Extending the Vocabulary Available for Cross-Disciplinary Searching of Earth Science Data

James C. French Worthy N. Martin *
Department of Computer Science
University of Virginia
Charlottesville, VA
{french|wnm}@cs.virginia.edu

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

ABSTRACT

This paper discusses the development of a prototype search assistant designed to aid cross-disciplinary searching of Earth science data. The goal of the project is to provide aids to help searchers overcome the vocabulary problem: the vocabulary used by the searchers is not the same as that used by the indexers. The work was motivated by vocabulary issues existing in the EOS Data Gateway (EDG)[1] at the time this work began. We describe the status of the project and give examples of the techniques that we are using. We also discuss the way in which we are implementing a new search paradigm, multiple viewpoints, within our prototype.

1. INTRODUCTION

Linking disparate information resources into cohesive federated information systems involves solving a number of vexing interoperability issues. Even if we adopt unambiguous standards for required metadata and agree completely on the semantics, we still have a serious problem in resolving individual field values. The quality of the information is the issue.

When investigators try to search the scientific literature in domains other than their own, they are often faced with an additional problem: the vocabulary used in the new domain is unfamiliar to them and they are unable to compose effective queries. This problem is not confined to the scientific literature; rather it occurs whenever heterogeneous collections of information are aggregated into a common information resource. We know what we want, but we are often unable to cast our information need into the proper jargon for effective search. We need help in the form of a translation service that maps our queries into forms more usable in these other environments. This problem is sometimes called "vocabulary switching."

The EOSDIS (Earth Observing System Data and Information System) 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. It might also be the case that the lay term "dust" is used instead and what is needed is some mapping to the notion "aerosol particulate matter." Another example is the term "precipitation." To some this is simply a synonym for "rain," but to others snow, sleet, *et cetera* should be included. The specific semantic difficulty is that the system does not know what the user "intends" when a specific term is used in a query. To be effective the system should learn this. The system should provide the means for discovering, often with user guidance, the intended meaning. The EVOC[2] project at the University of Virginia has this goal.

The conventional attack on this problem is to attempt to enforce a controlled vocabulary for use in searching. This approach has many shortcomings. History has shown repeatedly that conformity cannot be legislated. Nevertheless many approaches to this problem rely implicitly on conformity. Our approach is to use strategies to mitigate the problem. We understand that any controlled vocabulary, however well-intentioned, will slowly grow out of date. Our approach uses adaptive methods to continuously evolve the vocabulary to increase its utility for searching.

We illustrate some of the difficulties with an example. We conducted a search using the EOS Data Gateway at GSFC[3] in search of data on "atmospheric pollution." We were provided with a search interface that allowed us to enter Earth

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

[1]The EDG is the interface for searching and ordering Earth science data products from NASA and affiliated centers. (http://redhook.gsfc.nasa.gov/~imswww/pub/imswelcome/

[2]http://www.cs.virginia.edu/~cyberia/EVOC/

[3]The interface to the EOS Data Gateway has been changed since this search was performed but the search is still a useful illustrative example.

science "terms." A query on "atmospheric pollution" resulted in the following message.

No data sets were found that have terms matching your query. Try typing in another term or picking from the list of valid terms.

It is inconceivable that there are no data related to the topic of our query. However, we were offered no help beyond being told to try a term from the list of valid terms. "Valid terms" refers to the controlled vocabulary used to index the datasets and available to the user as the search vocabulary. Within the EOS Data Gateway these are known as "valids." This is the principal problem: the users may not be familiar with the valid terms.

When we recast the query as the disjunction "atmospheric; pollution" we are told: "*Found 89 datasets corresponding to the terms matching your query.*" The valid terms (valids) identified are the following.

```
ATMOSPHERIC EMITTED RADIATION
ATMOSPHERIC HEATING
ATMOSPHERIC PRESSURE
ATMOSPHERIC STABILITY
HYDROLOGIC ATMOSPHERIC PILOT EXPERIMENT
```

But, we are not made aware of the valid `TROPICAL OCEANS AND GLOBAL ATMOSPHERE` because we have used "atmospheric" as one of our terms. The term "pollution" does not appear in any of the valid terms. Apparently we are at a deadend without some additional knowledge.

A new attack starting with the query "dust" yields two data sets cataloged under the valid term `DUST/ASH`. We ask for "detailed document" for the data set `GTE-A3A-TOWER` and reach a page with the link *LDAAC-datacenter.html*. After accessing that document we find in the first paragraph a reference to "the Measurement of Atmospheric Pollution from Satellites (MAPS) satellite experiments." Now there is indeed a "valid term" `MAPS`, but clearly it imparts a different intuition (i.e., geographic or other kinds of maps), and moreover it is also a substring of the valid term `FAO SOIL MAPS`.

