Noesis: A Semantic Search Engine and Resource Aggregator for Atmospheric Science

Sunil Movva, Rahul Ramachandran, Xiang Li, Phani Cherukuri, Sara Graves
Information Technology and Systems Center
University of Alabama in Huntsville
Huntsville, AL 35899

Abstract - **The goal for search engines is to return results that are both accurate and complete. The search engines should find only what you really want and find everything you really want. General search engines (even meta-search engines) lack semantics. In this paper Noesis, which is a meta-search engine and a resource aggregator that uses domain ontologies to provide scoped search capabilities will be described. Noesis uses these ontologies to help the user scope the search query to ensure that the search results are both accurate and complete. The users can refine their search query using these domain ontologies and thereby achieve better precision in their results.**

I. INTRODUCTION

There are two types of search engines based on the characteristic of resource: Open Web and Hidden Web search engines. Traditional open web search engines provide syntax based text string search. Open web consists of static web-pages that are hosted on different servers around the world. Keywords from the content of these web-pages are indexed. The basis for search is term matching between the user's query and these indexes. Semantics associated with the search string are not captured. Thus, a search query is typically broad and often requires the user to evaluate the results and iterate using different modifications to the search query to find the appropriate resource (*Fig. 1*).

The hidden web consists of content such as science data catalogues. These catalogues are typically built using a standard vocabulary. Efficient searches on these catalogs are only possible through using the appropriate terms from the controlled vocabulary. Meta-search engines, which are used to provide larger search coverage, cannot handle searches on multiple resources if these catalogs use different vocabularies. In this paper, we introduce Noesis - a search tool for Atmospheric Science designed to address these issues.

II. SEMANTICS AND ANNOTATION

The fundamental reason for the above mentioned issues is the lack of semantic understanding of the resources by the search engines. This is in general a problem tackled by the semantic web [5]. The vision of the semantic web is to enable machines to not just present web resources but also be able to understand the semantic meaning of these resources. Such web architecture is possible only through annotating all the resources with semantic information. One of the major hurdles of the Semantic Web is harvesting this semantic information (Semantic Annotation) for existing content on the web. Ontologies are defined as "explicit formal specifications of the terms in the domain and relations among them" [6] and have been successfully adopted to provide semantic information in similar situations. It has been purported that the Semantic web can only become a reality with contributions from smaller communities [3]. Ideally, each of these communities should develop ontologies for a relatively small domain. Eventually, most of the resources available on the web can then be annotated using the collection of these small ontologies. But, annotating every web resource manually is not possible. Considerable research efforts in the semantic web community are being investigated to provide a semi-automatic / automatic solution to annotate these resources [2]. Yet, there is no elegant solution for this problem. Fortunately, for hidden web resources, it is not a hurdle. Since the hidden web consists of limited science data catalogues which hold metadata about the data stored in the archives and each of these archives uses a specific controlled vocabulary; the process of annotating these data archives is reduced to annotating the vocabulary with ontologies.

III. NOESIS ONTOLOGIES

As mentioned in the previous section, smart search capabilities require semantic description of concepts of the specific domain. Ontologies in general are very appropriate to fulfill this requirement. An ontology captures and encodes knowledge of concepts, constraints and the relationships among them, for use in a machine-readable fashion. Noesis uses two classes of Ontologies namely, Domain Ontologies and Application Ontologies implemented in Ontology Web Language (OWL). Domain Ontologies are used to describe concepts in a domain and their relationships. Noesis uses a set of core domain ontologies for describing concepts in Atmospheric Science.

Better use of the data catalogs by the search engines can be achieved by annotating the metadata vocabulary used in the catalog. Thus, in our approach, we provide application ontologies for the different control vocabularies used by different resources. These application ontologies enable flexible querying by bridging the gap between semantic concepts and the application specific vocabulary. These ontologies annotate the terms used in the catalogs with their conceptual meanings. The concepts in these ontologies are linked with core domain ontologies through the 'owl:equivalentClass.'

IV. Ontology inference service

The power of Ontologies comes from their machine understandability. In order to use the semantic information from the Ontologies, they should be coupled with an Inference Engine. Noesis uses an Ontology Inference Service for this purpose. 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 [4]. Pellet is an OWL DL reasoner based on the tableaux algorithms. The reasoner is pre-loaded with the Noesis Ontologies (Core and Application Ontologies) 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 coupled integration using standard web services protocol.

