Thesaurus Support when Searching Earth Science Data

James C. French
Department of Computer Science
University of Virginia
Charlottesville, VA
french@cs.virginia.edu

Lola M. Olsen
Code 902, Bldg.32, S130D
NASA/GSFC
Greenbelt, MD 20771
olsen@gcmd.nasa.gov

Worthy N. Martin
Department of Computer Science
University of Virginia
Charlottesville, VA
wnm@cs.virginia.edu

Abstract— **Keyword hierarchies are known to assist in searches for Earth science data sets. The level of assistance, however, is dependent on the common semantic interpretation by the indexer and the searcher. A strategy to improve search results may lie in the use of the elusive thesaurus. Thesauri have been discussed as the solution to semantic interpretation over the years. However, integrating their use within an interactive search has proven to be more difficult, as the thesauri built to date have most often been stand-alone. To negotiate this hurdle, we describe the integration of a thesaurus built on a well-functioning operational locator of Earth science data.**

I. Introduction

THE EOSDIS is a multidisciplinary data store. This leads to difficulties in cross-disciplinary searching due mainly to differences in terminology. This may manifest itself in several different ways. Users may be faced with unfamiliar jargon when searching in another discipline. They may also use a term that has a different meaning in another discipline. Consider, for example, the term "aerosol." It might refer to gases only or particulate matter or both. There is no way *a priori* to know what a user means by the term, what is included, or what is excluded. The specific semantic difficulty is that the system indexes using one vocabulary and it might be quite different from the vocabulary being employed by any particular searcher.

One attack on this problem is to provide a controlled vocabulary for use in searching. This approach has been used very effectively in some Earth science data systems, for example, the GCMD[1] (Global Change Master Directory). Another approach to mitigate this problem is to provide a thesaurus to help suggest useful search terms to searchers. The work reported here is focused on the latter approach. The full context of our research on these problems is outlined in [2]. In a companion paper [1] we describe our conceptual framework and additional approaches to mitigate these vocabulary problems.

Our current prototype work has developed and demonstrated an integrated thesaurus service for Earth science data systems. Our initial prototype was demonstrated in connection with the GCMD. One of our objectives is to provide an integrated thesaurus server that can be accessed from other NASA ES data systems such as the EOS Data Gateway (EDG). Term suggestion and thesaurus support can be useful in any interface.

II. DLR Thesaurus

The German Remote Sensing Data Center (DFD)[2] of the German Aerospace Center (DLR)[3] developed the original thesaurus that forms the foundation of this work.

The DLR thesaurus uses Oracle 8i on the server-side to manage the thesaurus data structure. A fragment of that data structure is depicted in Figure 1. This data structure captures the important thesaural notions by conceptually linking terms in a graph with appropriate relationships, e.g., synonyms, broader and narrower terms, and related terms. Each node in Figure 1 is shown with a label and the number of synonyms contained in the node. In a sense the node is known by $n + 1$ equivalent labels. Note that the DLR thesaurus contains English and German synonyms but our counts only show the count of English synonyms. This is to give the reader an idea of the richness of the thesaurus in a monolingual mode.

A. Search Assistant

We have written a new client-side Java applet to access the thesaurus. A screen shot of our interface is

This work supported in part by NASA Grants NAG5-8585 and NAG5-9747 and NASA GSRP NGT5-50062.

[1] `http://gcmd.nasa.gov/`

[2] `http://www.dfd.dlr.de`

[3] `http://www.dlr.de`

shown in Figure 2. The screen shown in the figure is in response to a user query for the string "atmospheric pollution." Note that the entry is labeled "air-pollution." Any of the synonyms listed will retrieve this node; the string "air-pollution" has simply been designated as the node's canonical name. The display of Figure 2 is "located at" the `air pollution` node in Figure 1. The bold lines shown in the data structure (Figure 1) correspond to the terms enumerated in the display (Figure 2).

The thesaurus data structure provides for the following relationships. They are described with respect to the current concept. For examples of each refer to Figure 2.

Synonyms: An equivalence class of strings denoting this concept. One of the strings is used as a label for the class.

Top terms: The terms at the top of the hierarchy. These are the broadest terms containing this concept.

Broader: Immediate predecessor terms in the hierarchy.
Narrower: Immediate successor terms in the hierarchy.
Related: Arbitrary terms in the thesaurus structure. Provide an alternative method for navigating the structure.

Each of these categories is represented in the display for a concept, for example see Figure 2.

Our interface supports two main activities:
1. Adding terms to the current query.
Any term in any category can be added to the current query. Simply mouse over the term and right-click the mouse. The selected term is added to the query.
2. Navigating the thesaurus structure.
The display can be refocused to any of the latter four categories from above (top terms, broader, narrower, or related). Simply mouse over the term and left-click the mouse.
When the `Finish` button is clicked, the Search Assistant returns to the form from which it was invoked with the modified search string substituted for the initial search string.

Figure 3 shows how simply the thesaurus Search Assistant can be implemented into the GCMD interface.

