Geophysical Research Abstracts
Vol. 12, EGU2010-3689, 2010
EGU General Assembly 2010

Keyword Ontology Development for Discovering Hydrologic Data

Michael Piasecki (1), Rick Hooper (2), and Yoori Choi (2)
(1) Drexel University, Civil & Env Engrg, Philadelphia, United States (mp29@drexel.edu), (2) CUAHSI, Medford, MA, USA

Supply of adequate keywords in data searches is a key element to building search clients. This is particularly true for science or engineering areas that straddle many subjects such as the hydrology field. The main difficulties that arise when trying to search across many disparate data holdings is that there are no norms that describe data sets uniformly (this is a lack of common metadata profiles) in particular when it comes to identifying them with recognizable labels. This leads to problems associated with hyponymy (a word or phrase whose semantic range is included within that of another) and synonymy (several terms are used to describe the same thing or parameter). Particularly the latter is endemic in the hydrologic data world and poses a substantial obstacle when trying to build information systems that can search for data across multiple data sources.

This paper discusses the effort that has been undertaken within the Consortium of Universities for the Advancement of Hydrologic Sciences Inc. (CUAHSI: http://www.cuahsi.org) Hydrologic Information Systems (HIS) development group to overcome these semantic heterogeneities by developing a keyword ontology that can be navigated to identify search keywords of ascending or descending generality to identify parameter sets or fairly specific parameters that the search engine should be searching for. The general is to tag or connect any variable name to a presented leaf concept in the ontology that best describes what a specific data set represents. While the search environment is not part of this paper, we will describe the underlying ontology, its extent, the way it is organized and why, and what sources and considerations were taken into account in developing the current version.

An initial ontology of 4033 leaf concepts describing physical, chemical and biological properties has been developed. These leaf concepts cover the vast majority of the records contained in major data sources such as the US Geological Service (USGS) National Water Information System (NWIS) and the Environmental Protection Agency's STORET data system . In order to avoid overwhelming returns when searching for more general concepts, the ontology's upper layers (called navigation layers) cannot be used to search for data, which in turn prompts the need to identify general groupings of data such as Biological, or Chemical, or Physical data groups, which then must be further subdivided in a cascading fashion all the way to the leaf levels. This classification is not straightforward however and poses much potential for discussion. Finally, it is important to identify on the dimensionality of the ontology, i.e. does the keyword contain only the property measured (e.g., "temperature") or the medium and the property ("air temperature").