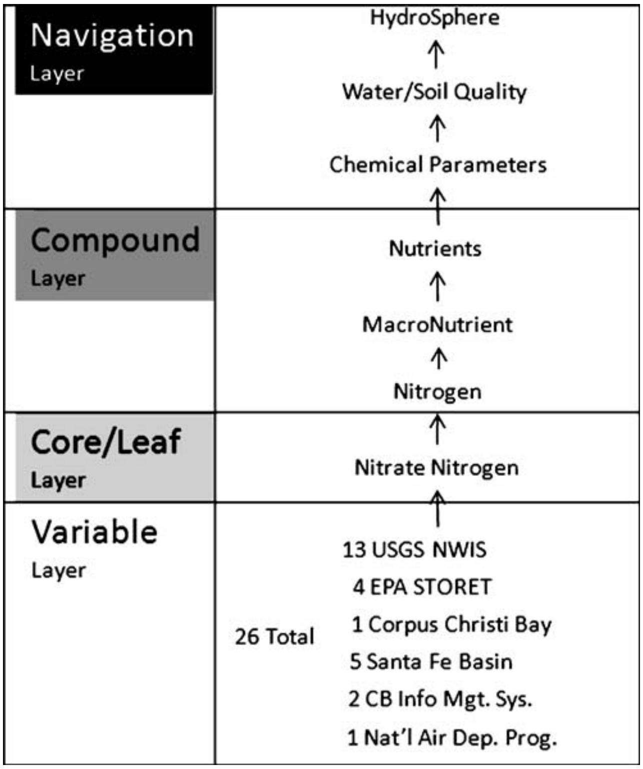


Fig. 3 Example trace for Nitrate Nitrogen from the top concept "HydroSphere" in the Navigation Layer to Leaf Concept Layer. The Leaf concept Nitrate Nitrogen currently has 26 individual variable names (*Variable Layer*) originating from 6 different sources attached to it

(example of "Nitrate Nitrogen") the concept "Nitrogen" is broken down into the commonly used or measured forms of nitrogen, of which "Nitrate Nitrogen" is one. The granularity of the break down at the leaf layer is largely determined by the number of variables (or instances) one can expect to register with each leaf; a rule of thumb suggests more than a dozen but less than 10 dozen, i.e. a number between 12 and 120. Currently, there are 26 nitrate nitrogen variables registered with the leaf concept "nitrateNitrogen".

The implementation of the registration within the HIS central registry happens through the creation and addition of a table that maps the concept identifier (each of the ~400 concepts has its own ID, even though the ones at the top level and compound levels are not used for the registration process), to a unique variable identifier (the central metadata catalogue assigns a unique ID to each newly registered variable) thus providing a unique pair of IDs across a global information system. Through the ID pairings (as shown in Fig. 4 for the example of the nitrate nitrogen variables from EPA STORET and USGS NWIS), the central registry can also identify all ancillary information associated with each registered variable ID as provided for by the ODM design; for example variable name, medium it was collected in, units, regular sampling intervals (or not), and so on (for details of the ODM design see Tarboton et al. 2007). This information is of importance when carrying out the tagging (or registering) step so each variable can be properly tagged a specific concept.

Aspects of tagging

The main attraction of using ontology for discovering data is the fact that an ontology is a graph, as opposed to a purely hierarchical tree, with the ability to define multiple parent concepts. In other words, a leaf concept can exist at multiple places in the ontology having different parent concepts. For example, the concept "Carbon as Nutrient" appears as a child of "MarcoNutrients" but is also a concept group that collects all forms of carbon. Another example is "Arsenic" which is a "Heavy Metal" but also belongs to the concept "Priority Pollutants" (a concept incorporated from EPA's STORET classification system). This flexibility in fact allows many subjective "views" to exist in parallel (different researcher like to see their variable tagged at different locations). It also introduces a high degree of flexibility when further expanding the ontology because leaf concepts can be repeated many times over always taking their "tagged" variables with them. As a consequence, it is fairly easy to expand and add; so long as the "old" concepts are not taken out backward compatibility is always ensured and new tagging is kept to a minimum. Currently the ontology has about 1,100 variables tagged to 128 leafs, i.e. an average of about 8-9 variables per leaf concept. However, efforts are underway to expand the list of leafs to encompass those listed by EPA's substance registry system (SRS 2009), as well as additional biological information as listed in the Texas Commission for Environmental Quality (TCEQ). This effort will add some 1,500 substance codes (leafs) to the current ontology for which some 8,500 USGS and 2,700 EPA STORET variables have been registered. Once completed this will be the most comprehensive collection of environmental variable ⇔ concept pairings in the US.

When defining the rules of how to best tag variables to ontology concepts several questions and issues arose:

- Can a variable also be tagged to a higher level concept, for example nitrate nitrogen variables to "MacroNutrient"?
- Can a variable be tagged to multiple leaf concepts or should a variable be tagged to a single concept and then use the multiple parent concepts to place a variable at different discovery locations?
- Is there is a logical way to resolve the hyponymy problem, i.e. lack of specificity of a variable name and multiple appearances thereof in different contexts?