V. Introducing semantics into the search process

The Noesis ontologies provide semantic descriptions of concepts in the domain. The general search process can be improved by leveraging this semantic information. Instead of a user trying different search queries to get to the desired results, Noesis provides the user with three sets of additional terms that can be used to append or rephrase the search query (Fig. 2, 3). This process of adding more terms to the search query is called *Query Expansion*.

A. Specializations/Generalizations (SP/GN)

Ontologies are organized in tree-like taxonomies, where the child nodes represent the *Specializations* and the parent nodes represent the *Generalizations* of a node (concept). Specializations can be used to provide more detailed search, while generalizations are used to make the search broader (relaxed). For example, as seen in Fig. 2, a search for "Cyclone" shows specializations, "Hurricane" and "Typhoon". Thus, using "Hurricane" as a search query will narrow down the results. Similarly, "Atmospheric Circulation" can be used for generalizations. This process of traversing the concept hierarchy to refine the search query is called Search Scoping.

B. Synonyms (SN)

Synonyms are different terms that have the same meaning. In ontological terms these are the equivalent concepts. 'owl:equivalentClass' allows linking two syntactically different terms to one semantic concept (synonyms). For e.g, as seen in Fig.3, a search for "Reflectance" shows synonym, "Albedo". Appending this term to the query expands the search, thus providing better search coverage.

C. Related Terms (RT)

Every concept has a set of Property concepts that are neither in the same inheritance hierarchy (SP/GN) nor equivalent (SN). These are called the *Related Concepts* and they are captured in the ontology through the property relationships. If the user intends to search for resources on a concept with respect to a particular property, these terms can be appended. For example, as seen in Fig.2, a search for "Cyclone" shows "Rain" as a Related Concept. Appending this term to the search narrows the search to resources that contain information about "Cyclone" within the context of "Rain."

VI. The search algorithm

Noesis uses a three step algorithm to search resources. The three steps are Query Analysis, Semantics Presentation and Resource Search. The algorithm architecture is depicted in Fig.4.

A. Query Analysis

In this step, the user-provided search query is broken down to identify concepts that are defined in the domain ontology. Once they are identified, they are annotated with the associated concepts from the ontology.

B. Semantics Presentation

The annotated concepts from the query string are used to search the Ontology Inference Service (OIS) for associated concepts (Specializations, Generalizations, Synonyms and Related Terms). The Specializations and Generalizations are shown in a tree structure to allow user to navigate through the hierarchy. Synonyms and Related terms are shown in separate categories and a check box is provided to let user select the term to append to the search (Fig.2, 3). The user uses these terms to refine the search query.

C. Resource Search

The selected terms are then used for searching the resources. For open web resource searches, the refined query is directly used to provide results since no semantic information is encoded (annotated) in these resources. For hidden web resources like data archives, an Application Ontology is added for every new vocabulary used. The concepts in the refined query are used to search the Ontology Inference Service to obtain equivalent terms in the associated Application Ontology. The obtained terms are then used to search for resource in that particular catalog.

The obtained results from searching different resources are presented to the user along with the semantic information from the second step (Fig.6). The user can modify the query string by adding and removing the associated terms to see how it alters the search results.

VII. RESOURCE AGGREGATION

Noesis is a meta-search engine. Meta Search Engines simultaneously search multiple Open Web and Hidden Web resources to provide increased search coverage. Noesis searches for web-pages, data, education material and publications related to Atmospheric Science. Noesis uses the refined search string to fetch resources, through search web services provided by third parties like www.yahoo.com and www.google.com (Fig. 6, 7, 8). The resources found by search are categorized based on their sources. These categories are provided to the user (Fig. 5) to enable or disable searching a particular source for searches.

VIII. CONCLUSION

The Noesis tool presented here uses ontologies to associate semantic information with the search process. It provides a *guided refinement of search query* producing successful searches and reducing the user's burden to experiment with different search strings. Obtaining semantically accurate results for open web searches is not yet possible due to the lack of semantic annotations for the open web resources. But, Noesis provides a better solution than traditional search engines by appending semantic information to the search query. Hidden web resources can be annotated with the semantic information and Noesis leverages such a method of semantic annotation in an efficient way for providing meta-search engine capabilities.

IX. REFERENCES

[1] G. Jian, Z. Xiang, D. Jianming, and Q. Yuzhong, "An Ontology-Driven Information Retrieval Mechanism for Semantic Information Portals," *IEEE Proceedings of the First International Conference on Semantics, Knowledge, and Grid (SKG'05)*, pp.63, 2005.