B. Update Assistant

We have also prototyped an Update Assistant to facilitate maintenance of the thesaurus data structure. The interface, not shown here, is very similar to the Search Assistant and provides familiar add, change, delete functionality for introducing new thesaurus terms or updating existing terms. The Update Assistant is intended for restricted access by personnel responsible for the thesaurus maintenance.

III. Deployment Strategy

In our initial prototype we provide for direct access to the thesaurus service as shown in Figure 4. The thesaurus service is stateless with respect to the invoking interface. The client-side Search Assistant is responsible for managing the modified query string. The specific ES data system is unaware of the existence of the thesaurus server.

We are currently reworking the thesaurus interface so that it can be packaged as a web service and exported to ES data system applications via ECHO (EOSDIS ClearingHOuse)[3].

References

[1] J. C. French, A. C. Chapin, and W. N. Martin. Using Multiple Viewpoints to Improve Access to Earth Science Data. In *Proc. Earth Science Technology Conference*, 2002.

[2] J. C. French, W. N. Martin, and L. M. Olsen. Extending the Vocabulary Available for Cross-Disciplinary Searching of Earth Science Data. Technical Report CS-2002-04, Department of Computer Science, University of Virginia, 2002.

[3] R. Pfister, R. Ullman, and K. Wichmann. ECHO Responds to NASA's Earth Science User Community. In *HCI International*, 2001.

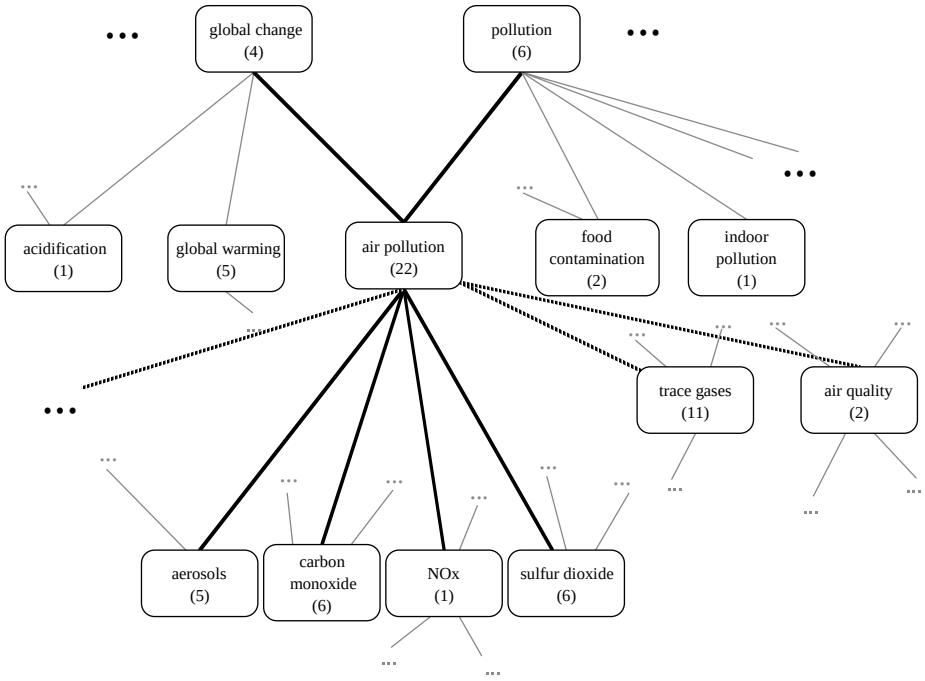

Fig. 1. DLR thesaurus data structure.