A query on the term `MAPS` results in 27 data sets, 21 of which are soil related and irrelevant to our search. So we are left with 6 data sets that bear on atmospheric pollution. The document referenced above also made mention of "aerosols" and that might be a fruitful search direction to try next. Other possible valid terms are "gas" (`GAS EXCHANGE SYSTEM, NDIR GAS ANALYZER, SOIL GAS, SOIL GAS CHAMBER, TRACE GASES`) and "particulate" (`PARTICULATE MATTER`) and, of course, there might be others. The chief difficulty is that there is really no way to know which of the valid terms might be fruitful. What is needed is a way to expand the search vocabulary to encompass a broader range of terms so that users can express their needs more naturally. Moreover, this should be done by a process that accretes metadata incrementally in the normal operation of the system rather than requiring expensive manual campaigns to populate metadata.

There is another interesting detail to note. When we asked the system for the definition of the data set, MAPS-OSTA3-CO5X5-HDF, we were told "Measurement of Air Pollution from Satellites Office of Space and Terrestrial Applications - 3 (OSTA3) Carbon Monoxide 5 degree by 5 degree data in Hierarchical Data Format." So it seems that the "AP" in `MAPS` might also be "air pollution" and we should consider that as another search term along with "atmospheric pollution," although it too is not a valid term.

Searching should not have to be this laborious. Search aids are clearly indicated. The process should not hinge on the persistence and ingenuity of the searcher. The system should provide assistance in the form of suggestions, perhaps learned from earlier user interactions, or perhaps derived from processing ancillary documentation, or from both approaches.

The existing search interface (EOS Data Gateway) offers a list of 785 "valid terms." These constitute the totality of the controlled vocabulary. As we have seen the terms are not complete nor are they intuitive. Many will only be accessible by specialists. This situation leads to ambiguity in data requests. More specificity is necessary and a broader vocabulary is the means to that end. The key is to maintain relationships between the valid terms (controlled vocabulary) and the newly admitted terms (uncontrolled vocabulary). By this device we believe we can help guide users to relevant data.

```
 ERB-SCANNER
ERBE-SCANNER
ERBE
ERBE-NONSCANNER
ERBS
```

(a)

```
PHOTOSYNTH. ACTIVE RADIATION
PHOTOSYNTHESIS ACTIVE RADIATION
```

(b)

```
FIRE OCCURANCE
FIRE OCCURRENCE
```

(c)

Figure 1: Examples of types of string variants: (a) variety; (b) abbreviation; and (c) misspelling.

The list of valid terms suffers from some of the usual problems with such metadata. Figure 1 gives examples of three types of problems. Figure 1(a) shows minor variants of the same concept or its negation while Figure 1(b) and Figure 1(c) show variation due to abbreviation and misspelling respectively.

Figure 2 shows another problem that users encounter. The terms in the vocabulary capture semantic nuances and as a result are often related in subtle ways. The figure shows how we are attempting to help users visualize this aspect of the vocabulary. Using techniques for automating the construction of authority files [6, 7, 8, 10] that we developed in collaboration with astronomers studying publication trends

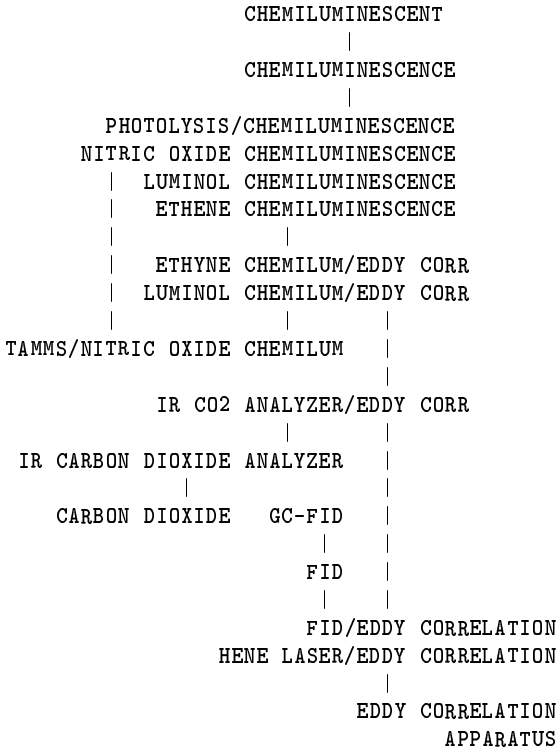

Figure 2: Visualizing the indexing vocabulary.

in their discipline [22, 23, 24], we can build a network of term relationships. Figure 2 depicts a fragment of this network. The hope is that users will find related terms by browsing this vocabulary term network.

So the objective of our work was to build a prototype search assistant to help guide users to relevant data. The search assistant has two key aspects. It is preloaded with relationships derived from existing descriptive metadata associated with each dataset and it will have the capacity to accept user-suggested relationships. This latter capability may be explicit: users state relationships, or implicit: the system infers relationships. We have implemented the first capability into the search assistant; the second is ongoing research.

2. THE VISION

We are building a prototype search assistant for use by EOSDIS users. We are creating thesauri for the purpose of expanding the meaningful search vocabulary available to users by: mining the EOSDIS for descriptive documents related to each dataset; and building associations between valid terms and the expanded vocabulary that is mined from the descriptive documents.