[2] L. Reeve, and H. Han, "Survey of Semantic Annotation Platforms," *The 20th Annual ACM Symposium on Applied Computing (SAC'05)*, Santa Fe, New Mexico, March 13 -17, 2005.

[3] N. R. Shadbolt, W. Hall, and T. Berners-Lee, "The semantic web: Revisited," *IEEE-Intelligent Systems*, vol. 21, issue 3, pp. 96–101, May 2006.

[4] B.C. Grau, B. Parsia, and E. Sirin, "Tableau Algorithms for E-Connections of Description Logics," 2004.

[5] T. Burners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific American*, vol. 284(5), May 2001.

[6] T. R. Gruber, "A Translation Approach to Portable Ontology Specification," *Knowledge Acquisition*, vol. 5, pp. 199-220, 1993.

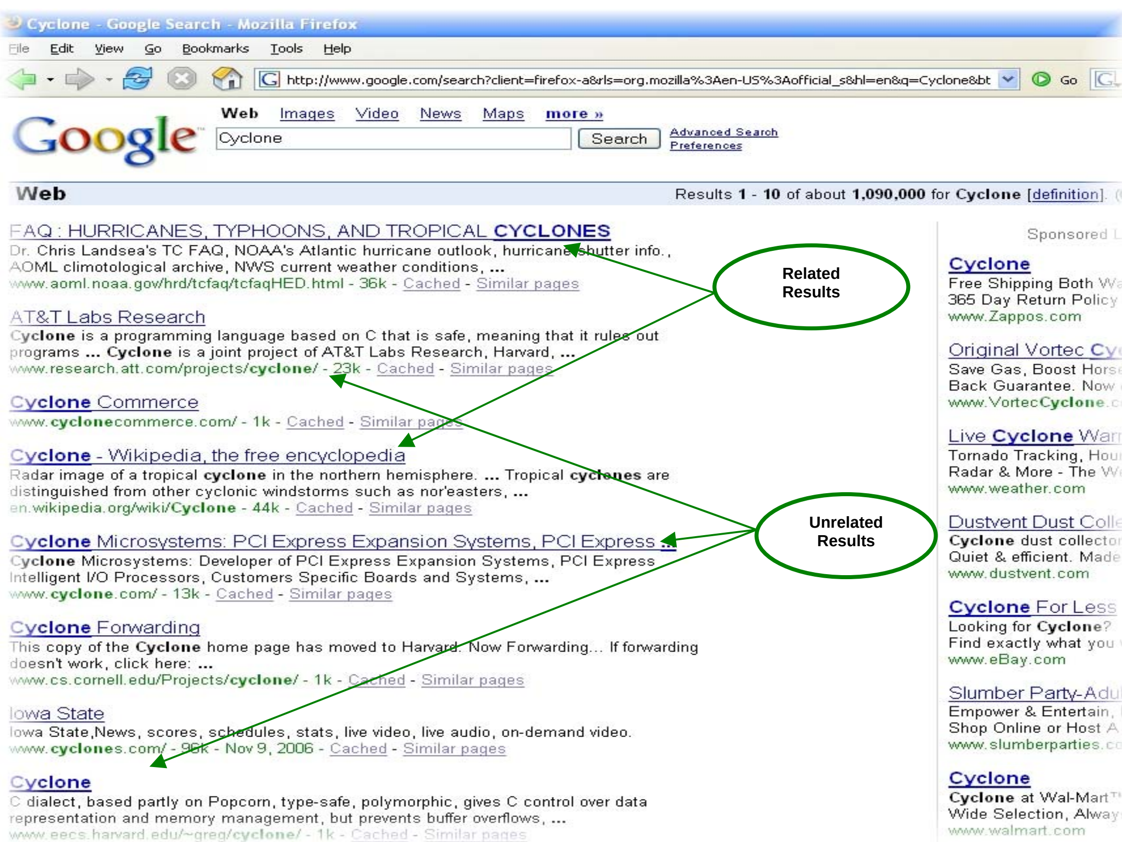

Fig. 1. Traditional Web Search.

