10.4 ONTOLOGY-BASED SEMANTIC SEARCH TOOL FOR ATMOSPHERIC SCIENCE

Rahul Ramachandran*, Sunil Movva, Sara Graves and Steve Tanner
University of Alabama in Huntsville

1. INTRODUCTION

The web is an enormous source of information containing resources such as web pages, data files, metadata catalogs, publications etc. Search tools typically used by a scientist can be broadly categorized into two categories based on the methodology used to collect the metadata. Search tools such as Google, Alta Vista and Lycos use special software robots called spiders. These spiders crawl the information space in the World Wide Web to extract metadata from web pages and rank these web pages based on the extracted metadata. These spiders typically rank the page based on the words within the page and where the words were found. Words occurring in the title, subtitle and meta tags in the HTML document are given higher weight. This approach results in completeness where every web page that mentions the search term is retrieved. This approach works well for documents but is typically not suitable for scientific data.

The second approach typically used by different scientific data catalogs involves creating a formal metadata. A formal metadata is metadata that follows some standard specification that provides a common set of terminology, definitions and information about the values to be provided. Examples of such specifications include Dublin Core Initiative, ISO standards and Federal Geographic Data Committee (FGDC). Government agencies such as U.S. Geological Survey (USGS) and National Aeronautics and Space Administration (NASA) archiving geospatial data are mandated to follow the FGDC specification. In addition to the specification that addresses what needs to be stored in the catalog, a set of keywords or a control vocabulary such as Climate and Forecast (CF) Metadata Convention or the Global Change Master Directory (GCMD) is also required to populate these catalogs. Queries on such catalogs normally produce very accurate search results.

* Rahul Ramachandran, Information Technology and Systems Center, University of Alabama in Huntsville, AL 35899

email: rramachandran @itsc.uah.edu

The metrics used to measure the quality of an ideal search result are both accuracy and completeness. Thus, an ideal search should find every thing and only the things a user really wants. Both the approaches mentioned here fail in this regard. Web search engines are simple to use but end up retrieving too much information. Searches on catalogs at data archives are more accurate but also more complicated and typically not complete. The user has to know the control vocabulary to effectively search the catalog. Non experts such as students have a hard time framing the right query to yield the correct result.

In this paper we present Noesis, an ontology-based semantic search tool that addresses these problems. Noesis is much more than just a simple semantic search engine, but is also a resource aggregator collating relevant information from distributed resouces. The tool's architecture and functionality are described after a brief overview on ontologies.

2. ONTOLOGY AND THE SEMANTIC WEB

An ontology was described by Aristotle as the science of being. From Machine Learning/Artificial Intelligence/Intelligent Systems perspective "an ontology is a formal, explicit specification of a shared conceptualization" (*Gruber, 1993*). Therefore, an ontology should contain concepts and constraints of use that are explicitly defined. It should be formal to be machine understandable and must be shared such that it captures consensual knowledge. Ontologies have two distinct components. They contain the names for important concepts for a specified domain. For example, an ontology for biology can contain "Elephant" as a concept whose members are a kind of "Animal and a "Herbivore" as a concept whose members are exactly those animals who eat only plants or parts of plants. Similarly, an "Adult Elephant" can be a concept whose members are exactly those elephants whose age is greater than 20 years. In addition to definition of concepts, an ontology also specifies the background knowledge and the constraints of the domain. Thus, the biology ontology could contain constraints and relations such as an "Adult Elephant" weighs at least 2,000 Kg, All "Elephants" can either be "African Elephant" or "Indian Elephant", no individual can be both "Herbivore" and a "Carnivore", etc. Thus, an ontology defines concepts in a domain and relationships between these concepts.

Ontologies will play an important role in achieving the vision of the Semantic Web (*Berners-Lee et al., 2001*). In this vision, machines will not just present data but also understand and use the data. Web pages with XML tags around ontology terms will enable machines to ascertain their meanings by examining the ontology content referenced by the namespace in the tag. Such a scenario will allow machines to be able to perform better searches and allow the machines to use the information without human intervention. Thus, searches in such a scenario will significantly reduce the number of false hits and increase the number of successful hits.