Our general approach to search assistance proceeds as follows. As a user engages in a search session we will have the system monitor the queries being submitted. If a user requests assistance, we will use the current query to search the thesaurus for "similar" terms. The user's query might be similar to several term classes so we will need to interactively decide among them by displaying the candidates to the user and requesting feedback as to the most appropriate term class(es); that is, the term class(es) that most nearly capture the user's intent. The system will take four steps after this interaction: (1) use the new terms to augment the user's query; (2) use the augmented query to search the metadatabase built by mining the EOSDIS for descriptive text; (3) display the results of the metadatabase query; and (4) record the new association in the metadatabase for use in subsequent term suggestion.

The prototype metadatabase is an auxiliary index formed during the initial mining phase and maintained incrementally during subsequent insertion of new data into the system. It associates terms with datasets, terms that have been derived from descriptive documentation supplied with each dataset or that have been explicitly nominated by some human indexer. Of course there needs to be some editorial review process and the system will be able to track new additions, modifications, or deletions for the purpose of having them confirmed by some authority before they are actually accepted. Users can be alerted to the fact that some associations are provisional if that seems appropriate.

The metadatabase serves to offer term suggestions and to provide a broader search vocabulary. So-called "valid terms" can still be used as in the past, but now we can associate an arbitrary number of new terms to a dataset and bring the full power of modern information retrieval techniques to bear on the retrieval problem. We will use inexact matching techniques, specialized clustering algorithms, and other machine learning techniques to build and refine the thesaurus classes maintained in the metadatabase.

To help illustrate our strategy we offer one possible interaction scenario based on the example in the introduction. Suppose a user enters:

```
atmospheric pollution
```

The system will respond: *There are no terms exactly matching your query. In the past the query,*

```
<MAPS and not (FAO SOIL MAPS)>
```

was useful in answering a similar query. You might also try:

```
<AEROSOL>
<air pollution>
<DUST/ASH>
      .
      .
      .
```

In the example above, things enclosed in < brackets > are intended to denote links; ALL CAPS denotes valid terms while lower case is a term from the expanded vocabulary.

One of our approaches is to build an initial thesaurus using the clustering algorithms that we developed in our authority control work. These algorithms are robust to spelling

variants, spelling errors, transliteration variants, etc. This initial thesaurus will serve as the starting point for expanding the search vocabulary. At a user's request we will search the thesaurus for terms present in the query and display alternative terms for the user's consideration. The new terms can be incorporated into the search query at the user's discretion. (An example of this is given by the terms `AEROSOL`, `air pollution`, and `DUST/ASH` above.)

Many users will search in an iterative fashion, reformulating queries to refine their information needs. We will log this activity to determine the initial and final queries. Subsequent user queries having similar initial formulations can be informed by the experience of earlier searchers. The system will provide users the opportunity to view queries that in the past proved useful given the initial query. Gathering these relationships, i.e., the association between an initial query (e.g., "atmospheric pollution" above) and its final formulation (e.g., "`MAPS and not (FAO SOIL MAPS)`" above), might be fully automated, might involve human interaction, or both. We will provide a mechanism for manual intervention and assess the utility of that mechanism in the search environment.

3. THIS WORK IN CONTEXT

Our initial investigation began with an examination of the EOS Data Gateway interface to NASA's Earth science data. The vignette played out in the introductory part of this paper grew out of direct experience with that interface. It should be noted, however, that substantial advances have been made toward the solution of the problems outlined in the introduction. These solutions have been implemented in the Global Change Master Directory (GCMD)[4] interface to Earth science data[16]. The GCMD is a comprehensive interface to a much larger array of Earth science data and includes as a subset the data stored in the NASA Distributed Active Archive Centers. We describe the functionality of the GCMD briefly next.

The GCMD maintains a carefully organized set of subject headings hierarchically organized into: category, topic, term, variable, and optionally, detailed variable. Each dataset has one or more subject headings assigned and these are recorded in the DIF[5] record associated with the dataset. An example follows.

```
Category: EARTH SCIENCE
Topic:    RADIANCE OR IMAGERY
Term:     INFRARED WAVELENGTHS
Variable: BRIGHTNESS TEMPERATURE
```

The category, topic, term, and variable components of the hierarchy constitute the controlled indexing vocabulary. The detailed variable is the means by which users may extend the hierarchy. Detailed variables are user assigned and constitute an uncontrolled vocabulary. In addition, users may assign keywords that they think will be useful in subsequent retrieval of the datasets. Again, the keywords are not controlled. This philosophy of combining a controlled and uncontrolled indexing vocabulary provides a richer search vocabulary and helps overcome some of the problems identified earlier.

Part of the problem with the EOS Data Gateway stems from the way in which it employs the GCMD subject hierarchy. Rather than adopt the GCMD controlled vocabulary in full, the developers of the EOS Data Gateway have chosen to use only the variable component, i.e., the leaf level of the hierarchy. These indexing terms constitute the "valids" used by the EOS Data Gateway. This decision results in the loss of precision and specificity of the indexing vocabulary and introduces ambiguity.