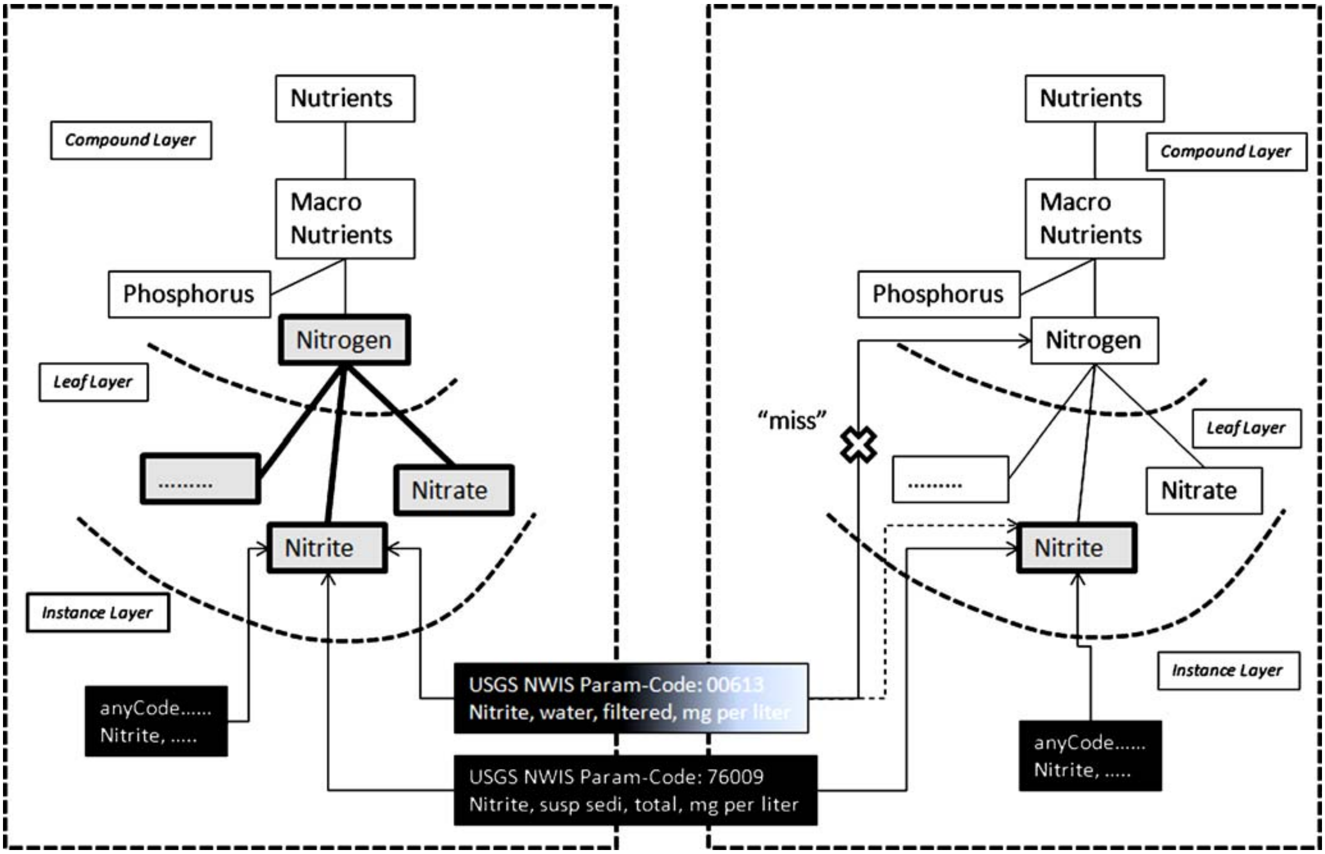


Fig. 4 Comparison between tagging to a Leaf concept only (*Nitrite*) and allowing higher level concept tagging. In the right panel higher level tagging would cause the NWIS parameter (*code:00613*) to be missed when searching for Nitrite

The first question concerns the possibility to create instances of classes that are higher up in the branch system of the ontology as shown in Fig. 4 (example of two USGS nitrite parameters is shown). Theoretically a variable can be tagged to any concept in the ontology. However, it is also clear that there is little to be gained from permitting an unconstraint tagging approach. Because of the fact that the parsing is carried out from highest to lowest concept, i.e. from concepts in the "Compound" layer to the "Leaf" layer, complete coverage (or "pick up") of all subsequent more specific concepts is ensured (depicted on the left side of Fig. 4 with bold framing; search keyword used: *Nitrogen*). On the other side, if the variable is tagged up higher in the branch system it may be missed entirely (as shown in the right side of Fig. 4; search keyword used: *Nitrite*) because the search concept was more specific than then one to which variables are tagged. Secondly, if one would allow the arbitrary tagging at any location along a major branch, a high level concept could end up with hundreds of tagged instances which negates the objective of keeping instance numbers to a manageable level (manageable here means in the double digits, preferably in the lower double digits). Finally, arbitrary tagging at higher branch concepts would prohibit the possibility of returning fully classified (along "Leaf" concept definitions) parsing results; a feature that is explored in the HydroSeek application providing efficient navigation of returned search results. Hence, we impose the constaint that variable can only be tagged at the "leaf" layer.

The second question concerns the need for a variable to appear at different places, i.e. different concept locations in the ontology. More specifically, this means if a variable can be an instance of two different concepts or if the variable should be the instance of just one concept and the concept having multiple parents, as shown in Fig. 5. While it is theoretically possible to carry out any number of tagging actions, i.e. tag variable "X" to concept "A" and then "B" and then "C" and so on (this is equivalent to allowing a variable to be the instance of multiple concepts), from a practical point of view this also poses a considerable extra burden for the individual (for example, a network data manager) carrying out the tagging.

