TERMINOLOGY PROJECT REPORT

Contract Number: EP02-332

Prepared for:

U.S. Environmental Protection Agency
Office of Environmental Information

Delivery Order Project Officer:
Jonda Byrd

Prepared by:

Joel Tochterman (GCI Information Services)
Valorie Lee (GCI Information Services)

May 2000

TABLE OF CONTENTS

Introduction to the Terminology Project Report

This report was written to provide the EPA management with an outline of the work performed by the GCI Terminology Team in support of the TRS (Terminology Reference System) and GEMET (GEneral European Multilingual Environmental Thesaurus) projects. The Report provides an overview of the context of the Team's work, work procedure and work products. The Report is meant to be both informative and evaluative. Included are the Team's observations of both TRS and GEMET, focusing on data quality and potential areas for improvement and future work.

Report highlights include an introduction to TRS version 1.1 in Section 2.3, a summary of planned data collection methods for TRS 2.0 in Section 5.2, and conclusions and data development possibilities in Section 6.0.

Information for this report was gathered from a variety of documents including conference presentations, reports, scholarly articles, internal correspondence and correspondence with Agency partners, and meeting notes.

Appendices are referenced in the text and include the 57 work product sets submitted to EPA (Linda Spencer and Brand Niemann) and to Consiglio Nazionale delle Ricerche (CNR), GEMET focal point during the course of the Team's involvement with the GEMET project. These sets include definitions and notes written by the team as well as a table of consulted sources.

2.0 History/Background

2.1 GEMET (Introduction)

GEMET, the GEneral Multilingual Environment Thesaurus was originally developed as an indexing, retrieval and control tool for the Catalogue of Data Sources (CDS) of the European Environmental Agency (EAA). It featured more than 6,500 controlled terms in English, German, Italian, Dutch, Spanish, French, Norwegian and Danish. Because of the different language needs, GEMET was designed as a "general" thesaurus in which a core of general terminology for the environment could be used throughout Europe.

GEMET was compiled by merging environmental thesaurus documents from Germany, Italy, Spain, France and the Netherlands and also UNEP's EnVoc Thesaurus and descriptor terms from relevant EEA documents. This complex task included the identification and scoring of coinciding environmental concepts. Experts from the organizations involved selected which core environmental terms would be included in GEMET and then organized the terms so users could access the thesaurus through a group hierarchical list, a thematic list or an alphabetical list.

2.1a The EPA's Involvement with GEMET

GEMET's potential as a global thesaurus has long been recognized by non-European countries. Sometime before the release of GEMET 1.0 in December 1998, the EEA and U.S. EPA agreed to collaborate in expanding and refining GEMET. This was the beginning of a move towards a thesaurus that many envisioned would be used globally.

GEMET was seen as a way to facilitate the global dissemination and exchange of environmental information or data. As part of this effort, the U.S. EPA has actively worked with the Asia and Pacific Economies Cooperation (APEC). It was hoped that APEC economies would cooperate in the expansion of GEMET and would utilize that system as a standard reference language and would tailor it to regional needs. Chinese Taipei has already headed an effort to translate the GEMET core into Mandarin. Proposals have been launched to include the translations into Arabic, Bahasa (Indonesian), Thai and Vietnamese.

For its own purposes, U.S. EPA believed GEMET had potential as a reference thesaurus that would offer assistance in organizing or bringing together U.S. EPA's terminology. It was adopted as the reference thesaurus for the Terminology Reference System (TRS). GEMET's List of Lists (GloL) is used to link to TRS and other terminology collections.

As a reference tool, it was believed that GEMET would have a number of different functions. It would assist U.S. EPA's cataloging and indexing efforts, particularly for its Web based inventories. It would also assist in harmonizing and standardizing

terminology. Plans were made to incorporate GEMET terminology within U.S. EPA's search engine in order to improve information retrieval.

GEMET was also seen as a reference for the organization of environmental terminology into classification schemes, ontologies, glossaries and other terminology sources. It would act as a map of environmental knowledge indicating how concepts relate to each other. Its definitions would serve as a guide in mapping terms and synonyms to concepts and database fields. It would also serve as a resource in the translation of environmental documents.

As part of this effort, U.S. EPA contracted with GCI to assist in the development of definitions and the incorporation of American equivalent terms. Earlier work by GCI involved the evaluation of GEMET terms for an American English version. GCI was provided with a list of British English terms. It was asked to assist in changing the British spellings and in changing plural terms to singular terms. GCI also drafted volunteers to assist in mapping GEMET terminology with the Concerned Citizens Controlled Vocabulary for the Web Inventory cataloging vocabulary.

In April 1999, the U.S. EPA provided funding to expand the Terminology Team to three people. The two new members of the Team worked almost entirely on providing new or revised GEMET definitions in preparation for the final release of GEMET 2.0 in August 1999.

The Team has worked closely with the Consiglio Nazionale delle Ricerche (CNR) of Rome, Italy. In May and October 1998, Dr. Bruno Felluga and Dr. Paolo Plini of CNR spent a total of eight weeks at the U.S. EPA. They were involved in planning, organizing and carrying out the initial GEMET work. CNR has been responsible for handling and upkeeping the GEMET reference master file, including the foreign translations. They have provided terminology to GCI and they have reviewed GCI's suggestions and definitions before incorporating them in the American English version of GEMET. **[For further information, please see the attachment for Appendix I for CNR's Report.]**

2.1b Questionnaire

In November and December 1999, CNR made a questionnaire available to users of GEMET through the Catalog of Data Sources (CDS) web site at http://www.mu.niedersachsen.de/cds/ At the meeting in Santa Fe, the results of the questionnaire were reported. A high percentage of survey respondents reported satisfaction with GEMET. 71% said that GEMET is NOT "too much or overwhelming." 71% also said that there is NO need to delete some terms. A large percentage of users believe there is NO need for more specific terms. This high rate of satisfaction leads GEMET developers to believe that GEMET's "content should remain stable" and only "minor improvements are justified."

But 50% of the respondents would like to use the interface in their own language. 57% need connection in their own application. These findings lead GEMET developers to

believe that there is a “need to add new functionalities to the tools” so that users can “customize their own ‘thesaurus system.’”

2.1c Santa Fe Meetings/Agreement

Present at the GEMET meeting in Santa Fe on January 14, 2000 were principals representing the partners -- Consiglio Nazionale delle Ricerche (CNR), the US Environment Protection Agency (U.S. EPA), the European Environment Agency (EEA), the United Nations Environment Programme (UNEP).

The primary agenda item of the meeting was to actuate "globalization" of the GEMET Thesaurus. Discussion was devoted to a feasible application of this concept, and following a general consensus on this, a detailed work statement was put forth on the planning and procedure for this endeavor or new project phase. The partners decided upon the creation of a global thesaurus steering committee, the stabilization of GEMET to concentrate on translation work and the addition of national and regional extensions among other planned tasks.

On February 3, 2000, the EPA issued a press release titled "Common Global Environmental Vocabulary Being Developed" which provided a general explanation of the global thesaurus project and enumerated the benefits in information organization and access. **[Please see the attachment for Appendix II, the EPA Press Release, “Common Global Environmental Vocabulary Being Developed.”]**

2.2a TRS Background/History

An introductory paragraph on the TRS home page (**http://www.epa.gov/trs/**) dated March 26, 1999 reported that the Terminology Reference System (TRS) was created in collaboration with the European Environmental Agency’s GEneral Multilingual Environmental Thesaurus (GEMET) system. The paragraph also stated that TRS, using GEMET as its foundation, would provide a single master source of environmental terminology for the Agency by incorporating the use of EPA-specific terms and lists into the European product.

TRS, implemented by the Systems Development Center (SDC), has evolved through six releases. TRS Version 1.1, released in March 2000, is the most recent release. In the new version, TRS has expanded to include a wider range of environmental terms and the goal appears to be to include more data from the fifty states. In TRS 1.1, GEMET is utilized as a primary resource. But the GEMET Thesaurus as transported into TRS 1.1 contains only the English Language terms and no longer appears to be the central focus of TRS.

In a presentation given by Linda Spencer on January 19, 2000 in Santa Fe, New Mexico, TRS was envisioned as a repository of well-defined environmental concepts as a tool for cataloging data and documents, retrieving web documents, developing data elements and integrating databases.

2.2b SEDES

The State Environmental Data Exchange Strategy (SEDES) Project has played a role in the development of TRS.

The purpose of SEDES is to facilitate the sharing of state and federal environmental site remediation data. The partners involved with this project include U.S. EPA's OSWER/OERR Information Management Center (IMC), and the Association of State and Territorial Solid Waste Management Officials (ACTSWMO) with meeting facilitation provided by the Marasco-Newton Group (MNG). State remediation agencies that agreed to participate in SEDES include the Illinois Protection Agency, the Missouri Department of Natural Resources, the New Jersey Site Remediation Program, South Carolina and the Texas Natural Resources Conservation Commission. Other state remediation agencies have been invited to join in this work.