The GCMD offers Earth science users many different search strategies. The home page provides a search of the subject hierarchy together with a carefully organized set of entry points into the hierarchy that are designed to provide users with very fast access to a large array of the data. In addition, the GCMD maintains a number of experimental search interfaces. These are enumerated below together with a sketch of the associated functionality.

- Free-text Search (Isite)
  - Navigate by typed-in text.
  - Search whole document or specific fields.
  - Map applet embedded for spatial search.
- Hierarchical Keyword Search
  - Navigate by a hierarchy of keywords.
  - Search can be narrowed by a typed in word or phrase.
  - Start from Science Parameters, Location, Platform (Satellite), Instrument (Sensor)
- Supplementary Information Guided Search
  - Search descriptions of Data Centers, Campaigns or Projects, Sources(Platforms), and Sensors (instruments).
  - Point and click only – no typing allowed.
- Java Matrix Search Applet
  - Preview database contents with 10 different categories.
  - Categories dynamically update as you narrow your search.
  - Instantly view relevant titles while searching.
- Query Preview Search Applet
  - Preview by Topic, Time, Location
  - Refine by Source, Sensor, Data Center
  - Several alternative overviews of entire database.
- Advanced Field-based Search
  - Navigate via multiple fields.
  - Typing allowed.

[4] *www.gcmd.nasa.gov*

[5] Data Interchange Format, a cataloging record that contains much of the metadata associated with a dataset.

– Map applet embedded for spatial search.

The performance of the interfaces varies from moderate to extremely fast.

In addition to these interfaces, the GCMD would like to integrate a thesaurus capability into its search processes. We are presently examining ways to access the DLR[6] thesaurus from any search interface offered by the GCMD.

Despite its wide array of search interfaces, the GCMD is constantly pursuing better access to its holdings. The experimental work reported in this paper is complimentary to the ongoing GCMD work and will hopefully add useful functionality in some areas. Thus, the work reported here constitutes an alternative attack on problems of interest to the GCMD.

4. ABSTRACTING THE PROBLEM

For our purposes we assume that science users pose queries of the following general form:

$$\mathcal{Q} = \mathcal{K} \wedge \mathcal{P} \wedge \mathcal{T}.$$

That is, users are interested in some kind ($\mathcal{K}$) of data observed at some place ($\mathcal{P}$) at or over some time ($\mathcal{T}$). The spatial ($\mathcal{P}$) and temporal ($\mathcal{T}$) aspects are only pertinent once you have identified the kind of data of interest. They can be employed to reduce the space of kinds of data as has been done by Plaisant *et al.*[17], but we are not concerned with that here. In our work the focus is on improving the ability of searchers to locate relevant data. Accordingly we focus on improving the effectiveness of ($\mathcal{K}$) with the understanding that spatial and temporal filters may be employed as pre or post processes.

5. RELATED WORK

Computer systems that depend on words for their correct operation suffer from the "vocabulary problem" described by Furnas *et al.*[9, p. 964].

> Many functions of most large systems depend on users typing in the right words. New or intermittent users often use the wrong words and fail to get the actions or information they want.

Simply stated this means that users will often use different words to access information than the designers of the systems anticipated. This is particularly acute in information retrieval systems, where indexers assign terms and searchers may use other terms. Entire books (c.f. Lancaster[14]) have been devoted to the topic.

Unfamiliar search vocabularies is a long standing problem in information retrieval[2]. The difficulty can be compounded by data quality issues[8].

The use of thesauri to help mitigate this problem dates to the earliest days of information retrieval systems[13]. A number of approaches have been proposed for automating the construction of thesauri (e.g., Crouch[4]), or linking different thesauri[1]. Our approach is centered around "concept spaces." These have appeared in several guises[3, 21] over the years. We elaborate our use of concept spaces and give examples of their use in Section 6.

Another related activity is automatically assigning index terms from a controlled vocabulary[15] or expanding the keywords associated with data items (e.g., Garfield and Sher[11]). Gey *et al.*[12] have used entry vocabulary indexes (EVI) as a search aid. EVI's have been investigated by French *et al.*[5] for collection selection and document retrieval effectiveness.

We also intend to support "vocabulary switching" as discussed by Schatz[20]. Our approach is to use past user queries to help disambiguate current queries. This approach will let users establish a pertinent search context in which to pose their queries. This aspect of our work is ongoing and is not the focus of the present paper.

6. OUR APPROACH TO CONCEPT SPACES

The base concepts are as follows. NASA *datasets* are stored at Distributed Active Archive Centers (DAAC). These datasets record observations from *projects* and store these observations as *granules* which are data files and are the means of distributing the data to users. The granules often differentiate the data along spatial and/or temporal dimensions.

The datasets are described by subject heading metadata called *valids*. The valids are a controlled vocabulary and are used for searching within EOSDIS.