3. NOESIS TOOL COMPONENTS

The system architecture for the Noesis tool is presented in Figure 1. There are three main components in this tool and these components are described next.

3.1 LEAD Ontology

The Noesis tool uses the ontology being developed as a part of the Linked Environment for Atmospheric Discovery (LEAD) Project (*Droegemeier et al., 2004; Droegemeier et al., 2005a; 2005b*). The LEAD ontology is being built using the Semantic Web for Earth and Environmental Terminology (SWEET) ontology (*Raskin and Pan, 2005*). The SWEET ontology was developed to allow discovery and use of Earth science data, through software understanding of the semantics of web resources. SWEET contains a collection of ontologies in the Ontology Web Language (OWL) (*Bechhofer et al., 2004*) that include both orthogonal concepts (space, time, Earth realms, physical quantities, etc.) and integrative science knowledge concepts (phenomena, events, etc.). SWEET is based on the NASA Global Change Master Directory (GCMD) which includes approximately 1000 controlled Earth science keywords, represented in a taxonomy. SWEET has been designed as a high level ontology allowing domains within Earth Science to create specialized ontologies leveraging the SWEET concepts. The LEAD ontology will focus on concepts relevant to Atmospheric Science. The American Meteorological Society glossary is being mapped into the SWEET ontology by using the concepts listed in the glossary and defining relationships between them. Therefore, the LEAD ontology will be a specialized ontology for Atmospheric Science and will extend the concepts defined in the SWEET ontology.

There are two main reasons for building the LEAD ontology. First, it will serve as a Knowledge Resource for the educational and research community in Atmospheric Science?. The LEAD ontology will be more than just a static glossary, it will contain definitions and relationships between atmospheric phenomena, parameters, data, services and others high level concepts. The eventual goal for the LEAD ontology will be to create a topology linking these high level concepts. Thus with the help of this ontology, an investigator searching for a term 'Mesocyclone' will be able to discover that this phenomena is defined by a physical parameter 'Vorticity'; that the NEXRAD radar datasets contain the physical parameter 'Vorticity' and a Data Mining service can be applied on this data field to extract the 'Mesocyclones'.

The second reason to build the LEAD ontology is to support semantic searches. The use of an ontology will allow tools such as Noesis to extend the search capability on the metadata catalog and other web resources beyond just using keywords.

3.2 Ontology Inference Service

The Ontology Inference Service (OIS) is a SOAP-based web service interface to an inference engine. It is built on Apache Axis SOAP engine. The inference engine used at the backend is Pellet (*Grau et al., 2004*). Pellet is an OWL DL reasoner based on the tableaux algorithms. The reasoner is pre-loaded with LEAD ontology and provides T-Box and A-Box querying capabilities on the ontology. T-Box queries cover specializations, generalizations and equivalence of a concept. A-Box queries search for all satisfying instances of a concept and querying for property fillers for an instance. Every search request to the OIS is translated to one or more queries for the reasoner. The OIS interacts with the reasoner through the description logic reasoner interface (DIG). The DIG interface is a standard for providing access to description-logic reasoning through an HTTP-based interface. The query results are returned back to the OIS through this interface. OIS has been designed to allow loosely couple integration using standard web services protocol with other systems such as the Query Service in the LEAD Data Subsystem. In Noesis, OIS communicates with the Smart Search Broker which is described next.

3.3 Smart Search Broker

The Smart Search Broker is responsible for managing and coordinating the requests between the user via the client, the OIS and the other distributed resources. Once a search term is entered by the user, the broker passes these terms to the OIS. The results from the OIS are presented back to the user as options via the broker. Once the user has selected the terms to perform the search, the broker uses this list of terms to query the different distributed

resources. These resources can be a search engine such as Google, metadata data catalogs such as the LEAD resource catalog and SURA Coastal Ocean Observing Program (SCOOP) catalogs, educational resources such as the Digital Library for Earth System Education (DLESE) catalog and others.

4. NOESIS USE CASE SCENARIOS

Two examples demonstrating the use of Noesis tool by end users are presented next.