The participating states agreed to collaborate in providing hazardous waste terminology and definitions for TRS. The Terminology Team was tasked to collect this data. It built databases locally in order to create flat files for SDC to import into TRS.

The SEDES partners were involved in the Open Forum meetings that took place in Santa Fe, New Mexico. Larry Fitzwater of the U.S. EPA presented and discussed how states could share cleanup information with a wide range of potential users. TRS was one of the technologies displayed. The TRS presentation demonstrated how users could search, view and download terms, definitions and term collections supplied by the five state partners. It was explained that TRS is in a test-pilot stage and that it would be valuable as a tool for regulation writers and for legal researchers who need to compare terms.

2.2c TRS 1.0 September 1999

TRS 1.0 was released in September 1999. In response to needs by state partners from SEDES, the new version included hazardous waste terminology and definitions from five states, including Illinois, Missouri, New Jersey, South Carolina and Texas.

The TRS 1.0 home page includes a brief introductory paragraph, with its vision for TRS. It reports that TRS "has been created to provide a single source of terminology for the Agency by incorporating EPA-related terms and lists as well as other environmentally-related terminology sources." The paragraph also reports that TRS is "currently under development" and "contains only a few terms, but additional information is being accumulated."

Below the introductory paragraph is a section titled Glossary Search. This section includes links titled Alphabetical Search and Keyword Search. By clicking on Alphabetical Search, the next page would retrieve a row of linked alphabetical letters. Each alphabetical letter would lead to a list of linked terms. The linked terms would in turn lead to definitions listed along with their resource titles and authoring organizations.

The Keyword Search link leads to a search page in which users can type in search terms. Below the search term box are two sections titled Search Options and Selection Options. Under Search Options, users may choose between Containing, Beginning With or Exact Match. Under Selection Options, users may choose between Term Only, Definition Only or Both Term and Definitions. When search terms have been entered and the options have been selected, the user would then click on Begin Search. The user also has the option of clicking on a button that would Clear and Start Over.

At the TRS 1.0 home page, the bottom section includes a link for Search by Resource. Through this link, users can view a term or list of terms by specific resource from a list of available organizations, terminology resources or classification of terms.

2.3 TRS 1.1 March 2000

TRS 1.1 was released on March 31, 2000. This new version of TRS contains 19,642 terms and definitions.

The User's Guide in section 4 of this Report provides detailed information on how to use TRS 1.1. Screenshots of the database have been included within section 4.

Unlike TRS 1.0, the new version seems to be less cumbersome and allows users to access data by retrieving fewer web pages. The data arrangement has been altered so that TRS has more of a dictionary-like format.

TRS 1.1 includes a row of linked alphabetical letters on the Glossary Search page, which is at the TRS home page. In TRS 1.0, a user would need to click on Alphabetical Search in the Glossary Search page in order to retrieve this row of linked alphabets. It would require clicking on two more links in order to get to the definitions. But in TRS 1.1, an alphabetical search takes users directly to a listing of both the terms and definitions—just as if users were viewing a dictionary.

The Glossary Search page also includes a different look for the Keyword search feature. Under the row of linked alphabetical letters is a search box that users would use to search for all terms containing a search term. TRS 1.1 allows users to perform searches from the Glossary Search page. In TRS 1.0, the Glossary Search page required users to click on Keyword Search in order to go on to a Search page.

A user would also have the option of clicking on a link for a more advanced keyword search. The advanced keyword search page is the same as the TRS 1.0 search page—in which users have the option of selecting Search Options (Containing, Beginning with, or Exact Match) and Selection Options (Term Only, Definition Only, or Both Term and Definitions).

Search results no longer include duplicate definitions, as is noted at the top of each search results page.

As with TRS 1.0, a Resource Search option is available at the bottom of the Glossary Search page. A user would click on Resource in order to retrieve a comprehensive listing of resources. The Search by Resource section appears to be similar with its previous version, only more resources have been included.

Unlike TRS 1.0, the new version includes the GEMET resource. GEMET is included within the Search by Resource section. After clicking on the link for the Resource Search at the Glossary Search page, the next page would retrieve a wide selection of GEMET related links. Under the Organization column, the user could click on the link for the "European Environment Agency (EEA), European Topic Centre on Catalogue of Data Sources (ETC/CDS)" in order to retrieve information about the organization responsible for producing and maintaining the GEMET resource. Under the Resource column, the user could click on "General Environmental Multilingual Thesaurus (GEMET)" in order to retrieve information about the GEMET resource and also to access an overall viewing of the GEMET hierarchy. And finally under the classification column, users would have access to links for 35 GEMET topical groups and 40 GEMET themes.

By clicking on links for one of the GEMET topical groups, users would retrieve a display of that group's hierarchical arrangement. The hierarchical relationships are displayed by using a dot system. One dot refers to the highest level of terms. Two dots refer to the terms in a level below. While three dots refer to the level below the two dot terms. The dots range from one to five dots.

The dot system hierarchy displays linked GEMET terms. By clicking on a linked GEMET term, users would retrieve a page that displays the term's definition and links for the term's resource, classification, broader terms, narrower terms and related terms.

TRS 1.1 includes American English definitions that the Terminology Team submitted, as incorporated by CNR, for GEMET 2.0, which was released in August 1999.

The Resource section also includes the U.S. EPA "Terms of Environment" glossary. This widely used glossary defines terms in non-technical language for the more commonly used environmental terms appearing in U.S. EPA publications, news releases and other Agency documents to the general public, students, the media and Agency employees. Another new resource that has been included is a set of terms from the United States Code.

More resources have been included from the five participating state partners: Illinois, Missouri, New Jersey, South Carolina and Texas. These resources were collected by the Terminology Team and submitted to SDC after the release of TRS 1.0.

TRS 1.1 includes 1,889 U.S. EPA terms, 61 United States Code terms, 73 terms from the State of Illinois, 22 terms from the state of Missouri, 518 terms from the state of New Jersey, 122 terms from the state of South Carolina and 401 terms from the state of Texas. Additionally, there are 16,556 GEMET terms.

In the new format, after a search is executed, much more information is downloaded. Conceivably, rather than 25 linked terms listed on one screen, a single term and its definitions could fill an entire screen and could require a user to scroll down and track the term and definition needed. For instance, if a user needs definitions for "herbicide," this person could click on "H" in the alphabetical listing. The computer would then retrieve all terms beginning with "H." The "H" terms would be included with their definitions and resource information. To get to "herbicide," a user would need to scroll down past definition information for widely used terms such as "hazardous substances" and "hazardous waste." It is likely that the user would be required to scroll down through more than a few screens before he lands at "herbicide." This approach is somewhat similar to a user who needs to browse through a dictionary in order to find the term and definition that are needed.

At the Glossary research page, a "simple" keyword search would also download much more information and would require extensive scrolling. For example, a user searching for "waste" would retrieve all terms including "waste" such as "food waste," "hazardous waste," "solid waste management," and "wastewater." It is good that a user would be able to browse through potentially useful terms. But since both the terms and definitions are included on the same page, along with the resource information, the user might experience an overload of information and might be required to sift through unwanted information.

However, skilled users or those who have greater experience with TRS would utilize the advance search option in order to retrieve more precise search results.

TRS 1.0 allowed users to download terms and definitions. TRS 1.1 was modified to make it easier to download data for users of Microsoft Internet Explorer Version 3 and for all versions of AOL's browser prior to 5.0.

In general, the improvements made for TRS 1.1 will help it become a repository of well-defined environmental concepts. The database already includes a wide range of terminology—including data from five states, the U.S. EPA's Terms on Environment list, and GEMET. Its framework allows for it to incorporate much more data—especially data from the fifty states.

2.4 TRS Data Maintenance Facility 1.1

For the TRS database, the Terminology Team currently copies and pastes glossary terms/definitions into dBase files and sends the dBase files as flat files to the Systems Development Center (SDC). Since the Team needs a more efficient way of transporting the data, a Maintenance Facility System, Version 1.1, was developed by SDC.

The Team evaluated the Maintenance Facility System on March 9, 2000. It was discovered that this system was originally designed for GEMET. The system has the flexibility to incorporate GEMET's uniqueness, particularly its hierarchical relationships and its foreign translations. In addition, the system establishes relationships with a term's

definitions, its GEMET group category, its glossary or dictionary source name, the authoring organization for the glossary or dictionary source, and its ISBN number or URL addresses. Because of the need to establish different relationships, the System has numerous data entry screens in which data must be entered.

Clearly, the Maintenance Facility System was built to handle the complexity of GEMET. It may be more sophisticated than what is needed for TRS. Most glossaries or dictionaries that would be used as sources for TRS are not likely to have the same complex relationships as a multi-lingual thesaurus such as GEMET. These sources would have few data entry requirements.