6.1 Our Model

DEFINITION 1. *A valids-set, $V = \{v_1, v_2, ..., v_k\}$, is the set of valids assigned to a dataset.*

DEFINITION 2. *$DS_V = \{DS_1, DS_2, ..., DS_n\}$ is the set of datasets labelled by the valids-set V.*

DEFINITION 3. *The similarity of DS_S and DS_T the sets of datasets labelled by valids-sets S and T respectively is given by*

$$sim(DS_S, DS_T) = sim(S, T) = \frac{|S \cap T|}{|S \cup T|}. \qquad (1)$$

Equation 1 is the Jaccard coefficient of similarity and measures the proportion of valids that the two datasets have in common. Here we are assuming that

$$sim(s,t) = \begin{cases} 1 & \text{if } s = t \\ 0 & \text{otherwise} \end{cases}$$

where $s \in S$ and $t \in T$ are individual valids. This supports exact matches among the valids. We can also to loosen this requirement to include partial matches among the valids. In that case we consider similarity coefficients like edit distance, $e(s,t)$, or Jaccard coefficient, $J(s,t)$, where each valid s and t is now treated as a set of words.

[6]The German national aerospace research center.

For our purposes, a concept space is an m-dimensional index space induced by a vocabulary of m indexing terms. Each indexed item is represented as a vector in this space. We are using the term "concept space" in preference to "vector space" to differentiate multiple spaces in which we conceptualize the datasets differently.

We are specifically interested in a *free text* space, t-space, and a *valids* space, v-space. The free text space is derived from descriptive text passages associated with datasets. The valids space is determined by the EOSDIS valids assigned to the datasets. Note that all the objects of interest to us (queries, datasets, and valids) are representable in each space. An example of each is given below.

6.2 Multiple Viewpoints

In earlier work, Powell and French[18] have demonstrated the potential of *multiple viewpoints* to increase retrieval effectiveness by enhancing the discovery process. We have explicitly provided a mechanism in our prototype for switching from a t-space search to one in v-space to examine the hypothesis that retrieval effectiveness can be improved by searching initially in one space and then switching to the other. The notions of t-space and v-space are made more concrete in the following sections.

6.2.1 *t-space*

We constructed a text space (t-space) by associating descriptive texts with datasets and then using the vector space model (VSM) of information retrieval[19]. In the VSM we represent an object, O_i as a vector, $(w_{i1}, w_{i2}, ..., w_{in})$, in an n-dimensional term space derived from the terms in all the objects. The vector component, w_{ij}, is a weight representing how well term j characterizes object i. We use a $tf \times idf$ weighting strategy where weights have the general form

$$w_{ij} = tf_{ij} \cdot \frac{N}{df_j}.$$

Here tf_{ij} is the *term frequency* of term j, that is, how often term j occurs in object i. The denominator, df_j is called the *document frequency* of term j and denotes the number of objects containing at least one instance of term j.

In our early experiments we associated descriptive texts from two sources with each dataset resulting in two distinct t-spaces. In the first we used selected sections from the Guide documents describing each dataset.[7] For the second we used the dataset summary taken from the DIF entry associated with each dataset. These two sources resulted in spaces with different properties. We are currently evaluating both for their suitability in the final search assistant. An example of a query in each space is shown below.

Figures 3 and 4 show the results of processing the query *atmospheric pollution* in the Guide-generated[8] and DIF-generated t-spaces respectively. Each search result is a ranked list of datasets. Although they produce different results, both strategies suggest data pertaining to atmospheric pollution. In Figure 3 the first and second ranked datasets are MAPS (Measurement of Atmospheric Pollution from Satellites) data. The first five ranked datasets in Figure 4 are MOPITT (Measurement of Pollution in the Troposphere) data, while the sixth and seventh ranked datasets are MAPS data. Both strategies suggest potentially relevant datasets, an improvement over the initial query response asserting that no data exists.

The Guide-generated t-space (Figure 3) often shows multiple datasets at each rank. This is because a Guide document often describes multiple datasets, and therefore we derive the same internal representation (vector) in the t-space for these datasets. This situation does not occur in the DIF-generated t-space (Figure 4) because we have a single DIF for each dataset. However, if two DIFs have the same summary section, their associated vectors will be the same.

Note that each figure shows the valids associated with each ranked dataset. This is to enable a transition from the t-space to the v-space described in the next section.

6.2.2 *v-space*

We form the v-space by creating a vector, $(v_1, v_2, ..., v_n)$, for each dataset where $v_k = 1$ if valid k is assigned to the dataset. We use the Jaccard coefficient given in Eqn. 1 to measure similarity in the v-space.

As stated in the last section, the valids associated with a dataset are used to transition from the t-space to the v-space. Figure 5 shows an example where the v-space has been entered with a focus on dataset D_1. We show the five most similar datasets to D_1 in the figure. The following conventions are used in Figure 5. ALL CAPS in the "matched" field of D_k indicates a term that has been assigned to both D_1 and D_k; lowercase indicates a term that is assigned to D_1 and not D_k. The "unmatched" terms are those assigned to D_k and not to D_1.