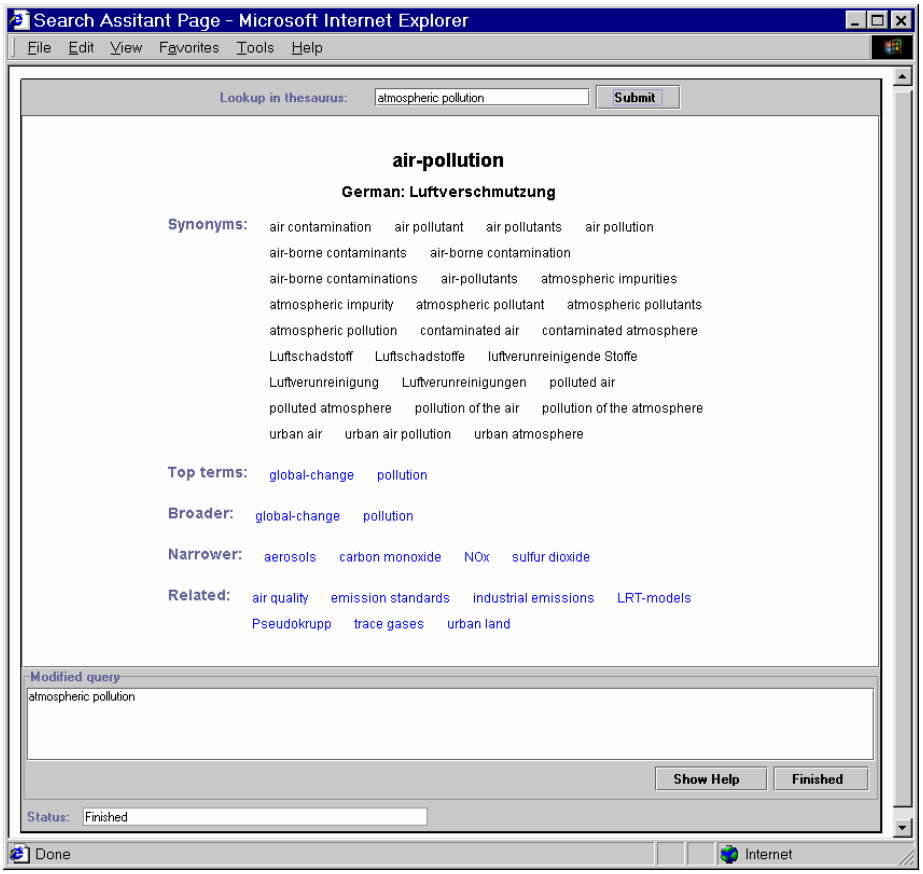

Fig. 2. Interface to Search Assistant.

Fig. 3. GCMD interface with thesaurus Search Assistant added.

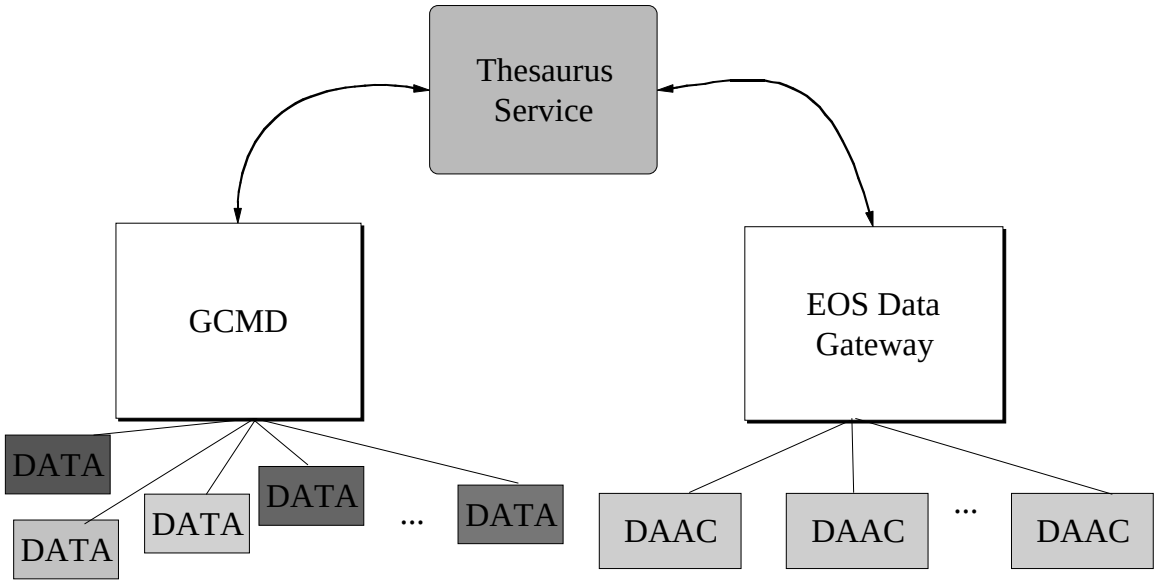

Fig. 4. Stand-alone server accessed directly from applications.

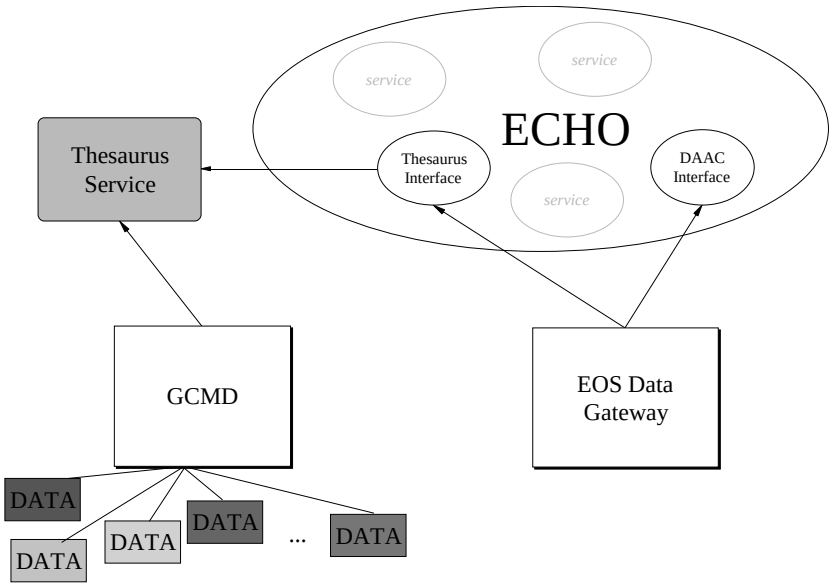

Fig. 5. Stand-alone server with direct access from applications and also packaged as a web service via ECHO.