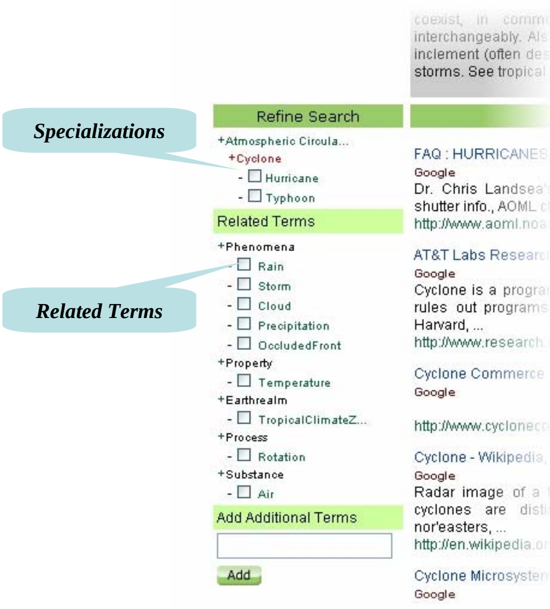

Fig. 2. Specializations and related terms for the search term ("Cyclone") presented to the user.

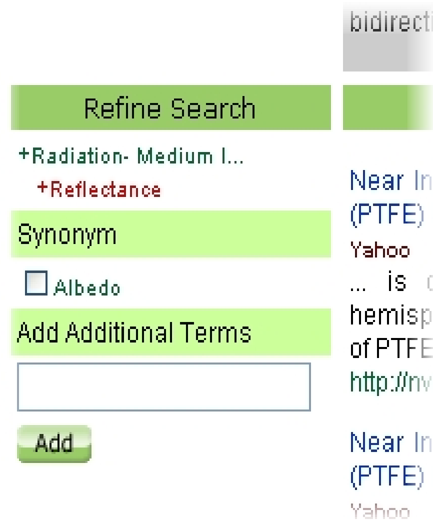

Fig. 3. Synonyms for the search term ("Reflectance") presented to the user.

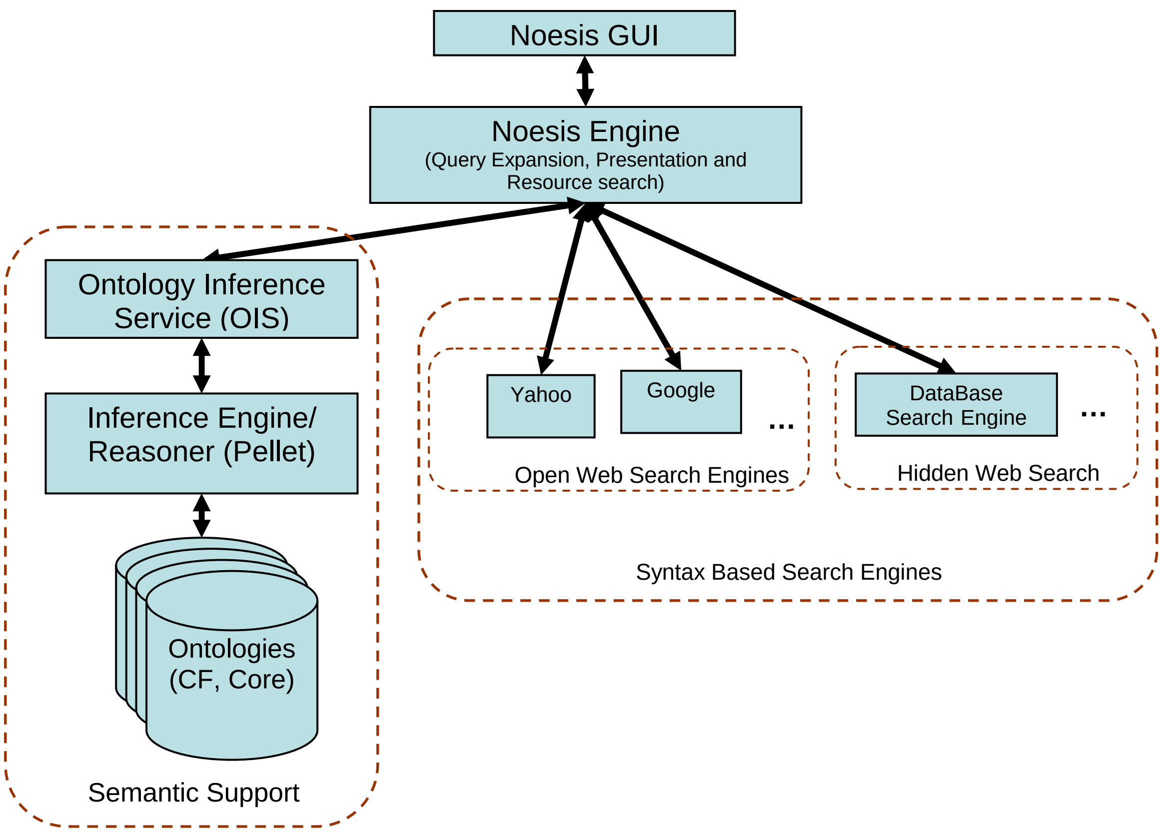

Fig. 4. Noesis Search Architecture.