The third question concerns the lack of specificity for a particular variable because it does not describe the context it was measured in. For example, the variable "Temperature" is measured in air, soil, water, and organic

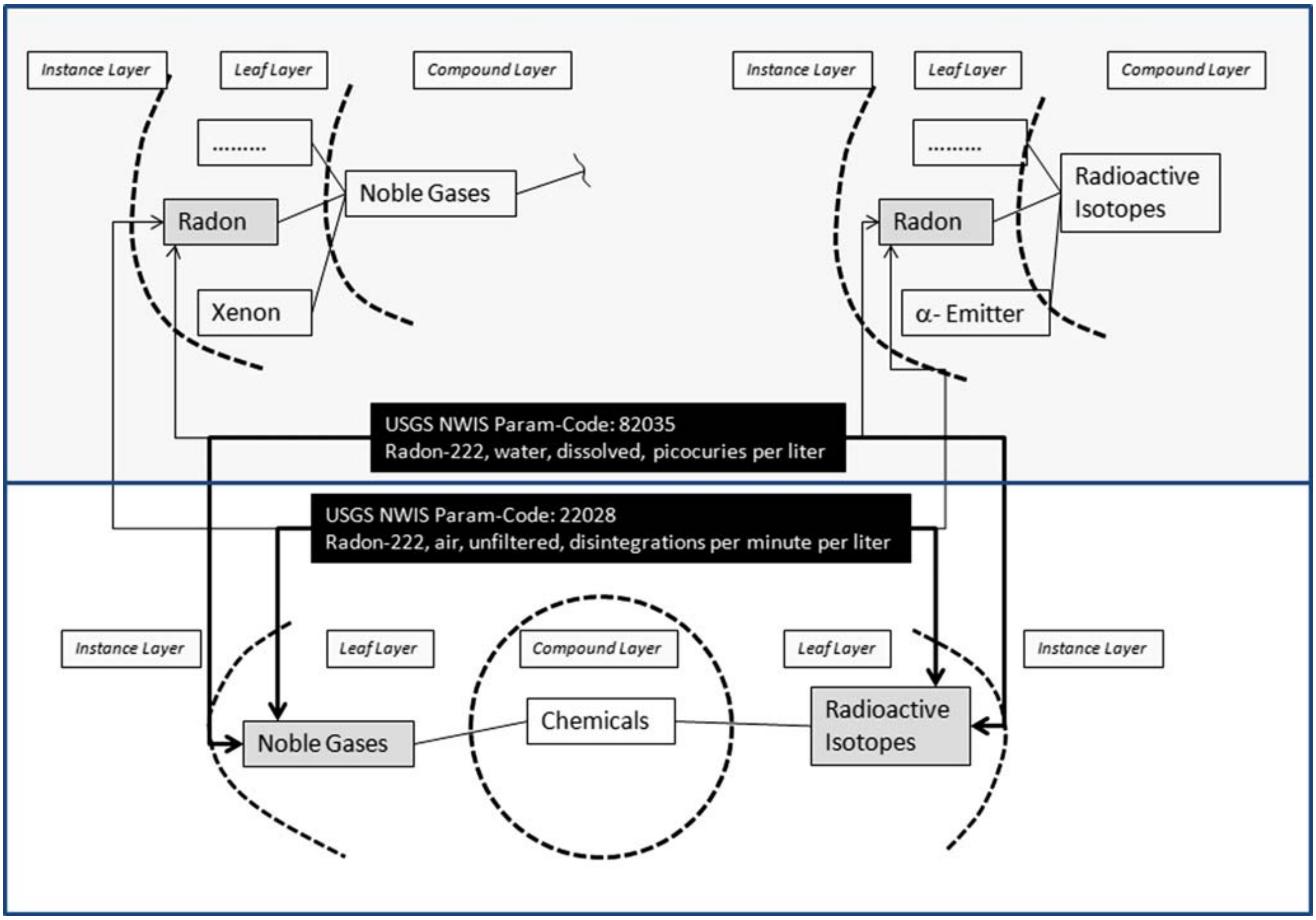


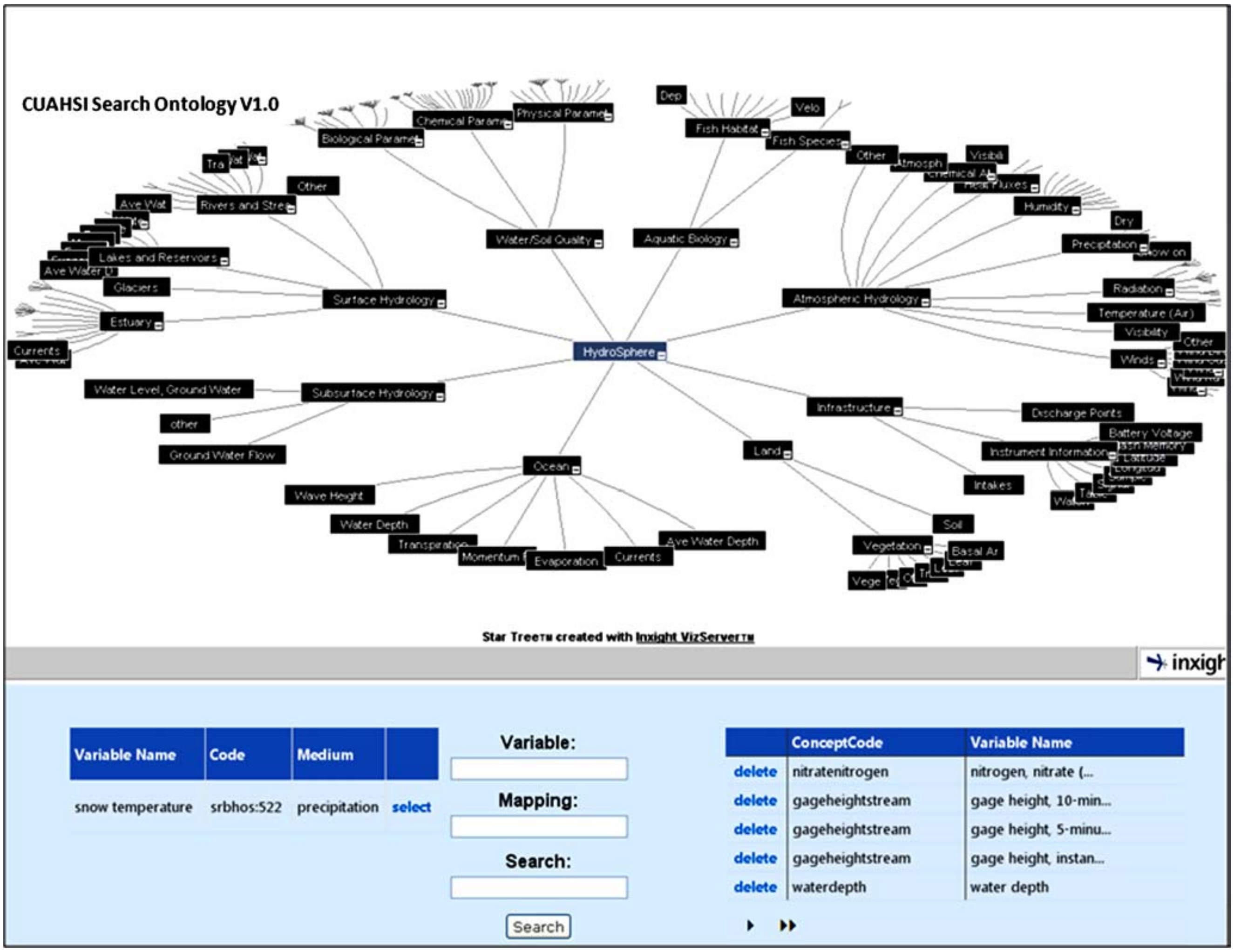
Fig. 5 Tagging to a single concept (*shown is Radon*) which then can have multiple parents (*as shown in the upper panel*) is less work intense than having to execute multiple tagging actions for individual radon variables acquiring multiple parent realizations (*lower panel*)

matter thus needing additional qualifiers (in this case it would require identification of the medium) to uniquely identify a temperature measurement in the appropriate sections in the ontology. The current ontology resolves this dilemma by requiring the definition of concepts such as ‘airTemperature’, ‘waterTemperature’, ‘soilTemperature’, ‘snowTemperature’ and so on thus removing the ambiguity and avoiding the hyponymy problem. However, this convention runs counter to the desire to logically identify only one variable name for each physical quantity: in this sense *temperature* is the physical quantity measured and as such should be strictly described as Temperature only without additional qualifier.

Tagging application

Design and features

The introduction of the hyperbolic StarTree viewer (URL reference see above) for viewing the ontology solved a serious problem in that it permitted efficient visual “access” to the ontology structure without having to use applications requiring long and tedious scrolling actions for traversing the various branches. It