As a tool for GEMET, the Team believes the Maintenance Facility System would be useful for editing. The Team could use it to revise previously entered GEMET definitions or to update information.

Currently, the maintenance facility system is only for internal usage at SDC. If it is to be used by the Terminology Team, it would need to be reconfigured so that it is compatible with U.S. EPA's computer system. The Terminology Team would also need assistance and guidance in how to use the maintenance system.

3.0 GEMET Definition Sets

3.1 Background (workflow)

3.1a Terminology Team's Tasks

Initially, the Terminology Team was tasked to work on revising English definitions written by the International Society for Environmental Protection (ISEP) from Vienna, Austria (http://www.isep.at). The developers of GEMET were eager to have these definitions revised in time for the release of GEMET 2.0, which was expected in August 1999.

The Team was provided with a dBase file that included all of the terms used in the GEMET hierarchy. This dBase file also included columns for the group category name, sources, notes, foreign translations, codes for preferred terms and ID numbers. A column was included to access the definitions by clicking on an icon. A column titled "XEPA" was used by CNR to indicate which ISEP terms/definitions needed to be revised.

It became clear that most of the assigned ISEP definitions required major revisions. They needed fine-tuning especially to ensure that they would be acceptable to American users of GEMET. If a term needed revising, the Team would revise the term and provide a note within the "NOTE" column in dBase that said "suggest change (defworkxx.doc)." In the Word document, the Team placed the term and newly revised definition under a section titled "New/Revised Definitions."

Some of these ISEP definitions were satisfactory. If they were satisfactory, the Team provided a note within the "NOTE" column in dBase that said "suggest to retain (defworkxx.doc)." In a Word document, the Team placed these retained terms within a section titled "Approved Definitions."

Sometime late in the summer of 1999, CNR asked us to work on a new set of definitions. In a new dBase file, they marked which terms needed to be revised by entering hash-mark symbols within a "Q" column.

In the winter of 1999, the Team completed its work for the hash-marked terms. A few weeks before the expected completion, the Team received permission to contact CNR to request a new set of terms to work on. It took approximately a month to receive a new set of terms. During this waiting period, the Team was tasked to revise previously submitted definitions. By the end of December, the Team received a list of annex terms. The Team learned that these terms are not currently part of the GEMET hierarchy. CNR will consider whether to enter these terms and definitions within the hierarchy at a later date. Some of the terms and definitions might remain hierarchy-less but could be entered in GEMET's alphabetical listing and its thematic list.

The annex terms presented a new challenge for the Team, since it could no longer rely on the hierarchy for clues in what a term might mean. It meant definitions needed to be defined broadly to incorporate all possible meanings.

3.1b Definition Writing Procedure

Typically, the Team follows a certain procedure for defining GEMET definitions. Each Team member is assigned a topic category (for example, "Atmosphere [air, climate]"). Using dBase, the Team member creates a dBase Report for this particular topic category. When printed out, the Report lists the GEMET terms and indicates where the term fits within the hierarchy (using a dot system). Also displayed on this Report are symbols, such as hash-marks, which highlight the terms needing revisions or new definitions.

The first step in defining a term is to study how it fits within the GEMET hierarchy. For this purpose, we refer to ThesShow. As a demonstration, we could use the term "waste disposal." In ThesShow, we would refer to the alphabetic search and we would type in "waste disposal." Then we would click on "Systematic" to retrieve the systematic arrangement. The systematic arrangement would tell us that "waste disposal" is placed under "overburden." But it is a broader term to eleven other terms, including such terms as "battery disposal," "dumping," "incineration of waste," and "waste water disposal." **[For further information about ThesShow, please refer to the ThesShow slide show at** http://www.mu.niedersachsen.de/cds/etc-cds_neu/thes_slide_home.html**]**

It is important to determine how "waste disposal" is used in the GEMET hierarchy. The Team would compare it with its broader and narrower terms and also any of its sibling terms or terms on the same level. A term often has several different meanings. So by looking at the organized hierarchy, a writer can narrow it down to certain meanings.

At times, a term is not immediately understood or recognized by a writer. Since the terms have been translated from foreign languages, it is often necessary to refer to a foreign language translation resource. The Team uses EuroDicAutom, a foreign language translation service that is available at the following URL address: http://eurodic.ip.lu/cgi-bin/edicbin/EuroDicWWW.pl The Team also uses foreign language dictionaries available in print in the library's reference section. As a back-up source, we use AltaVista's Babel Fish at http://babelfish.altavista.digital.com/cgi-bin/translate?

Additional translations in English provide the writer with clues. Often it becomes clear that there is an American equivalent term or the term that would be used in the United States. Any proposed American equivalent term is suggested in the "Notes" section of the Word document.

With clues derived from the hierarchy and the foreign translations, a writer can then research for the term's meaning. A term such as "waste disposal" can easily be found within one of the Team's print sources. But usually finding a definition presents a challenge. If a definition cannot be found after consulting print glossary or dictionary sources (or from CDROMs for RandomHouse and Oxford English Dictionary), then a writer would consult the World Wide Web. The Team prefers to use the AltaVista search engine for its searching (http://www.altavista.com/).

After a writer has found a definition for a term, he or she can start the process of re-writing this particular definition so that it fits within the GEMET hierarchy. In writing a definition, there are several considerations.

If the definition comes from a non-government source, there is a need to re-phrase the definition to avoid copyright infringement problems.

Often a word within a compound term has been defined previously by the Team. So to avoid inconsistency, a writer often needs to incorporate language from previously written definitions. The writer also needs to be consistent by conforming to the Team's writing style.

A final step often includes entering citation information for new sources in a source document format that CNR has provided. Consistency in citation writing is another practice that is emphasized.

3.1c GEMET Challenges

CNR has emphasized the importance of finding definitions for the terms that have been provided. It is believed that they seek definitions from authoritative sources so they feel confident that the correct definitions have been provided. To ensure that high quality definitions are provided, the Team spends an extraordinary amount of time researching for definitions.

The Team has also learned to write detailed and comprehensive definitions that are written succinctly. These definitions attempt to incorporate all of a term's possible meanings or applications within one sentence. They also should have some meaning from an environmental or ecological standpoint.

It may seem that certain GEMET terms could easily be defined by referring to general dictionaries. For instance, a term such as "indigenous knowledge," could be defined by utilizing definitions from RandomHouse for "indigenous" and "knowledge" and then linking them together in a single definition. Whenever possible the Team has learned to avoid "building" definitions from the constituent parts using general dictionary sources. It is preferable to find authoritative definitions by extensively researching AltaVista or other web-based sources.

One typical challenge is to write definitions for terms that we are familiar with, yet definitions cannot be found for these terms in print or web-based sources. Many GEMET terms are compounds or include more than two words. A typical example is "environmental education equipment." Although one might know instinctively what this term probably means, finding a definition for it by searching AltaVista is very difficult. A solution might be to search for a definition for "environmental education" in AltaVista and then combine it with a general dictionary definition for "equipment."

At times, the Team is not able to find a dictionary or glossary definition for a term. But we are able to determine how a term might be used in documents found on World Wide Web. A typical example is "mountain farming." The Team was not able to find a definition for this term. But we were able to determine how it was used in several web documents and we cited one of these documents as a source.

A number of terms clearly have a focus on matters related only to Europe. The Terminology Team frequently wrote a note explaining that these terms would be better left to someone with more expertise and experience. Some examples of such terms include "Community Act," "EC regulation on eco-management and audit," "transposition of directive," "European Environmental Council," "European Recovery Programme," "social tourism" and "internal European market."

Some terms simply did not make sense from an American perspective. It is believed that many of these terms did not translate well in American English. For these terms, the Team extensively researched for the meaning of these terms within EuroDicAutom and other foreign language resources. We also researched the World Wide Web for possible American equivalents. But even after all of these research attempts, there were times when we could not determine how a term could be re-written. A couple recent examples would be "depositing agent" and "hermetic collection." In these circumstances, the Team often submits a note asking for further clarification.

Frequently, the Terminology Team is able to track down an American equivalent term. But this process consumes an extraordinary amount of time. It should be understood that

an evaluation of the GEMET terms is usually required and this process will slow down the definition writing process. An evaluation of GEMET terms is required mostly to ensure that they are appropriate for American users of GEMET.

3.1d Editing Review

The Team has an editing review process in which Team members exchange each other's definitions in the middle of the week for a mid-week review. Feedback or help is provided after each review.

At the end of each week, the definitions are submitted to an editor. On Monday of the following week, the editor combines all of the definitions in one file. This editor reviews definitions along with each team member's personal notes. The primary concern of the editor is to ensure that each term has the correct definition. The editor is also concerned with ensuring that there is consistency in all of the provided definitions.

By the middle of the week, after the definitions have been edited, the editor downloads the definitions and sources files into zipped files. The zipped files are then FTPed to a server by Internet Team members who have access to this server. When this process is completed, a form e-mail letter with an attachment is submitted to Sandra Lucke of CNR in Rome.