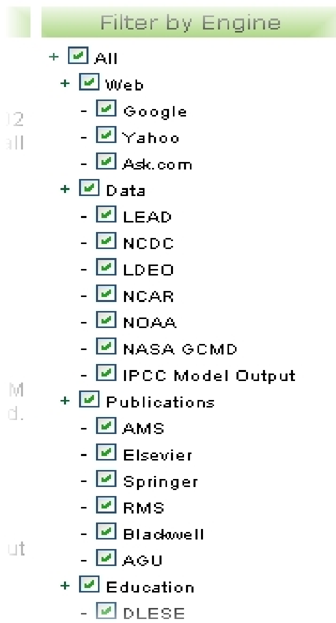

Fig. 5. Resource Aggregation.

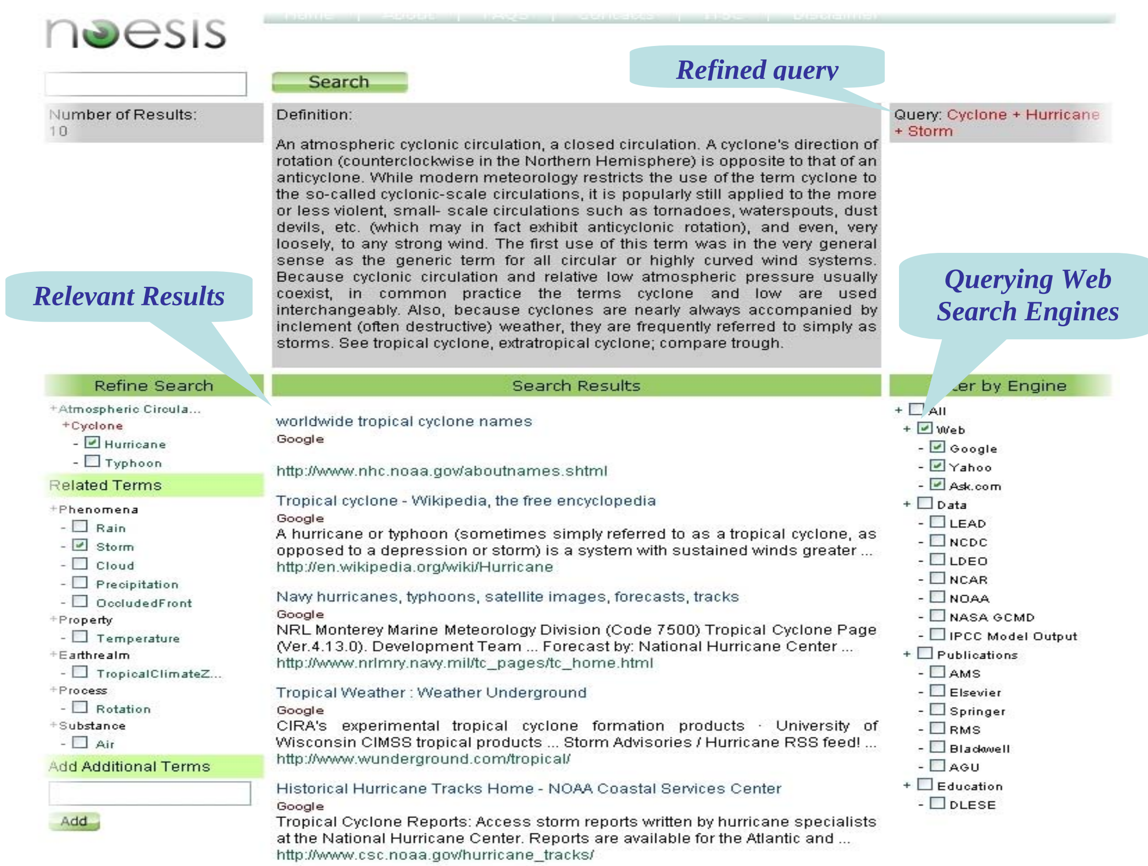

Fig. 6: Web Search Results (From step 3).

Fig. 7. Publication Search Results (From step 3).

Fig. 8. Data Search Results (From step 3).