4.1 Specialization Example

The Noesis tool can be used to browse the concept hierarchy in the ontology. While browsing, the users will be able to navigate and traverse the ontology from different points of access. In this scenario, a user may not know the exact name of the physical parameter but can provide the higher level concept. Thus, a user will be able to start at a general topic and be able to navigate to specific topics by selecting the concepts of interest. For example, a user can type a search term "Pressure", the Noesis tool uses the ontology to find specializations such as "Hydrostatic Pressure", "Total Pressure" etc and these results presented back to the user as options. The tool then utilizes the list of selected terms to search different distributed resources such as the Google search engine and the DLESE catalog of educational materials. The aggregated results from this search are then returned to the user. This example use of the Noesis tool is presented in Figures 2a-e.

4.2 Synonym Example

The Noesis tool can also resolve the problem of synonyms for the search term. For example, a user searching the metadata catalog for "Precipitation" will never find the data sets containing data fields with "Rainfall". The Noesis tool uses an ontology to solve this problem. Different synonyms are returned to the user in addition to specializations of the search term as options. Once the user has selected the different options, this selected list of terms is then used to search the different resources and the aggregated results are returned to the user. Thus, the Noesis tool returns both complete and accurate search results.

5. SUMMARY AND FUTURE WORK

The Noesis tool presented here represents the next generation of specialized search tool and resource aggregators that use domain ontologies. The domain ontologies will help guide the user and the search mechanism to make the search results both accurate and complete. The Noesis tool can play a useful role in geoscience research and education. It not only uses a domain ontology to provide the user with context to refine their search term, but also searches different resources that might be of interest to them. These resources will eventually include web pages, related educational material, data sets and relevant publications. This initial version of Noesis is using a limited LEAD ontology. As the LEAD ontology evolves, the newer versions of the ontology will be incorporated into the Noesis tool. The current version of the Noesis tool only searches the web via the Google search engine and the educational material stored in the DLESE catalog. Work is also underway to incorporate the LEAD and the resource catalog to allow users to find relevant datasets.

Even though this version of the Noesis tool is focused on Atmospheric Science, the tool itself can be configured to different domains. The reconfiguration would only require a different domain ontology and access to a different set of distributed resources.

6. ACKNOWLEDGEMENTS

The LEAD project is funded by the National Science Foundation under the following Cooperative Agreements: ATM-0331594, ATM-0331591, ATM-0331574, ATM-0331480, ATM-0331579, ATM03-31586, ATM-0331587, and ATM-0331578.

7. REFERENCES

Bechhofer, S., F. v. Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein, 2004: OWL Web Ontology Language Reference.

Berners-Lee, T., J. Hendler, and O. Lassila, 2001: The Semantic Web. *Scientific American*, **284,** 34-43.

Droegemeier, K., V. Chandrasekar, R. Clark, D. Gannon, S. Graves, E. Joseph, M. Ramamurthy, R. Wilhelmson, K. Brewster, B. Domenico, T. Leyton, V. Morris, D. Murray, B. Plale, R. Ramachandran, D. Reed, J. Rushing, D. Weber, A. Wilson, M. Xue, and S. Yalda, 2004: Linked Environment for Atmospheric Discovery (LEAD): A Cyberinfrastructure for Mesocyclone Meteorology Research and Education. *Interactive Information and Processing Systems (IIPS)*, Seattle, WA, American Meteorological Society.

Droegemeier, K., V. Chandrasekar, R. D. Clark, D. Gannon, S. Graves, E. Joseph, M. K. Ramamurthy, B. Wilhelmson, K. Brewster, B. Domenico, T. Leyton, D. V. R. Morris, D. R. Murray, B. Plale, R. Ramachandran, D. Reed, J. Rushing, D. Weber, A.

Wilson, M. Xue, and S. Yalda, 2005a: Linked Environments for Atmospheric Discovery (LEAD): Architecture, Technology Road Map and Deployment Strategy. *Joint Session on Cyberinfrastructure to support atmospheric and Oceanic Education: Examples and strategies, AMS Annual Meeting*, San Diego CA.