was decided that this viewing feature would be a crucial element in designing a visually supported tagging application, called HydroTagger, without having to access OWL editors such as Protégé or SCOOP or having to manipulate other, for example, table based applications requiring equally long scrolling actions. In addition to the viewer (that can also be navigated through a concept search feature) the graphical user interface also needs to accommodate three panels: one for listing all those variables that had been discovered as being “new” during the last updating run of the CUAHSI HIS central metadata catalogue (for more information on the central catalogue visit http://hiscentral.cuahsi.org) and thus needing tagging, one panel that shows the currently registered variable-concept pairs for the specific registered network, and one panel providing some click and execute actions to carry out the actual tagging, as shown in Fig. 6. It should be noted that the tagging is actually carried out between a unique



Variable Name	Code	Medium	
snow temperature	srbhos:522	precipitation	select

Variable:

Mapping:

Search:

	ConceptCode	Variable Name
delete	nitratenitrogen	nitrogen, nitrate (...
delete	gageheightstream	gage height, 10-min...
delete	gageheightstream	gage height, 5-minu...
delete	gageheightstream	gage height, instan...
delete	waterdepth	water depth

Fig. 6 HydroTagger application user interface; the upper panel shows the ontology while the lower panel tracks the outstanding tasks and also shows the already established mappings. In this example “snow temp” in the SRBHOS network has not been tagged and is awaiting the mapping