At times, CNR has provided the Team with their editing review feedback. They have also written to provide the Team with directions on how to use CNR's files**. [For further information, please refer to CNR/GCI e-mail correspondences in the attachment for Appendix III. Please note that the listing is meant to be representative and is NOT a full set of e-mail correspondences.]**

3.1e Future Workflow

The Team continues to work on Annex terms provided by CNR. Under the Team's current arrangement, it is believed that much of the work on the Annex terms will be completed early in the summer of 2000. But how much work that will be completed depends on other Team priorities. It will need to be determined how much time should be devoted to writing and revising GEMET definitions.

The Team believes it would be worthwhile to continue its cooperative efforts with CNR and its partners in Europe. The U.S. EPA could continue to have an important role in the development of GEMET into a global thesaurus. With the Terminology Team's support in future GEMET related projects and after some years of establishing a global thesaurus, it is believed that the terminology from this thesaurus could be accepted and standardized around the world. This kind of progress would help ease the exchange of environmental information across national boundaries and such progress will be needed as the world's economies become increasingly interdependent.

3.2 Results

Since the Team was first established, it has produced approximately **2,240** definitions for GEMET terms. **1,468** definitions were submitted to CNR prior to the release of GEMET 2.0 in August 1999. Of these definitions, **1,121** were accepted by CNR for GEMET 2.0. The Team continued to submit definitions after the work for GEMET 2.0 was completed. In this time period, the Team submitted more than **790** definitions. These definitions have been set aside for future review. Some of the definitions have already been evaluated and are being saved for the global thesaurus initiative. **[The Team's GEMET definition sets are available in the attachment for Appendix IV.]**

Much of the Team's work has involved evaluating and mapping the terms so they can be understood from an American perspective. For each set of definitions, the Team has submitted "Notes" that provide suggestions for American equivalent terms and hierarchy changes. A total of **516** "Notes" have been submitted. Currently, CNR is working to incorporate these "Notes" in their working master file. [**The "Notes" that the Team submitted are available in the attachment for Appendix V.]**

The Team also contributed to the GEMET 2.0 sources document. Using a format provided by CNR, the Team has continued to provide a listing of sources each time a set of definitions is delivered. This document now includes over **310** source citations and includes information about each source's author, title, publisher, publication location, publication year, ISBN or document number and URL address. The source citations are arranged in alphabetical order by their three lettered abbreviation codes (the abbreviation codes are included after each definition in the definition sets). Periodically, the Team reviews the document to ensure that the information is consistent and accurate. **[The Team's most recent Sources document is available in the attachment for Appendix VI.]**

3.3 Applications

3.3a GEMET on EPA's Intranet site

An older version of GEMET is available on EPA's Intranet site at the following URL address: http://intranet.epa.gov/epahqirc/thesaurus/ This page was last updated on February 18, 1998. The Terminology Team believes that this page should be updated again so that the latest version of GEMET is available through this page. It should also be noted that the link for "Comments" refers to Stu Gagnon's e-mail address and he is no longer working as a contractor for U.S. EPA. It might be useful to consider making this Intranet page as part of a larger web site that covers all of U.S. EPA's terminology.

3.3b GEMET on CD

CNR has been distributing CDs in which GEMET is available through PDF files and ThesShow. The CD includes a guided tour which discusses the background for the European Topic Centre on Catalogue of Data Sources (ETC/CDS), the need for

"metainformation," the CDS Cataloging Concept and the GEMET approach (with links about ThesShow and access to GEMET PDFS). The tour is also available through the ETC/CDS web site at the following URL address:
http://www.mu.niedersachsen.de/cds/Guided-Tour.htm

The GEMET 2.0 PDFs page available through this CD includes links to seven PDF files. The link for the "Systematic List of Descriptors" includes GEMET's organized hierarchy. The link for the "Thematic List of Descriptors" includes a list of terms under each of the forty GEMET topic categories. The link for the "Alphabetical List of Descriptors" includes a list of all GEMET terms in an alphabetical arrangement with definitions and related terms under each term. The "Concordance List" includes a list of singular terms in an alphabetical arrangement with all variations of a term listed under each term ("action" has a list that includes "action group," "class action suits law," "public action," and "urban action program"). The link for the "Mutilingual List of Descriptors" includes a list of terms in an alphabetical arrangement with foreign translations listed under each term.

The CD allows users to install ThesShow. But this software would only be available in a 30-day trial version. For installation, a user would need the number code of 12345678. Those who want to use the full version are advised to contact ETC/CDS by e-mail to receive a password.

3.3c ThesShow

ThesShow is a software that is suitable for viewing only. This Visual Basic based browser supports navigation through the GEMET database. A user can navigate through the systematic view that displays the hierarchical levels organized as a tree from the first to the last entry. Or a user can scroll down an alphabetical or thematic list of GEMET terms. Another option is to enter a term in a detail window in the top right corner of the window screen. By clicking on "Find Next," a user would retrieve a display of the term's related terms, groups, definitions and foreign translations in a window screen below where the term is entered. **[For further information about ThesShow, please refer to the slide show at** http://www.mu.niedersachsen.de/cds/etc-cds_neu/thes_slide_home.html**]**

3.3d GEMET within TRS

TRS was originally built to house GEMET. In TRS 1.1, GEMET's hierarchy is shown through its dot system. One dot refers to the highest hiearchial level, while two dots represent the second level down and three dots represent the third level down. This dot system ranges from one dot to five dots. TRS also includes GEMET terms and definitions under each of GEMET's topical categories. In TRS 1.1 and in future versions of TRS, GEMET is likely to become a widely used resource for the environmental community.

3.3e Future applications or possibilities

To make GEMET a useful tool for American users, the Terminology Team believes that further work needs to be undertaken to improve GEMET. The GEMET product as it is now includes many terms that would not be recognized by American users. It is suggested that along with definition writing, the Terminology Team should be involved in a close working relationship with CNR that would involve term mapping or evaluating terminology and definitions so that the product is more user friendly to Americans.

4.0 User's Guide to TRS

TRS is an extremely user-friendly system designed for easy access to terms, definitions, term resources and an international environmental thesaurus (GEMET). Users will find TRS functions as an on-line dictionary possessing an easy access term look-up, that it contains term phrases similar to an unabridged dictionary, that it includes source information as does a bibliography, that it shows semantic relationships as does a thesaurus, and that it provides one or more definitions for a single term that fulfills its primary feature, as a collection of specialized vocabulary on the environment. All totaled, TRS exemplifies a well-rounded information source that incorporates many functions of standard reference sources (dictionaries, directories, bibliographies). Its overall vision to supply environmental terms and definitions should appeal to and satisfy a variety of users.

For environmental specialists, TRS organizes environmental terms derived from official documents and other expert source literature. Inherent in the collocation, or "bringing together" function of TRS, is the capacity to analyze similarity and variation of the terms and definitions produced by various entities.

TRS has three options for conducting searches: 1) basic searching 2) advanced keyword searching and 3) search by resource. A basic search includes the simple, "click on letter" search (**See Fig A**). The user simply clicks on the letter from the displayed alphabet that begins the desired term. This may be an attractive feature for foreign users unsure of the spelling of terms, or more broadly, for any situation where there is uncertainty of how the term is represented in the display.

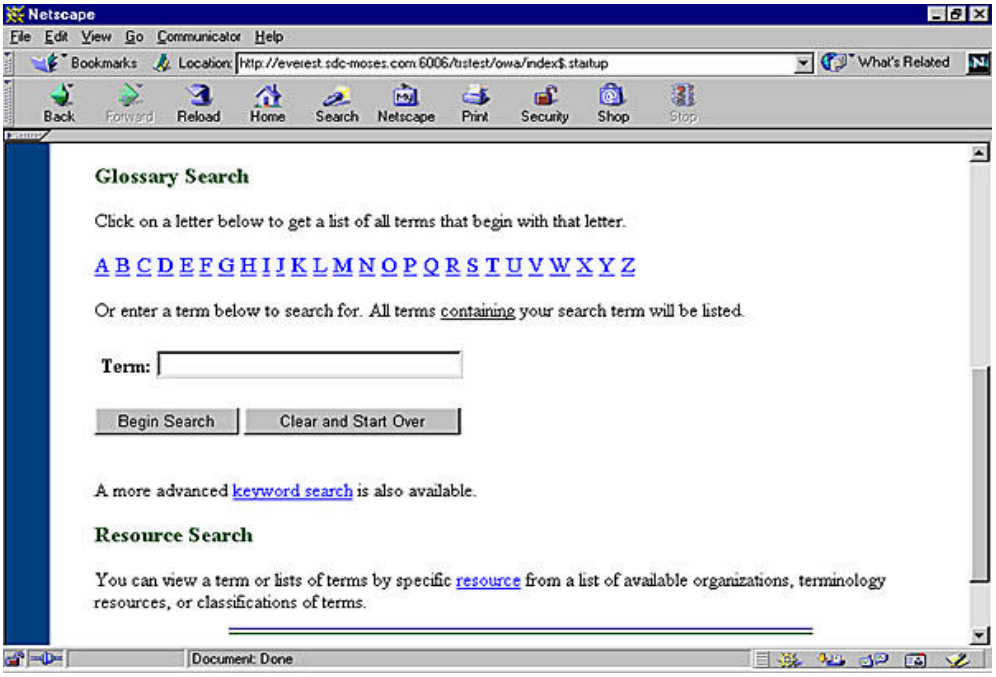

Fig. A: Click on Letter Search

The second basic search is the search term box **(See Fig B**). Users enter search terms into the search box and enter for the result display.