Droegemeier, K. K., D. Gannon, D. Reed, B. Plale, J. Alameda, T. Baltzer, K. Brewster, R. Clark, B. Domenico, S. Graves, E. Joseph, V. Morris, D. Murray, R. Ramachandran, M. Ramamurthy, L. Ramakrishnan, J. Rushing, D. Weber, R. Wilhelmson, A. Wilson, M. Xue, and S. Yalda, 2005b: Service-Oriented Environments in Research and Education for Dynamically Interacting with Mesoscale Weather. *IEEE Computing in Science & Engineering*, **7,** 24-32.

Grau, B. C., B. Parsia, and E. Sirin, 2004: Tableau Algorithms for E-Connections of Description Logics.

Gruber, T. R., 1993: A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition,*, **5,** 199-220.

Raskin, R. G. and M. J. Pan, 2005: Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). *Computers & Geosciences*, **31,** 1119-1125.

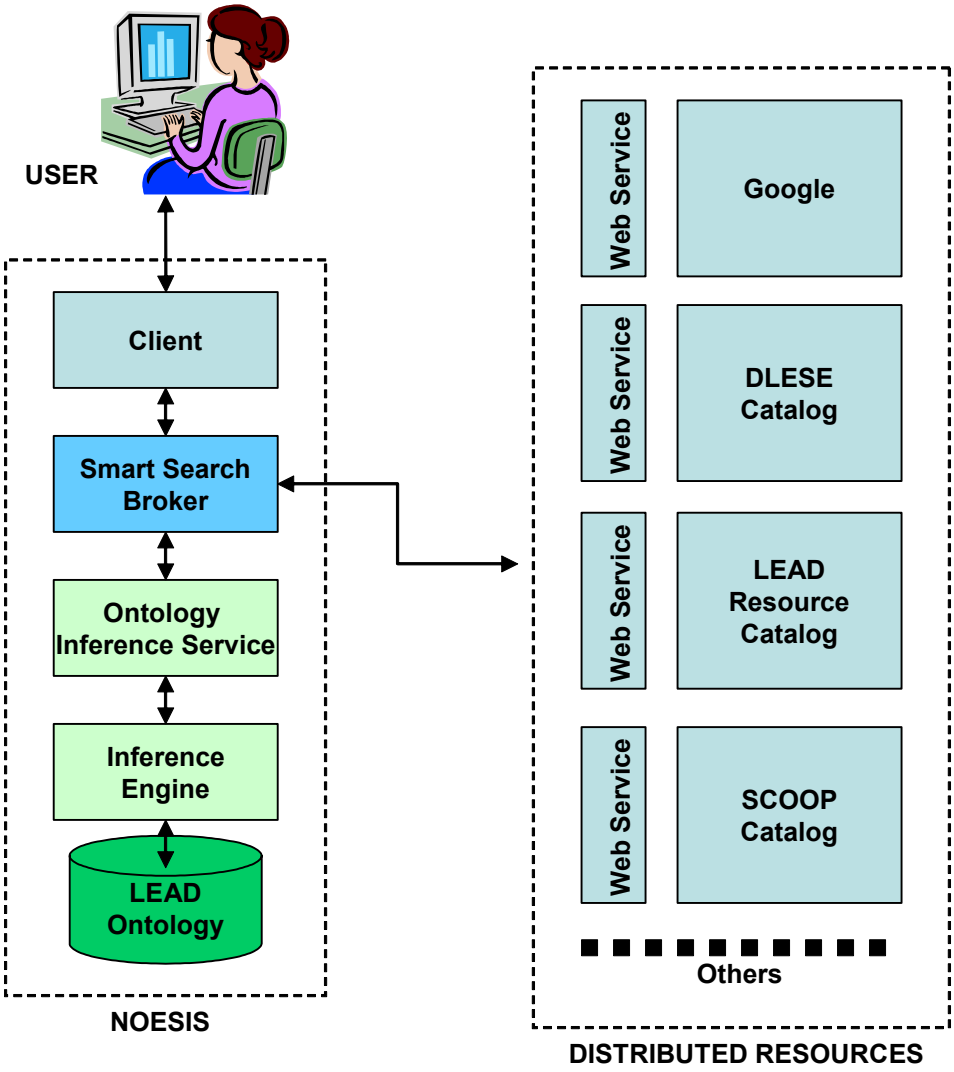

Figure 1: Noesis system architecture

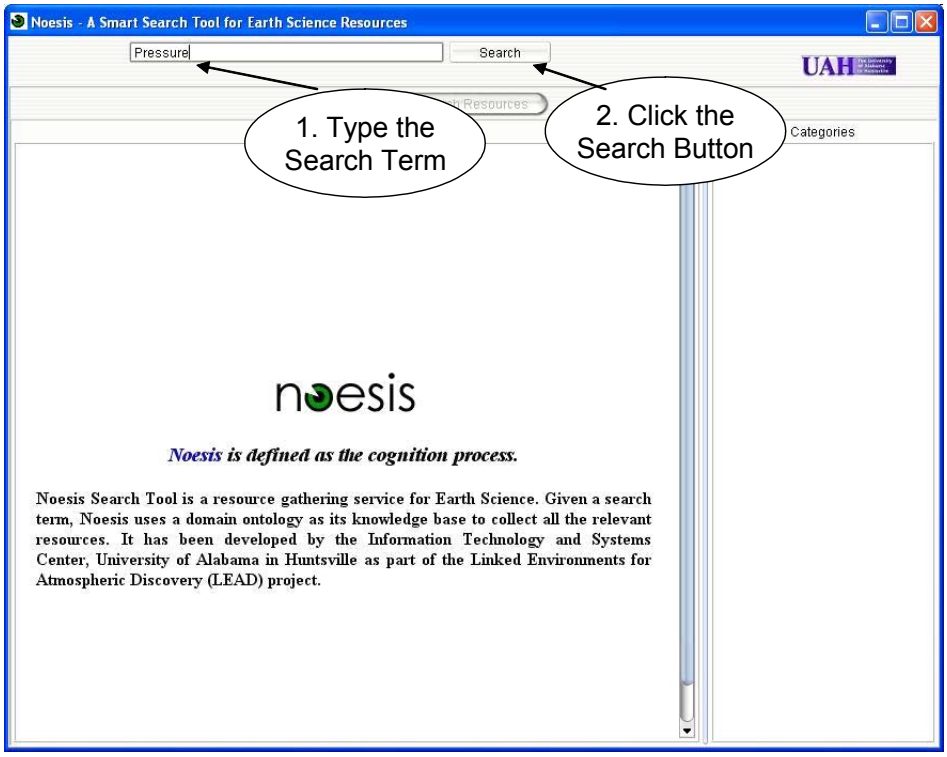

Figure 2a: Example use of the Noesis Tool