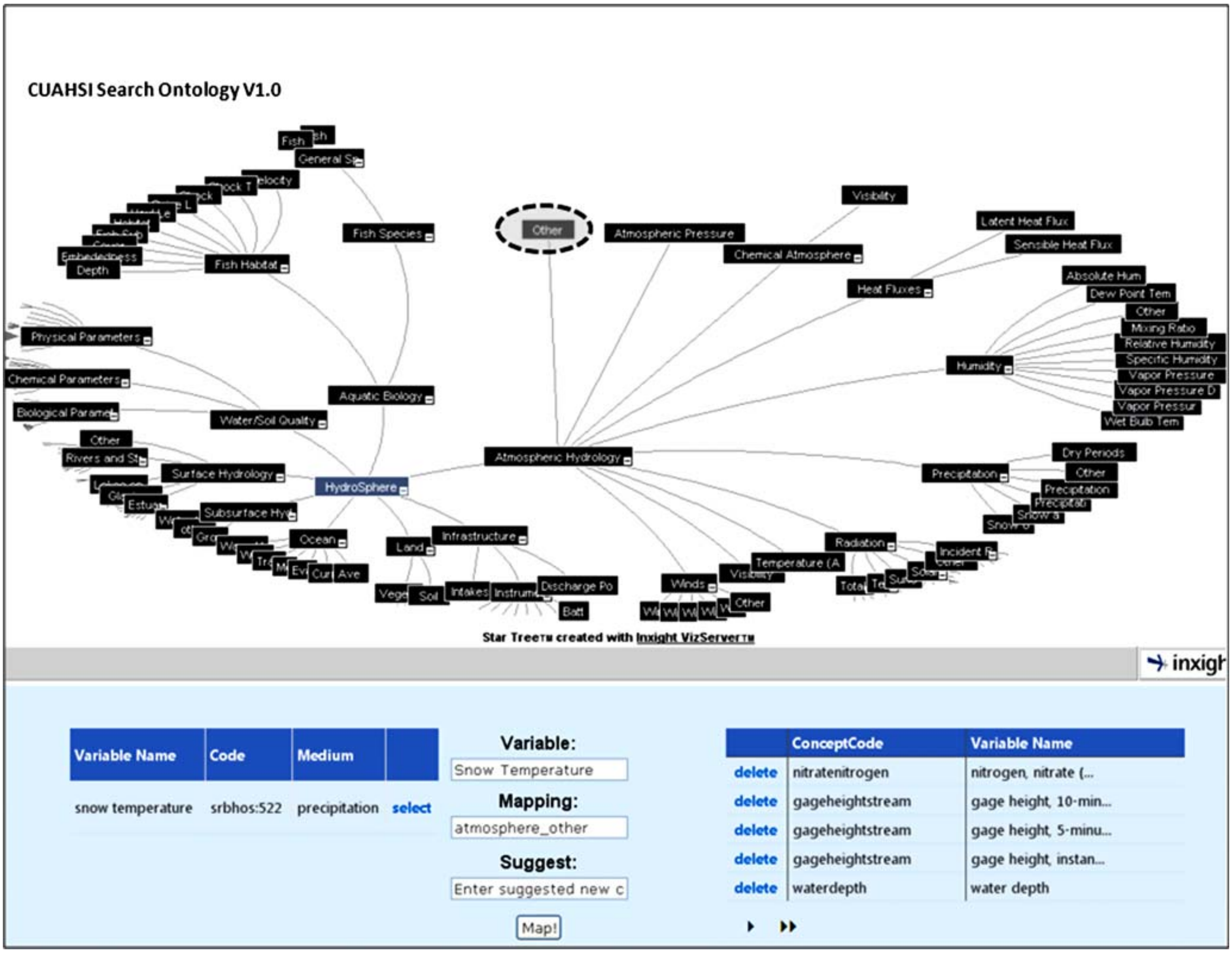
“variableID” and a unique “conceptCode” while only the “variableName” and “conceptLabel” are being displayed guiding the user through the tagging process.

The first version of the tagging application did not show any ancillary information in addition to the variable name which caused problems when trying to uniquely identifying variables. For example, the variable name “Temperature” does not specify where the temperature is being measured thus making it impossible to find the correct ontology concept. The newer version thus added two categories to the left panel: medium and variable code. While this addition helped in many instances it did not resolve issues that arose from lack of information other than “medium” and actually recognizing the unique “variableCode” and what variable it actually stands for. Multiple measurements of a variable at one site (for example turbidity measurements carried out by two different instruments) while uniquely identified through their different codes would still display the same variable name and medium. If one instrument data stream were to be censored (because of an unreliable calibrating procedure), the data manager would need to recall from the “variableCode” (typically an integer number) alone which one of the measurements to tag while leaving the other untagged. The observations data model would support a number of additional qualifiers that in theory would clearly identify each variable and its context, however, the issue is that of space and potentially cluttering the GUI with an excessive amount of information and pull down menus making it more difficult to navigate the tagging process.

In order to facilitate a minimum level of community involvement for expanding the ontology in case concepts are not present, the application features a number of placeholder concepts typically labeled “other”. These “other” concepts are placed in the Core/Leaf Layer because they are by definition tagging concepts and are intended to provide a place to map a variable to if a user cannot find the proper concept anywhere in the “Core/Leaf Layer”. In the example shown in Fig. 7 the variable “Snow Temperature” has no corresponding concept in the ontology, hence it would need to be added. The user identifies “Atmospheric Hydrology” as the main branch to place snow temperature with, and then looks for the “other” placeholder (Notice that there is no Compound Layer entry here, i.e. the Core/Leaf Layer connections directly to a “Navigation Layer” concept. After selecting the variable and clicking the “other” (identified by the dashed line in Fig. 7) concept the center panel switches and now displays a ‘Suggest’ box into which the user needs to type the new “conceptCode” (not the concept label because the tagging is carried on the conceptCode and not the conceptLabel), for example “snowTemperature”. Once the mapping has been carried out the new concept is stored in the HIS Central facility awaiting approval by the curator team. Notice that the chosen level is appropriate because it already hosts the concept “Temperature, Air”.

It should be noted that the ability to add new concepts at the Core/Leaf level is limited in that it does not permit to move concepts to other places (for example across layers), nor does it allow to start entirely new branches, nor does it allow multiple tagging, i.e. the possibility of identifying multiple parents. While this may seem to be somewhat of a limitation, it is also important to recognize that these types of alterations can completely break the variable-concept connections that have already been established. Hence, the only change permitted is a forward-change, i.e.an addition at the Leaf/Core level to ensure backward compatibility. However, these functions (in additions to others) would be essential to fully support a graphical ontology editor without active OWL file manipulations.