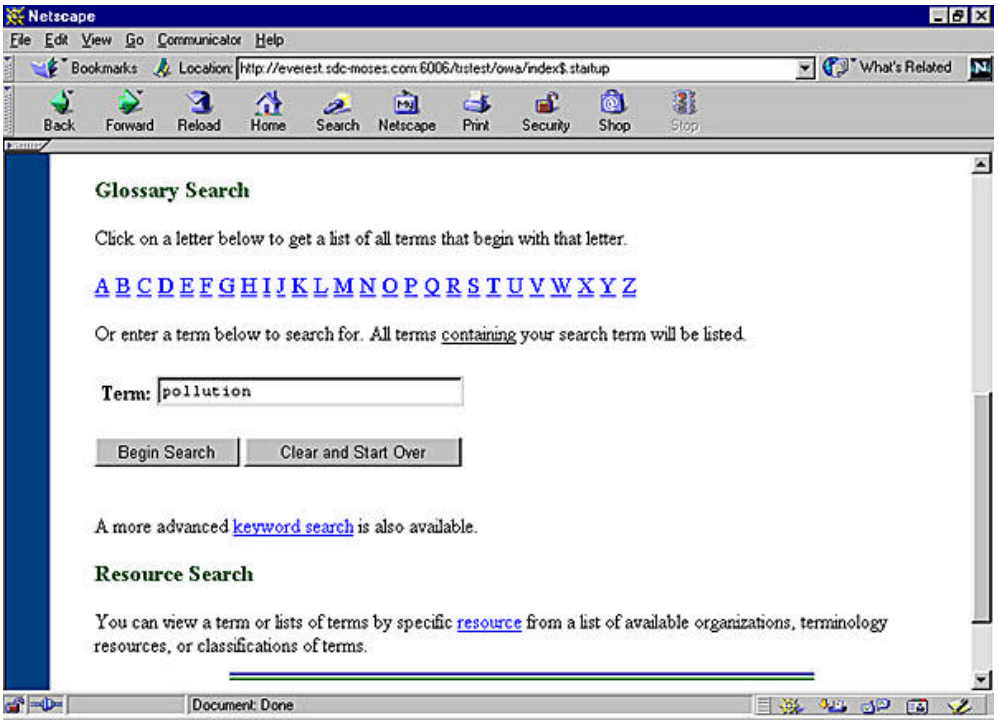

Fig. B: Search term entry

Results will include all phrases containing the term as well as definitions for the entered search term (**See Fig. C**).

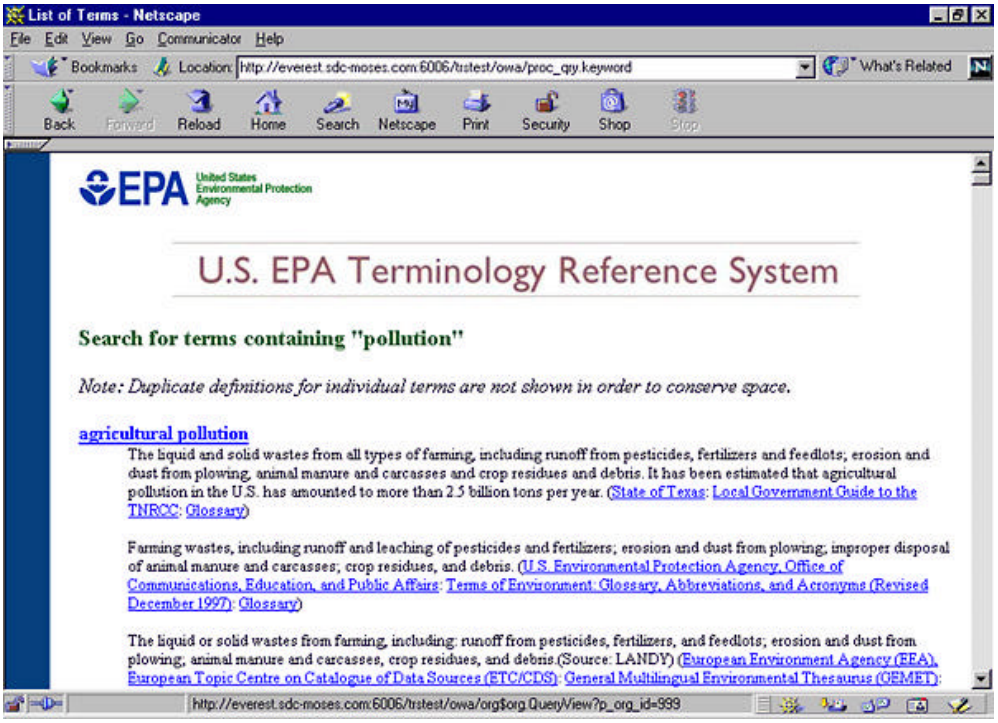

Fig. C: Results of Search by Term Entry Search

Advanced keyword searching features three search options: the search term is embedded in a phrase ("containing") the search term begins a phrase ("beginning with") and the search term comprises the entire phrase in full ("exact match") (See Fig D).

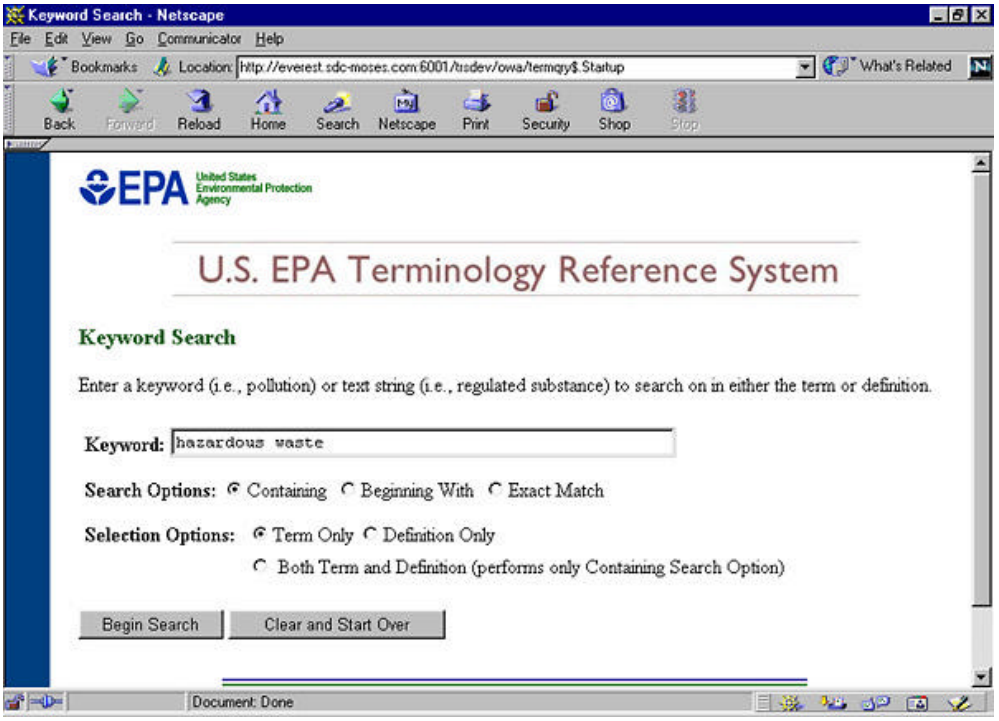

Fig. D: Search Options

These options increase the flexibility and promote precision in search results (See Fig E).