The user can select any dataset shown and "refocus" attention in the space to that dataset. In this way it is possible to explore neighborhoods of a dataset for other relevant data.

Note that we can represent individual valids in the t-space. By this device we can provide a transition to the t-space from the v-space. We can also enter the t-space via the current dataset under focus in the v-space. Both these entry mechanisms into the t-space can be used to support a multiple viewpoint interaction.

7. CONCLUSION

We are building a prototype search assistant with NASA support to help Earth science users deal with vocabulary problems in the EOS Data Gateway, that is, when the indexing vocabulary is unfamiliar to the searcher. We have described the status of our prototype and discussed a novel interaction paradigm called multiple viewpoints[18] which lets users investigate the system holdings via different indexing spaces.

We hope to improve search efficiency and effectiveness for end users, and hence, increase scientific productivity. We

[7] Guide documents are texts that serve as dataset documentation. These documents are available online.

[8] The sixth entry in Figure 3 has been truncated manually to include more entries in the space allowed for the figure.

expect that searchers will have an easier time locating data, and in some cases, we expect that searchers will find data that they might never have located without the search assistant.

It is also important to reiterate that this process accumulates metadata incrementally in the normal operation of the system rather than requiring expensive manual campaigns to populate metadata. This approach keeps the system current even in the face of declining budgets. Our prototype work has taken steps in this direction. We still need to evaluate the effectiveness of the system and that promises to be a significant challenge.

We are also collaborating with the GCMD to make this search assistant available to Earth science users through one or more of its existing interfaces and hope that we can contribute to the large body of work already done by the GCMD to solve the vocabulary problems faced by its users. In the end we hope that we will help enable EOSDIS to be the foundation upon which more and better science will be conducted.

8. REFERENCES

[1] S. Amba, N. Narasimhamurthi, K. C. O'Kane, and P. M. Turner. Automatic Linking of Thesauri. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, pages 181–186, Zurich, Switzerland, 1996.

[2] M. Buckland, A. Chen, H. Chen, F. Gey, Y. Kim, B. Lam, R. Larson, B. Norgard, and Y. Purat. Mapping Entry Vocabulary to Unfamiliar Metadata Vocabularies. *D-Lib Magazine*, 5(1), January 1999.

[3] H. Chen, T. D. Ng, J. Martinez, and B. R. Schatz. A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System. *Journal of the American Society for Information Science*, 48(1):17–31, 1997.

[4] C. J. Crouch. An Approach to the Automatic Construction of Global Thesauri. *Information Processing and Management*, 26(5):629–640, 1990.

[5] J. C. French, A. L. Powell, F. Gey, and N. Perelman. Exploiting A Controlled Vocabulary to Improve Collection Selection and Retrieval Effectiveness. In *Proceedings Tenth International Conference on Information and Knowledge Management (CIKM 2001)*, pages 199–206, 2001.

[6] J. C. French, A. L. Powell, and E. Schulman. Applications of Approximate Word Matching in Information Retrieval. In *6th International Conference on Information and Knowledge Management (CIKM'97)*, pages 9–15, Las Vegas, Nevada, 10-14 November 1997.

[7] J. C. French, A. L. Powell, and E. Schulman. Using Clustering Strategies for Creating Authority Files. *Journal of the American Society for Information Science*, 51, 2000. To appear.

[8] J. C. French, A. L. Powell, E. Schulman, and J. L. Pfaltz. Automating the Construction of Authority Files in Digital Libraries: A Case Study. In C. Peters and C. Thanos, editors, *First European Conference on Research and Advanced Technology for Digital Libraries*, volume 1324 of *Lecture Notes in Computer Science*, pages 55–71, Pisa, 1-3 September 1997. Springer-Verlag.

[9] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The Vocabulary Problem in Human-System Communications. *Communications of the ACM*, 30(11):964–971, November 1987.

[10] V. Ganti, R. Ramakrishnan, J. Gehrke, A. Powell, and J. French. Clustering Large Datasets in Arbitrary Metric Spaces. In *15th International Conference on Data Engineering (ICDE'99)*, pages 502–511, Sydney, March 1999.

[11] E. Garfield and I. H. Sher. KeyWords Plus — Algorithmic Derivative Indexing. *Journal of the American Society for Information Science*, 44(5):298–299, 1993.

[12] F. Gey, M. Buckland, A. Chen, and R. Larson. Entry Vocabulary – A Technology to Enhance Digital Object Search. In *Proceedings of the First Internation Conference on Human Language Technology*, 2001.

[13] T. Joyce and R. M. Needham. The Thesaurus Approach to Information Retrieval. *American Documentation*, 9:192–197, 1958.

[14] F. W. Lancaster. *Vocabulary Control for Information Retrieval, (2nd. edition)*. Information Resources Press, 1986.