The HydroTagger is based on number of web based technologies that have been combined in this application.



Variable Name	Code	Medium	
snow temperature	srbhos:522	precipitation	select

Variable: Snow Temperature

Mapping: atmosphere_other

Suggest: Enter suggested new c

Map!

	ConceptCode	Variable Name
delete	nitratenitrogen	nitrogen, nitrate (...
delete	gageheightstream	gage height, 10-min...
delete	gageheightstream	gage height, 5-minu...
delete	gageheightstream	gage height, instan...
delete	waterdepth	water depth

▸ ▸▸

Fig. 7 HydroTagger user interface demonstrating the feature for concept adding at the Leaf layer. In this case, since no concept “Snow Temp” exists, the user can utilize a placeholder “other” and initiate a new concept by providing a new concept name

The basic application is encoded using JAVAScript which defines the basic layout of the GUI and loads cascading style sheet (CSS) definitions to control the appearance of text. The hyperbolic StarTree viewer is a commercial product (InXight: http://www.inxight.com) that requires the JAVA SDK to be installed on the client's machine. It operates off a server (the single user license is installed at the San Diego Supercomputer Center) and is loaded to the GUI as a JAVA applet. Functionality implemented on the Server such as reading untagged variables and medium information from the central water catalogue, writing back concept/variable pairs, and the ability to delete concept/variable pairs from the catalogue are handled through code written in C# on MicroSoft's .NET platform. Also, the StarTree viewer application uses its own custom file format and cannot directly ingest and display a XML or OWL file. To this end a converter program (written in C#) was developed that translates the concept file (OWL type) into the custom file format.

The role of hydrotagger within CUAHSI HIS

While the HydroTagger is an important component to execute the semantic tagging it is just one piece in a larger set of cyberinfrastructure developed by CUAHS HIS, as schematically shown in Fig. 8.

The HIS Central facility is being hosted at the San Diego Supercomputer Center. This facility encompasses a group of MicroSoft servers that publish the WaterOneFlow (WOF) web services thus exposing the central water metadata catalogue; the facility also exposes applications that permit outside data managers to register the web services of their HIS nodes, it permits the testing of these services, it hosts the central water metadata catalogue in which all metadata for all registered services (or sources) are being compiled into an Observations Data Model DM database, it hosts a data search and discovery tool called HydroSeek, and also provides the semantic tagging application HydroTagger.

The workflow for the semantic mapping is as follows:

1) The harvester application (either scheduled or invoked via request) accesses the central registry and interrogates all registered services for the latest additions, in terms of sites, variables, and values.
2) It compares the discovered timestamps of all variables and their values with those that are registered in the