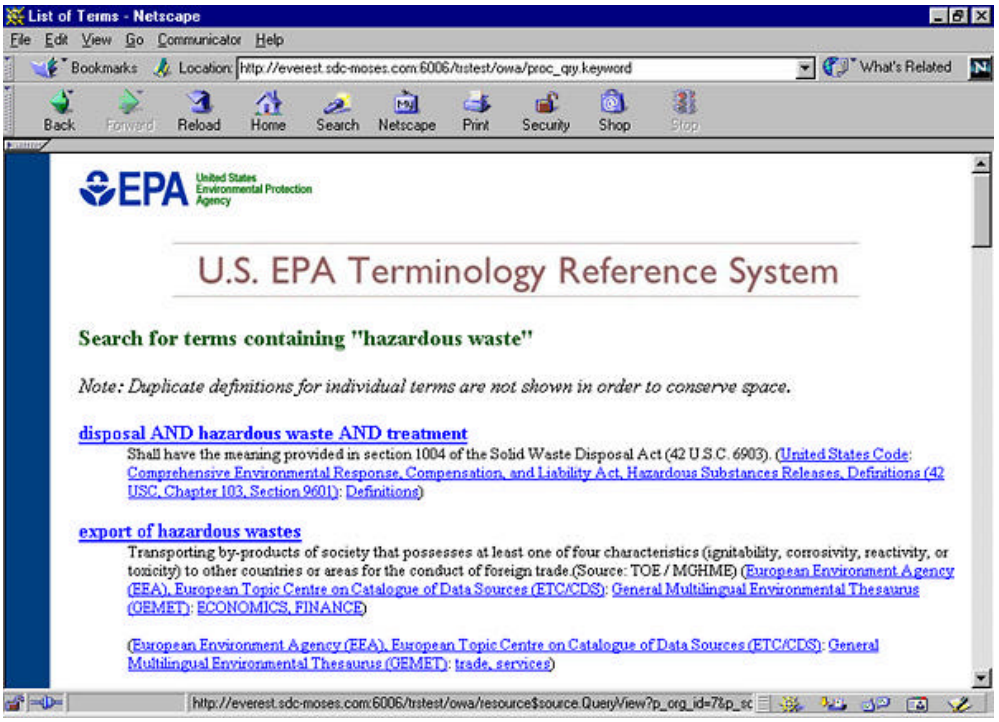

Fig. E: Search results for "containing" option

The search by resource option fulfills the need to view the entire set of terms contained within a single resource. To do so, the user selects the "resource search" option and views a list of resources used in TRS (See Fig. F). The user selects and clicks on a resource and selects the preferred search method to either access a specific term or view the complete list of terms for that resource. This option might be attractive for researchers having a source-based, rather than term-based, focus.

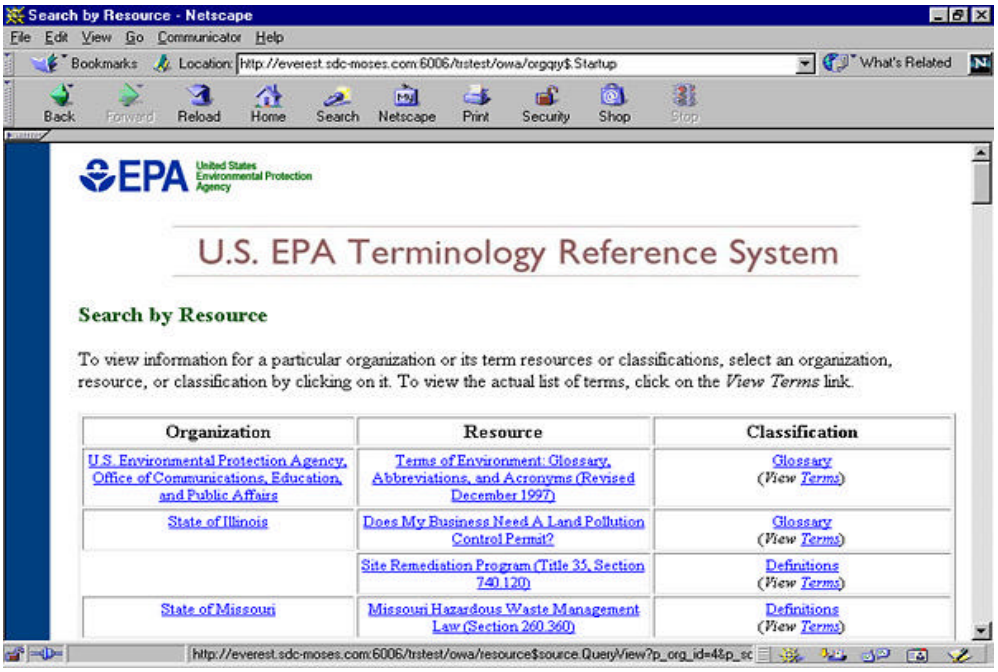

Fig. F: Search by Resource

There are three options for displaying search results. Users can choose to view terms without definitions, definitions only, or both the term and definition. Choosing the desire display option simply requires clicking on the desired choice from the advanced keyword search screen.

To illustrate the potential use of TRS in detail, the following are hypothetical cases for searching and retrieving TRS terms and definitions:

Case #1. A user wants to know all resources published by the State of New Jersey included in TRS, and those terms contained in these sources.

Action: The user selects the "resource search" option. Three columns are displayed: organization, resource and classification. It should be noted that the organizing element is an alphabetical listing of the organizations (i.e., responsible entity) rather than the resource title. The user scans the organization column for New Jersey ("State of New Jersey") and views the list of resources appearing in the second column. The user clicks on each resource link and then selects a search option to access the terms.

Case #2: A user enters in a search term and yields zero hits. He is uncertain of why the term does not yield results and needs to verify whether the desired term exists in the system.

Action: User selects the "alphabetical search" by clicking on the first letter of the desire term from the letter list. User browses the area where the term should appear if represented in the system and verifies whether there is representation.

Case #3: User wants to compare the State of New Jersey definition(s) for solid waste with the definitions produced for the State of Texas. He also needs to know the phrases each state produced which include "solid waste."

Action: User selects the keyword search option and enters "solid waste" in the search box. In the results set, he sees whether source information following the definition includes the State of New Jersey or the State of Texas. Optionally, the user can select the "resource search" option, click on sources for both New Jersey and Texas and following, conduct a search.

International Terminology (GEMET)

International environmental terms in the form of the GEMET Thesaurus (refer to previous sections for history), was imported into TRS in both hierarchical and non-hierarchical formats. The latter format simply resembles a dictionary format with the addition of the thesaurus classes where the term is positioned. The former, the hierarchy, was retained for TRS to enable a visual display of terms and relationships (broader terms, narrower terms, related terms and siblings) (See Fig. G & H). Search features which are available only for the international terms are 1) searching the term group or class 2) searching terms having a relationship to the search term 3) displaying a selected portion of the hierarchy containing the search term.