[15] C.-H. Leung and W.-K. Kan. A Statistical Learning Approach to Automatic Indexing of Controlled Index Terms. *Journal of the American Society for Information Science*, 48(1):55–66, 1997.

[16] L. Olsen. Helping Users Overcome Vocabulary Barriers in the GCMD, 2002. In preparation.

[17] C. Plaisant, B. Sneiderman, K. Doan, and T. Bruns. Interface and Data Architecture for Query Preview in Networked Information Systems. *ACM Transactions on Information Systems*, 17(3):320–341, 1999.

[18] A. Powell and J. C. French. The Potential to Improve Retrieval Effectiveness with Multiple Viewpoints. Technical Report CS-98-15, Department of Computer Science, University of Virginia, 1998.

[19] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, 1983.

[20] B. R. Schatz. Information Retrieval in Digital Libraries: Bringing Search to the Net. *Science*, 275(1):327–334, January 1997.

[21] P. Schauble. Thesaurus Based Concept Spaces. In *Proceedings of the 10th International Conference on Research and Development in Information Retrieval*, pages 254–262, New Orleans, LA, 1987.

[22] E. Schulman, J. C. French, A. L. Powell, G. Eichhorn, M. J. Kurtz, and S. S. Murray. Trends in Astronomical Publication Between 1975 and 1996. *Publications of the Astronomical Society of the Pacific*, 109:1278–1284, 1997.

[23] E. Schulman, J. C. French, A. L. Powell, S. S. Murray, G. Eichhorn, and M. J. Kurtz. The Sociology of Astronomical Publication Using ADS and ADAMS. In G. Hunt and H. Payne, editors, *Astronomical Data Analysis Software and Systems VI, volume 125 of the Astronomical Society of the Pacific Conference Series*, pages 361–364, 1997.

[24] E. Schulman, A. L. Powell, J. C. French, G. Eichhorn, M. J. Kurtz, and S. S. Murray. Using the ADS Database to Study Trends in Astronomical Publication. *Bulletin of the American Astronomical Society*, 28(4):1281, 1996.

```
Search terms: atmospheric pollution

1: 0.23
dataset: MAPS_OSTA3_COSEC_HDF_STS-41-G_GFC RADIOMETER
         MAPS_SRL1_COSEC_HDF_STS-59_GFC RADIOMETER
         MAPS_SRL2_COSEC_HDF_STS-68_GFC RADIOMETER
valids:  CARBON MONOXIDE;

2: 0.23
dataset: MAPS_OSTA3_CO5X5_HDF_STS-41-G_GFC RADIOMETER
         MAPS_SRL1_CO5X5_HDF_STS-59_GFC RADIOMETER
         MAPS_SRL2_CO5X5_HDF_STS-68_GFC RADIOMETER
valids:  CARBON MONOXIDE;

3: 0.11
dataset: CZCS LEVEL 1 FULL RESOLUTION_NIMBUS-7_CZCS
valids:  RADIANCE;

4: 0.11
dataset: CZCS LEVEL 1A GAC_NIMBUS-7_CZCS
valids:  RADIANCE;
dataset  CZCS LEVEL 2 BROWSE_NIMBUS-7_CZCS
valids:  PIGMENTS;
dataset: CZCS LEVEL 2 GAC_NIMBUS-7_CZCS
valids:  AEROSOL RADIANCE; LIGHT ATTENUATION; PIGMENTS; WATER-LEAVING RADIANCE;

5: 0.10
dataset: CZCS LEVEL 3 DAILY PST_NIMBUS-7_CZCS
         CZCS LEVEL 3 MONTHLY COMPOSITE PST_NIMBUS-7_CZCS
         CZCS LEVEL 3 MONTHLY COMPOSITE_NIMBUS-7_CZCS
         CZCS LEVEL 3 WEEKLY PST_NIMBUS-7_CZCS
valids:  AEROSOL RADIANCE; PIGMENTS; LIGHT ATTENUATION; WATER-LEAVING RADIANCE;

6: 0.09
dataset: TARFOX_UWC131A_UW C131_CLOUD CHAMBER
valids:  CLOUD CONDENSATION NUCLEI; NUCLEATION;
dataset: TARFOX_UWC131A_UW C131_ETHENE CHEMILUMINESCENCE
valids:  OZONE;
dataset: TARFOX_UWC131A_UW C131_FSSP
valids:  DROPLET CONCENTRATION/SIZE;

7: 0.09
dataset: ERS-1 SAR IMAGES - FULL RES_ERS-1_AMI-SAR; SAR; RADAR
         ERS-1 SAR IMAGES - LOW RES_ERS-1_AMI-SAR; SAR; RADAR
         ERS-2 SAR IMAGES - FULL RES_ERS-2_AMI-SAR; SAR; RADAR
         ERS-2 SAR IMAGES - LOW RES_ERS-2_AMI-SAR; SAR; RADAR
valids:  RADAR CROSS-SECTION; RADAR BACKSCATTER; RADAR IMAGERY;
```