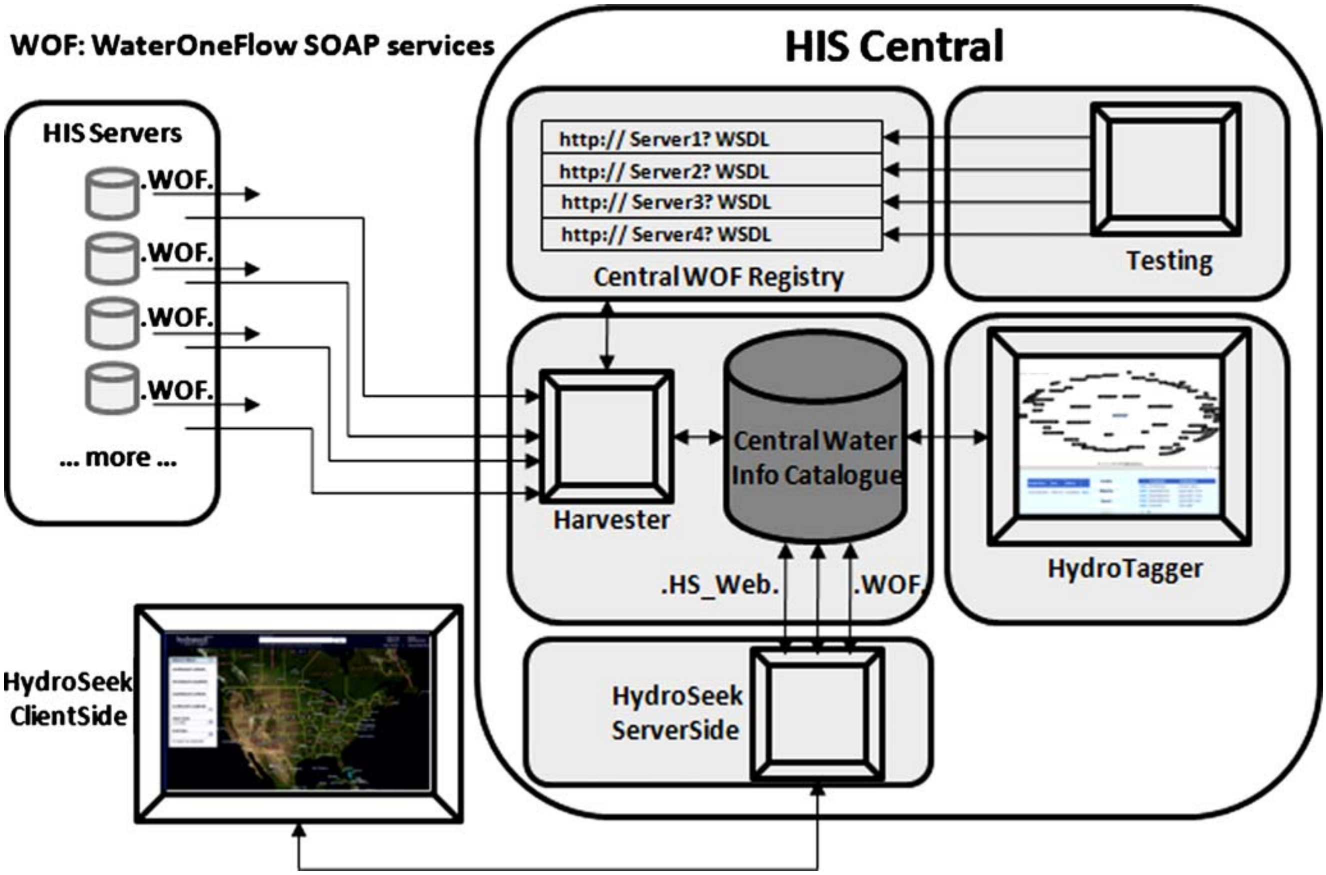


Fig. 8 The role and position of the HydroTagger application within CUAHSI's HIS Central environment, which manages the central water metadata catalogue, Web Service registrations, testing, harvesting, and the HydroSeek application

central catalogue and updates the catalogue with values not previously stored. It also compares the set of variables (and sites) found during the interrogation with those already registered.

3) If the interrogation yields new variables previously unknown to the central catalogue the variables (including their values) are stored in the central catalogue. In addition a flag is set that these new variables need to be mapped to a concept, this is done through an entry into a special table that is part of the central catalogue DB.
4) The next time the data manager logs into HIS Central he is informed that the harvesting has yielded a new variable for the network he is responsible for. He is then asked to visit the HydroTagger application in HIS Central to perform the semantic tagging.
5) Once the tagging has been carried out the variable-concept pair is stored in a general mappings table and the formerly non-tagged variable is removed from the pending table.
6) The variable-concept pairs are now accessible through a suite of specially designed discovery webservices (those are derivatives of the WOF) that are accessed through the search and discovery application HydroSeek.
7) In case new concepts have been added during the mapping process, the central administration needs to review the suggested concept for appropriateness and proper placement in the concept ontology. Once it has been approved, it is added to the variable-concept table and the ontology is updated.

The current version of the HydroTagger application is fully functional. It should be noted that the user group of this tool is fairly limited, i.e. it is restricted to registered data managers (preferably to one individual per registered data source). There are currently 34 registered networks (http://hiscentral.cuahsi.org/pub_services.aspx) from about 20 different data sources (nationwide sources as well as individual PIs). The smaller number of data sources results from some sources having multiple networks (such as USGS NWIS) and some data sources are being handled by HIS Central data managers directly. This means that there are currently about two dozen individuals who have either full (HIS Central data managers) or partial (data source specific) access to the tagging system. Consequently, there is very limited feedback and the HydroTagger has neither been exposed to a large user group evaluation nor are there any metrics this paper could report on. However, during the first couple of years of its operation some feedback suggests the need i) to expand the display of additional variable information during the tagging process (for better identification of the variable) ii) for a better text search capability within the StarTree so concepts can be found easier.

Summary and future outlook

The semantic annotation concept has proven to be extremely successful in aiding the CUAHSI HIS team overcoming the semantic heterogeneity. The idea of developing and identifying concepts that just one level more general then the actual variables collected by mission agencies and individual researchers has worked remarkably well, which has been recognized by many users of the system. Initial shortcomings have been the lack of sufficient branches and Leaf/Core concepts to map against. As a consequence, the ontology has been expanded significantly over the first 2 years eventually reaching a state where all variables collected at the participating sites as well as subsets (those considered most important) from the large mission agencies have been successfully mapped, plus all those variable names that have been registered in the controlled vocabulary of CUAHSI HIS have a matching concept.

We have found that it is more efficient to define Leaf/Concepts that can be moved around and duplicated so as to establish multiple parent relationships rather than to request multiple tagging to different concepts. The clear separation of ontology concepts (handled by HIS central administration) and their instances (the Variable Layer) provides a clean interface in which user involvement is minimized. This is also leaves the expansion of the ontology, definitions of synonyms, and creation of multiple parent relationships in the hands of a central facility that typically has better means for managing and making sure that the provenance is ensured.

The problem of synonymy has been addressed by permitting additional qualifiers to be used in the variable name. While this is a relatively simple approach, it also constitutes a breach with what are commonly accepted best practices, i.e. the desire to avoid concatenating variable names with additional metadata tags thus creating cumbersome and unreasonably long variable names that are, in the extreme, impossible to manage and remember. The CUAHSI HIS group has not yet found a solution that would permit the system to stay with a clean name and providing supporting metadata through other means (for example pull down menus).

The definition of synonyms (those are not visible in the HydroTagger application, i.e. a new variable is always mapped to a "master" concept for which perhaps one or more synonyms exist) has proven to be an adequate approach to reduce the need for an a priori knowledge of permissible keyword. However, the current collection of synonyms needs to be expanded to cover a wider range of nouns commonly used when searching for data. While the current collection works reasonably well for the hydrologic community, it is clear that if other communities want to use the HydroSeek search engine then more synonyms need to be incorporated.

