Controlled Vocabularies Boost International Participation and Normalization of Searches

Lola M. Olsen
NASA's Goddard Space Flight Center

For presentation at the:
7th International Conference on Hydrosciences and Engineering (ICHE)
Metadata in Hydroscience Mini Symposium

Bossone Research Center
Drexel University College of Engineering
Philadelphia, PA
September 10-13

More than a decade ago, the Global Change Master Directory's (GCMD) science staff set out to document Earth science data and provide a mechanism for its discovery in fulfillment of a commitment to NASA's Earth Science program and to the Committee on Earth Observation Satellites' (CEOS) International Directory Network (IDN). At the time, the dilemma of whether to offer a controlled vocabulary search or a free-text search was resolved with a decision to support both. The ease of developing an effective free-text search initially appeared to be the more attractive option. However, the feedback from international groups indicated that being asked to independently determine the appropriate "English" words through a free-text search would be very difficult. They preferred to be "prompted" for relevant keywords through the use of a hierarchy of well-designed science keywords. Because the controlled keywords are required for all data set descriptions, they serve to "normalize" the search through knowledgeable input by metadata providers.

During the last decade, Earth science keyword taxonomies were developed to assist in the search for related data and services. To reduce maintenance and streamline the process of developing and maintaining the hierarchies, rules for additions, deletions, and modifications were created. These rules have served the directory well in the ongoing maintenance of the keywords. In addition, secondary sets of controlled vocabularies for related descriptors such as projects, data centers, instruments, platforms, related data set link types, and locations, along with free-text searches assist users in further refining their search results. Through this robust "search and refine" capability in the GCMD, users are commonly directed to the data and services they seek.

The GCMD's keyword taxonomies are widely used by organizations within: (1) universities, such as the University of California, Drexel University, and George Mason University; (2) other federal agencies, including the National Oceanic and Atmospheric Administration (NOAA), the Federal Geospatial Clearinghouse (FGDC), and the Department of Energy (DOE); and (3) international organizations and projects, such as Geoconnections (Canada), the Australian Antarctic Data Center (AADC), the British Antarctic Survey (BAS), and the Intergovernmental Oceanographic Commission/International Oceanographic Data and Information Exchange (IOC/IODE). In addition, the GCMD's keyword taxonomies have been translated into French, Japanese, Chinese, and Spanish. As such, the vocabularies stand as a valuable and readily available contribution for search within other Earth science data systems around the world.

The next step in guiding users more directly to the resources they desire is to build a "reasoning" capability for search through the use of ontologies. Controlled keyword taxonomies, such as the GCMD's, are useful in the development of such ontologies. The GCMD has played an important role by sharing its taxonomies, which have been used as the source for the development of preliminary ontologies by organizations, such as: (1) the Marine Metadata Initiative: MMI; (2) the Semantic Web for Earth and Environmental Terminology: SWEET; (3) the GEOsciences Network: GEON; (4) AnnoTerra: Prototype semantic web application for Earth Observation System (EOS); (5) Upper Hydrologic Vocabulary (Drexel); (6) Enabling Parameter Discovery: EnParDis from the British Oceanographic Data Center (BODC); and (7) the Digital Library for Earth System Education: DLESE.

Incorporating twelve sets of Earth science keyword taxonomies has boosted the GCMD's ability to help users define and more directly retrieve data of choice. The public benefits through the higher precision of the search made possible by the controlled input of appropriate keywords by metadata contributors. By extending the knowledge for the relationships created among the variables for data sets and services within the directory, the knowledge of inherent relationships may help to provide baseline measures of the effectiveness of the semantic web. We are currently exploring and defining plans for the creative development of ontologies to assist with the discovery of Earth science data and services – using these valuable resources created in-house. In so doing, the GCMD is expected to play an important role in demonstrating the value and boosting the profile of the developing semantic web. Examples using the hydrosciences will be demonstrated.