Figure 3: Datasets returned for the query "atmospheric pollution." The dataset representations were mined from Guide documents.

```
Search terms: atmospheric pollution

1: 0.23
dataset: MOPITT Level-3 Data (Gridded CH4 Total Column): MOP05
valids:  METHANE;

2: 0.22
dataset: MOPITT Level-3 Data (Gridded CO Total Column): MOP07
valids:  CARBON MONOXIDE;

3: 0.22
dataset: MOPITT Level-3 Data (Gridded CO Mixing Ratios): MOP06
valids:  CARBON MONOXIDE;

4: 0.22
dataset: MOPITT Level-2 Data from EOS Terra (MOP02)
valids:  CARBON MONOXIDE; METHANE;

5: 0.20
dataset: MOPITT Level-1 Data from EOS Terra (MOP01)
valids:  AEROSOL RADIANCE; OUTGOING LONGWAVE RADIATION; INFRARED IMAGERY;

6: 0.14
dataset: Measurement of Air Pollution from Satellites (MAPS)
         Space Radar Laboratory - 1 (SRL1) Carbon Monoxide
         Second by Second data
valids:  CARBON MONOXIDE;

7: 0.14
dataset: Measurement of Air Pollution from Satellites (MAPS)
         Space Radar Laboratory - 1 (SRL1) Carbon Monoxide
         5 degree by 5 degree data
valids:  CARBON MONOXIDE;

8: 0.11
dataset: Priority Programme for China's Agenda 21
valids:  AGRICULTURAL RESOURCES; ENVIRONMENTAL INDICATORS; INDUSTRIAL
         RESOURCES; AGRICULTURAL EQUIPMENT; FARM STRUCTURES; CROPPING SYSTEMS;
         DAIRY PRODUCTS; LIVESTOCK PRODUCTS; POULTRY PRODUCTS; ANIMAL
         MANAGEMENT SYSTEMS; FIELD CROPS PRODUCTS; FRUIT PRODUCTS;
         HORTICULTURAL PRODUCTS; VEGETABLE PRODUCTS; AGRICULTURAL ECONOMICS;

9: 0.09
dataset: Directory of EuroMAB Biosphere Reserves
valids:  CLIMATE CHANGE; LAND CHARACTERISTICS;

10: 0.09
dataset: Atmospheric Profiles: TOVS - NOAA (FIFE)
valids:  OZONE; ATMOSPHERIC PRESSURE; AIR TEMPERATURE; CLOUD AMOUNT;
```

Figure 4: Datasets returned for the query "atmospheric pollution." The dataset representations were mined from DIF entries.

```
D1: 1.000
  Matched:
     AEROSOL EXTINCTION; AIR TEMPERATURE; CARBON MONOXIDE; METHANE;
     NITROGEN DIOXIDE; NITROGEN OXIDES; NITROUS OXIDE; OZONE; TRACE GASES;
     WATER VAPOR;
D2: 0.800
  Matched:
     AEROSOL EXTINCTION; AIR TEMPERATURE; carbon monoxide; METHANE;
     NITROGEN DIOXIDE; NITROGEN OXIDES; nitrous oxide; OZONE; TRACE GASES;
     WATER VAPOR;
D0: 0.643
  Matched:
     AEROSOL EXTINCTION; AIR TEMPERATURE; carbon monoxide; METHANE;
     NITROGEN DIOXIDE; NITROGEN OXIDES; NITROUS OXIDE; OZONE; TRACE GASES;
     WATER VAPOR;
  Unmatched:
     altitude; chlorine nitrate; chlorofluorocarbons; nitric acid;
D13: 0.545
  Matched:
     aerosol extinction; AIR TEMPERATURE; carbon monoxide; methane;
     NITROGEN DIOXIDE; NITROGEN OXIDES; nitrous oxide; OZONE; TRACE GASES;
     WATER VAPOR;
  Unmatched:
     nitric acid;
D11: 0.500
  Matched:
     aerosol extinction; AIR TEMPERATURE; carbon monoxide; methane;
     NITROGEN DIOXIDE; NITROGEN OXIDES; nitrous oxide; OZONE; TRACE GASES;
     WATER VAPOR;
  Unmatched:
     geopotential height; nitric acid;
D7: 0.417
  Matched:
     aerosol extinction; AIR TEMPERATURE; carbon monoxide; METHANE;
     nitrogen dioxide; NITROGEN OXIDES; NITROUS OXIDE; ozone; TRACE GASES;
     water vapor;
  Unmatched:
     chlorofluorocarbons; nitric acid;
```

Figure 5: Similar datasets in v-space using Jaccard coefficient. Fifteen datasets have nonzero similarity to dataset D_1; only the first five, $\{D_2, D_0, D_{13}, D_{11}, D_7\}$, are shown here. ALL CAPS in the "matched" field of D_k indicates a term that has been assigned to both D_1 and D_k; lowercase indicates a term that is assigned to D_1 and not D_k. The "unmatched" terms are those assigned to D_k and not to D_1. The value of the Jaccard coefficient is also shown.