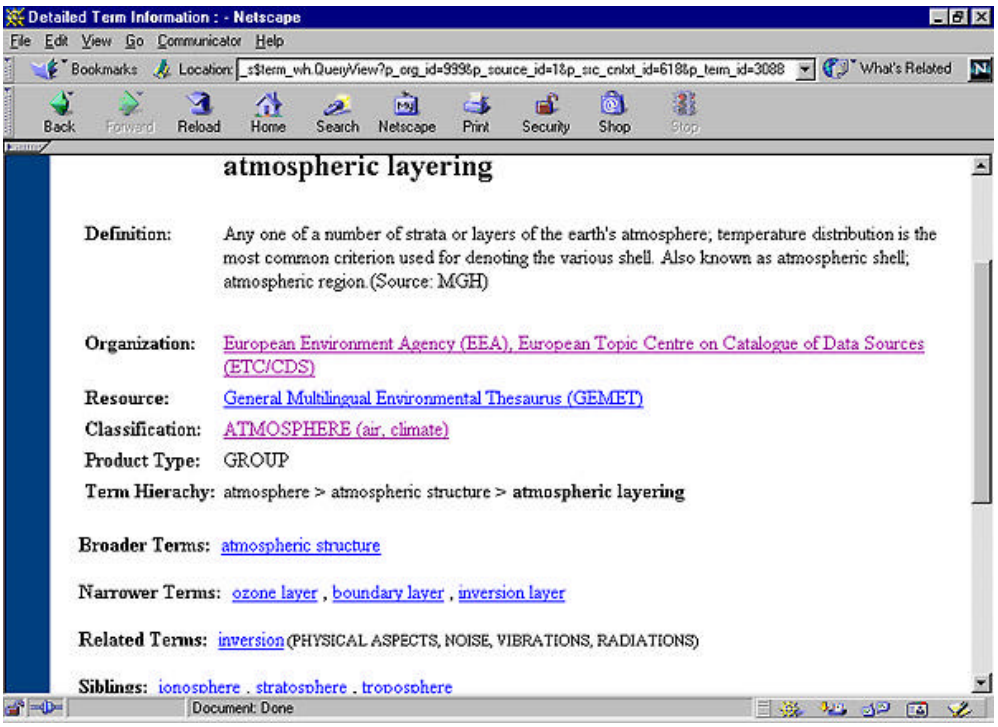

Fig. G: Search Results for GEMET Term

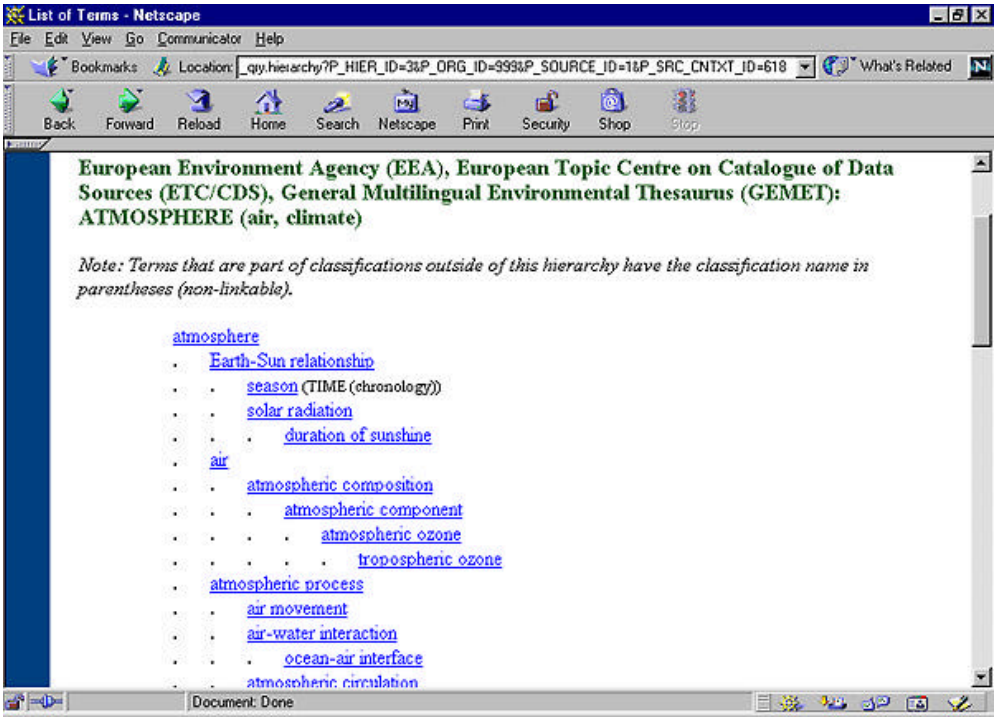

Fig. H: Hierarchy Display for GEMET Term

4.1 Writing Regulations Using TRS

One of the inherent functions of TRS is the collocation of environmental terms, term phrases and term definitions. As a database of collected and organized environmental terminology, TRS can facilitate the process of writing regulations and therefore save the writer valuable time and effort. In researching and creating terminology contained or cited in regulations, the regulation writer can utilize TRS. Any TRS term can function as a model for writing terms or definitions or in general act as an aid for regulation composition.

The writer can exploit the inclusion of the term source in TRS especially in cases where a specific organization or state's term is desirable. For example, the writer can, through a resource search, determine whether the state he is interested in (or writing for) previously created a definition or definitions for the desired term. Or the writer may need to expand his scope to other states or organizations. TRS allows term comparisons within a single state, among states and the EPA, and between the United States and certain European countries (member countries participating in the GEMET Thesaurus).

4.2 Information locator tool

One perceived disadvantage of natural language searching is the intellectual effort placed on the searcher to think of terms, that is search terms, that will extract desired documents from the database. The task and advantage of a thesaurus is that it eases this intellectual exercise of deciding what, and all, search terms to employ in document retrieval. A terminology collection, while less structured than a thesaurus, is nonetheless a potential supplier of search terms and therefore a means to economize the search effort. The organization of environmental terms in TRS can aid in document retrieval. Although there is no controlled vocabulary, no imposed authority or preferred term, TRS terms and phrases, because they are drawn from expert literature, can be employed as search terms for document retrieval.

One potential TRS has with respect to querying a database is the inclusion of phrases in the results set of a term search. This allows a searcher to refine their term or phrase, that is, to narrow the concept to an exact phrase. This will probably be effective with strongly represented terms, such as "waste" or "treatment." This honing process will likely generate precise results.

With the inclusion of the GEMET thesaurus in TRS, a broad range of European environmental terminology is made available to the searcher. Researchers looking for European produced literature written in English (or translated into English) will no doubt welcome the European counterpart terms.

For those with general interest in the environment, students who are writing reports, concerned citizens who want to increase their general knowledge of the environment, becoming familiar with terminology is good place to launch the learning process. This includes terms, concepts, semantic relationships, phraseology, definitions and context.

Terms identify and reflect issues, actions, characteristics, qualities and numerous other aspects existing within the universe of environmental science. To a good extent, simply reading definitions and understanding the semantic relationships among terms engenders at least an impressionistic understanding of the environment, and may be a gateway to in-depth learning.

As a vast term collection or repository of environmental terms, TRS can aid in individual or collective efforts to standardize environmental terminology. Term collocation is the necessary and preliminary activity of harmonizing terminology and this in turn, the preliminary effort of standardizing terminology. F.W. Lancaster, a renowned authority in the field of Library and Information Science states that:

"… the standardization of terminology, including some degree of international harmonization, is an area of increasing interest and activity. This type of standardization has traditionally been promoted through the use of technical dictionaries and glossaries. Multilingual versions of these tools assist translators and promote consistency in translation. Such tools, however, are expensive to produce and tend to become obsolete very quickly. An alternative, and one of growing importance, is the terminology data bank which De Besse… refers to as "A kind of living multilingual electronic dictionary containing hundreds of thousand of technical and scientific terms together with the appropriate terminological information."

Standardized terminology facilitates the communication process and therefore aids in cooperative understanding. Terms contributed from responsible organizations, whether local, regional, national or international, can make a contribution to the process of standardizing terminology by acting as a source for controlled vocabulary or preferred terms. Helen M. Hutcheson supports the standardizing effort in her essay, "Preparation of Multilingual Vocabularies":

> "It is the improvement of communications and the reduction of misunderstandings that are the ultimate goals of authors of multilingual vocabularies. All the terminology bodies consulted agree that multilingual vocabularies have a standardizing influence on the use of terminology. By recommending certain terms and equivalents and by promoting correct terminology and language usage, multilingual terminology organizations implicitly contribute to terminology standardization. Therefore multilingual vocabularies must be of the highest quality possible."

5.0 Environmental Terminologies for TRS September 2000

5.1 Preparation of New EPA and State Resources

TRS 2.0 Data Collection

For collecting state terminology, both the designated priority sources and proposed

collection method for TRS 2.0 will differ from versions 1.0 and 1.1. The prior method of mining terms from the states' web sites will be initially replaced by a more structured approach of identification and extraction of state definitions from the EPA's 13 National Systems targeted in EPA's *Reinventing Environmental Information (REI) Action Plan*. Seven of the 13 National Systems will be the focus of initial exploration. (See the attachment for **Appendix VII**). These seven systems contain pollutant data reported by the 50 states. Research may involve identifying how data are defined by the states and whether, or to what degree, the state definitions conform to their representation in the EPA Systems. This collection method recognizes that state data definitions in the EPA National Systems have strong potential for building TRS and in turn, that TRS can provide an additional avenue for access to this data.

In addition to extraction of state data definitions from the 13 National Systems, an additional and concurrent effort is to mine EPA glossaries. Glossaries created from EPA Offices responsible for the data in the 13 National Systems have priority. The Team has collected approximately 180 EPA glossaries to date and will begin evaluation in May 2000.

Data Mining Tools

In the coming months, the Team may research the features of a variety of data mining and display tools, such as XML topic maps, term-extraction software, graphical data analysis software, or document content characterization software. The primary effort would be to inform the client on the potential utility of these tools to facilitate data mining or graphical data display of TRS content.

The Team was able to view a demonstration of the SPIRE (Spatial Paradigm for Information Retrieval and Exploration) Software, visual text analysis software developed by Pacific Northwest National Laboratory. SPIRE mediates visual representation of information, or more specifically, the use of spatial proximity to represent conceptual similarity. From the software demonstration, it appeared that SPIRE would be especially valuable for any large-scale research project requiring massive organization and content interpretation of documents. The ability to not only display document content, but show conceptual similarities within a set of documents may reduce the enormous task of document organization and interpretation that is a preliminary effort to term standardization. In particular the cited need for standardizing terminology identified in the *Draft Summary of the Data Standards Priorities Workgroup Report* (April 5, 2000) is one basis that would justify further exploration of the potential role of information visualization tools.

The Team may research and evaluate the potential of topic maps (in XML) for TRS content. A topic map is an SGML or XML document that organizes or imposes structure over an information set by describing its topics, topic occurrences (information sources) and associations (relationships between topics). The function of topic maps is to filter information and optimize it for navigation and retrieval. In the same way that a back-of-the-book index directs the user to the desired section in the text, the topic map displays

topics for identification and selection. Like a thesaurus, the topic map shows how the topics are related. Further, one information set may have a multitude of topic maps and these maps can be associated to represent and reveal the complexity of relationships in an information set.