The ontology organization into various layers (4) that have different roles to fulfill while somewhat ad hoc at the beginning proved to be an excellent approach with permissible levels of generality and specificity surprisingly well defined. There has been very little criticism from the user community in terms of lack of generality when picking the search keywords, suggesting that many users have a fairly good understanding of the type of data they are looking for with pool of permissible search keywords (and as such concepts) being just at the right level.

The most challenging task ahead for the CUAHSI HIS team is the continued development of the underlying ontology concepts. This concerns three aspects; firstly the design of the upper ontology levels, i.e. that of the Navigation and parts of the Compound Layers, secondly the branch off into the lower Compound Layer and the definition of Core/leaf layer entries, and thirdly the scope of the ontology in terms of communities covered. It is clear that this is necessarily an iterative approach that will need the active participation of the community. This is not a trivial task and will require a continued commitment of energy, time and resources to engage a sufficiently large number of domain experts to ensure vetting and acceptance of the conceptual organization of domain keywords. While this manuscript is exclusively referring to the CUAHSI discovery ontology V1.0, efforts are under way to move this initial ontology to version 2.0 and then subsequent versions through the establishment of an ontology workgroup that is part of a number of advisory groups guiding the HIS group on its developments.

Acknowledgements The authors would like to thank David Valentine, Tom Whitenack and Ilya Zaslavsky from the San Diego Supercomputer Center at the University of California, San Diego for their feedback on the development of the application in general and the installation and integration into the HIS Central environment in particular. The migration from a prototype into a production type installation would not have been possible without their dedication and hard work. We would also like to extend our gratitude to the CUAHSI HIS team as well as the testbed managers for their continued beta testing of the application and the suggestions for improvement they have provided. Lastly, we would like to acknowledge the National Science Foundation for their generous support of this work; the development was supported under grant number EAR 0622374.

References

Beran B, Piasecki M (2008) “Engineering New Paths to Hydrologic Data”, submitted to *Computers and GeoSciences*, Elsevier, Accepted for publication December 2007.

EPA (2009) “STORET, Store and Retrieval, a Computerized Environmmntal Data Storage System”, US Environmental Protection Agency. http://www.epa.gov/storet/

Fox P, McGuinness D, Cinquini L, West P, Garcia J, and Benedict J (2008) "Ontology-supported Scientific Data Frameworks: The Virtual Solar-Terrestrial Observatory Experience", Computers and Geosciences, doi: 10.1016/j.cageo.2007.12.019.2009

US Geological Survey (2009) “National Water Information System: Web Interface”, USGS Water Data for the Nation, http://waterdata.usgs.gov/nwis

GEON (2009) The GEON Grid Portal, http://www.geongrid.org/index.html. accessed June 2009

Horsburg J, Berger J (2008a) “ODM Data Loader, Version 1.1”, An application for loading data into the CUASHI HIS ODM, http://his.cuahsi.org/documents/ODMDL_1_1_Software_Manual.pdf. June 2008

Horsburg J, Berger J (2008b) “ODM Streaming Data Loader, Version 1.1”, An application for loading streaming sensor data into the CUASHI HIS ODM, http://his.cuahsi.org/documents/ODMDL_1_1_Software_Manual.pdf. June 2008

ISTC (2009) Information Technology and Systems Center, University of Alabama at Huntsville, http://www.itsc.uah.edu/index.html. accessed June 2009

LEAD (2008) Linked Environments for Atmospheric Discovery, https://portal.leadproject.org/gridsphere/gridsphere?cid=portal-home. accessed June 2009

MMI (2009) Marine Metadata Initiative, Monterrey Bay Aquarium Research Institute (MBARI). http://marinemetadata.org. accessed June 2009

NASA (2009) “Global Change Master Directory, GCMD”, discover Earth Science data and services, http://gcmd.nasa.gov/, implemented as ontology at the MMI at http://marinemetadata.org/gcmd

Raskin R (2009) “SWEET, Semantic Web for Earth and Environmental Terminology”, Jet Propulsion Laboratory. http://sweet.jpl.nasa.gov. accessed November 2008

SESDI (2009) Semantically Enabled Science Data Integration, The High Altitude Observatory at the National Center for Atmospheric Research. http://sesdi.hao.ucar.edu/intro.php. accessed June 2009

SRS (2009) Substance Registry System, Environmental Protection Agency, The Environmental Exchange Network. http://www.exchangenetwork.net/exchanges/cross/srs.htm. accessed June 2009

Tarboton D, Horsburgh J, Maidment D (2007) “CUAHSI Community Observations Data Model (ODM), Version 1.0, Design Specifications”. http://www.cuahsi.org/his/docs/ODM1.pdf. May 2007

To E, Whitaker T (2008) “HydroGET, A web service Client for ArcGIS”, http://his.cuahsi.org/hydroget.html

Whitaker T (2008) “HydroEXCEl Version 1.1 Software Manual”, http://his.cuahsi.org/documents/HydroExcel_Software_Manual.pdf

Whiteaker T, Tarboton D, Goodall J, Valentine D, To E, Beran B, Min T (2007) “HIS Document 5: CUAHSI WaterOneFlow Workbook”, Version 1.0, http://www.cuahsi.org/his/manuals/HISDoc5_UseWebServices.pdf

World Wide Web Consortium (2006) “XML Schema Definitions”, version 1.1, http://www.w3.org/XML/Schema

World Wide Web Consortium (2007) “Web Ontology Language, OWL”, http://www.w3.org/2004/OWL/

Zaslavsky I, Valentine D, Whiteaker T (2007), “CUAHSI WaterML”, submitted to Open Geospatial Consortium as discussion document. http://www.cuahsi.org/his/docs/WaterML-030-forOGC.pdf. May 2007