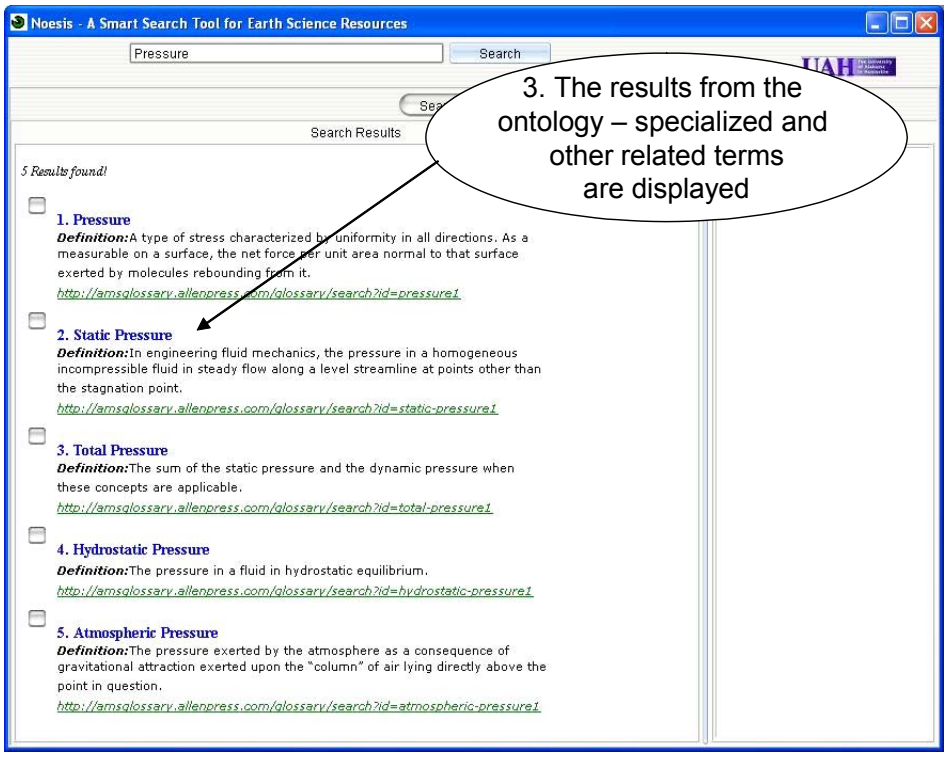

Figure 2b: Example use of the Noesis Tool

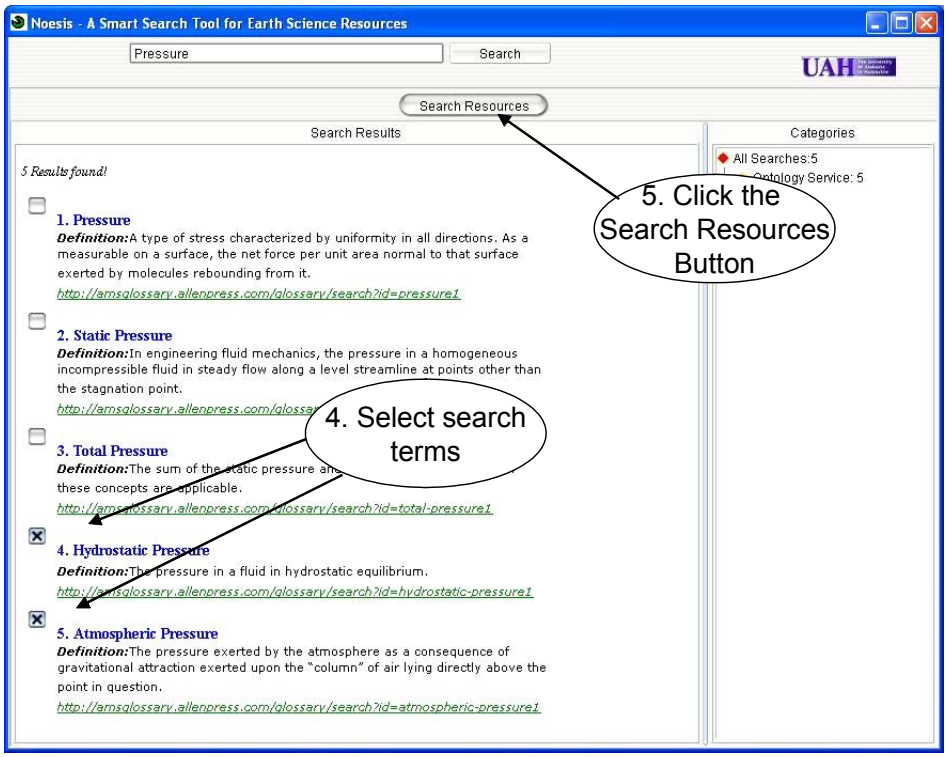

Figure 2c: Example use of the Noesis Tool

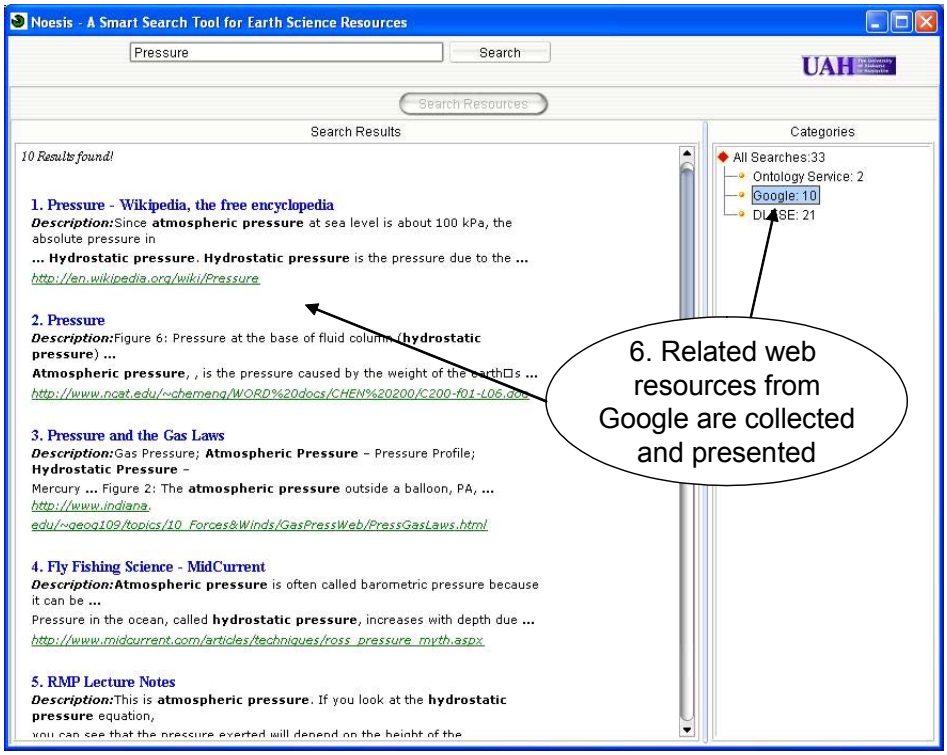

Figure 2d: Example use of the Noesis Tool

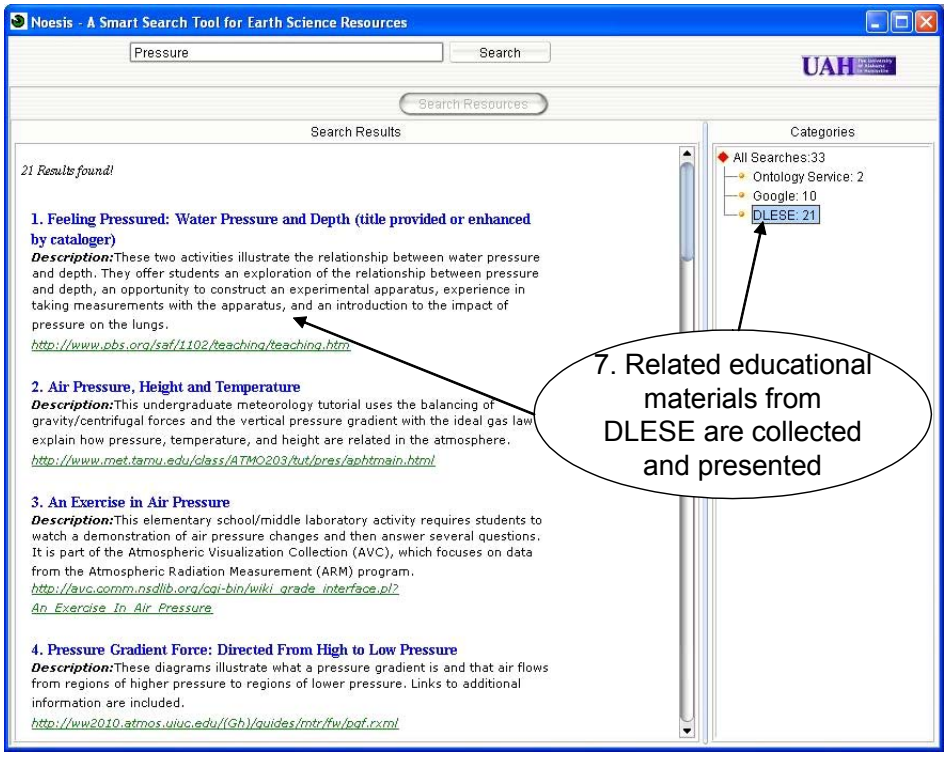

Figure 2e: Example use of the Noesis Tool