Second Stage Collection Effort

The Team will continue to work systematically to coordinate activities and exercise quality control of the data. Following the successful completion of the priority effort described above, extraction of terms from various terminology collections will continue. This second phase development effort is expected to focus on the following terminology sources:

- State Environmental sites linked to EPA Solid & Hazardous Waste site
- Federal Collection: Code of Federal Regulations
- EPA: Hazardous Waste Superfund Database Thesaurus
- EPA glossaries (in general, not related to the 13 National Systems)
- Environmental glossaries meeting pre-designed selection criteria

Identification/Collection: Terms included in TRS 1.1, excluding the GEMET and TOE, were obtained from Internet sources. For the second phase, the Team will continue to mine web-based sources and concurrently identify print resources that fit our selection criteria. In preparation for TRS expansion, the Team mined the web for environmental glossary collections and documented numerous web sites for later evaluation. To support the data mining effort, the Team will also seek to research and evaluate technology having the potential to beneficially affect all phases of the collection development and management processes, as resources for the effort become available.

Selection of Candidate Terms: The Team will continue to identify sources and review terms based on selection criteria developed for TRS 1.1. The criteria are standard criteria in source collection and can be briefly stated as 1) authority 2) currency 3) consistency 4) scope 5) relevancy. Authority ensures a credible source, an authoritative entity, having jurisdiction or empowered with certain functions, or a recognized expert in the chosen field. Currency would ensure that the source is relevant, not outdated, or with archaic terms, or terms superseded by subsequent edition, version or issuance. Scope requires that the parameters of the document content, the set of terms, do not conflict but conform to the scope of the database. Relevancy requires that the terms are consistent with the vision or theme of the database or consistent with overall data content.

Team Coordination: Repeating the procedure for TRS 1.1, each team member will be assigned glossaries to mine and be responsible for the accurate transfer of data into files. Team members will also self-assign glossary review to ensure they meet the pre-designed selection criteria. A team member acting as editor will review the work for consistency, accuracy and conformity with data content and entry rules.

Data Evaluation: Throughout the data identification, selection and inclusion processes, candidate terms and definitions will need to be evaluated. Evaluation is the key activity to ensuring data quality, integrity and viability; it is the quality of the data that ultimately determines how TRS qualifies as a robust terminology reference system. The data framework, the ease in using the system, the appearance of the interface, etc. will attract and generate interest among users; however, the utility, especially continued utilization of the system, is heavily dependent on the quality of the data.

In the initial evaluation process, the source is evaluated against pre-determined selection criteria as previously explained, and if approved, the source becomes a candidate for inclusion in the database. However the evaluation process does not cease with the identification of candidate sources. Like any viable information system, TRS will require an ongoing evaluation of data -- not only of individual terms, but in totality, as a large container of environmental terminology. Thorough data evaluation can generate questions regarding quality data as well as the overall database quality. For example, for TRS 1.1, the Team entered candidate terms that are not strictly regarded as environmental terms, but rather belonging to another discipline, such as politics or law. Although on an individual term basis they meet the criteria, their collective inclusion in TRS may have an unintended or undesirable impact on the system with respect to search and retrieval. With the inclusion of all states in TRS, heavily referenced terms such as "act," "department" or "commissioner" may overwhelm the search and retrieval process or at least the browse capacity of the TRS. This is but one example meant to illustrate the need for an overall and ongoing evaluation of data, rather than concluding evaluation at the point of induction into the system. For a further evaluation of TRS data, see Section 6.0, "Conclusions and Future Possibilities."

6.0 Conclusions and Future Possibilities

The Terminology Reference System, named by Larry Fitzwater and Linda Spencer, is an appropriate appellation for a robust terminological database. TRS can live up to its name by providing quality data from authoritative sources and satisfying a multitude of reference functions. A systematic effort will ensure that TRS 2.0 fulfills its vision as stated in the January 19, 2000 presentation by Linda Spencer:

Develop a repository of well-defined environmental concepts as a tool for:

- Cataloging data and documents
- Retrieving Web documents
- Developing data elements
- Integrating databases

TRS has excellent potential to expand on its primary asset of collocating or collecting and bringing together terms on the environment. The following evaluates TRS 1.0 with respect to data collection and evaluation and includes suggestions for improving the organization and collection of the data.

Data Collection and Evaluation

Data collection and evaluation are often oversimplified and the time required for completion is often underestimated. This may be attributed to the notion that mass data transfer handled by technology forms the bulk effort of populating a database while the identification and preparation of the data are relatively lesser tasks. It should be emphasized that the quality of the data ensures repeated use.

Selection criteria developed by the Team was used to evaluate candidate sources for TRS 1.0 and 1.1. (See Section 5.2). For TRS version 2.0, it is recommended that the Team narrow the selection focus from candidate sources to candidate terms. This will enable the Team to decide on a term-by-term basis which terms qualify for TRS inclusion, rather than automatic allowance on the basis of the term's inclusion in a selected source.

This recommendation is based on the observation that candidate sources selected for TRS 1.0 and 1.1 contained terms and definitions that did not appear to qualify as "well-defined environmental concepts." Terms in state regulatory documents in particular, were often found to be "document dependent"; that is, once extracted from the context, the function as environmental concepts was attenuated. The general characteristics of these terms are discussed in the following section.

Document-Dependent Terminology

General characteristic: In order for these terms to be fully comprehended, in order to understand the specific meaning or application of the term, the document in which it appears would need to be consulted.

Document dependent definitions may vary in their degree of semantic provision. Looking at the definitions below that come from CFR Title 40, the first term, "device", offers very little to the user without access to the referenced Act. The second definition, "federally enforceable," does give some meaning and may satisfy the user. In both cases, however, the terms may be said to be document dependent.

> Device: Means any device or class of device as defined by the Act and determined by the Administrator to be subject to the provisions of the Act.
>
> Federally enforceable: Means all limitations and conditions which are enforceable by the Administrator, including those requirements developed pursuant to 40 CFR Parts 60 and 61 …"

Data evaluation conducted by our Team would include determining which document dependent terms should be admitted as TRS terms. This analysis may be influenced by any number of variables such as user expectations, project vision, database features, access, etc. Data evaluation, to a large extent, requires consideration of possible scenarios where data do or do not meet user expectations. Consider the user who is looking for a dictionary definition of "device" or "federally enforceable" and instead

encounters the above definitions. However, the user may be an environmental specialist who possesses familiarity with the context and finds the above definitions adequate for his or her needs.

Another type of document specific term is one that is adequately defined, but is clearly specific to the subject matter of the document. Here is one example from a State of New Jersey water document:

> criteria: means ground water quality criteria.

If the user came across the word "criteria" in any other non-water document for New Jersey, it is likely that the word will not have the same meaning. When the above is removed from its document context, the intended definition may be misconstrued.

Non-definition Terms: Another type of term that might be evaluated for TRS inclusion is one that can be said to function as "see reference" term (also called admitted term). Here is one example:

> Board - See Board of Arbitrators.

The Team would decide whether these "see reference" terms should be included in TRS. On one hand, they have an obvious function. If a user does not know the exact phrase and searches "Board" he is directed to the preferred term, "Board of Arbitrators."

The disadvantage of these terms in a term database, however, is that they may generate user confusion. A hypothetical example is that two states have see references from the term, "Board" and the preferred terms are not the same, so that user is directed to two different terms. For example, The State of New Jersey may have a see reference from "Board" to "Board of Arbitrators," while the State of Texas may have a see reference from "Board" to "Board of Directors." They would appear in TRS as such:

> Board - See Board of Arbitrators. (State of New Jersey…)
> Board - See Board of Directors. (State of Texas..)

If the user specifically needs "Board" as it is defined by the state of New Jersey, he would know which preferred term to consult. If no state is preferred, he might be perplexed over the choices. The co-existence of two "see references" in TRS is mutually inhibiting. This is usually not a dilemma in an index, such as a back-of- book index, because any admitted term ("Board") will correspond to one preferred term ("Board of Directors"). In a term repository or collection of terms that is the basis of TRS, multiple instances and representations of one term are expected.

The above explanation and examples attempt to serve a dual purpose of highlighting the need for data evaluation to meet data quality standards and to outline possible actions based on data evaluation conclusions. It should be noted that the Team has not

implemented any of the above to date and is detailing them in this Report to the EPA for their consideration.

The following sections offer suggestions for the organization of TRS data into supplementary resources that may potentially enhance TRS utility.

Acronyms

TRS 1.1 contains acronyms with and without definitions. The four representations below show how acronyms and their full form have appeared in state regulatory documents:

1) Both acronym and phrase are represented
2) Acronym acts as a see reference
3) Acronym is defined (no phrase)
4) Phrase is defined, does not include acronym

Examples of the above:

1) PCB: Means Polychlorinated biphenyls

2) PCB: See Polychlorinated biphenyls

3) PCB: means a mixture of compounds composed of the biphenyl molecule which as been chlorinated to varying degrees.

4) Polychlorinated Biphenyls: means a mixture of compounds composed of the biphenyl molecule which as been chlorinated to varying degrees.

If a search is conducted for "PCB" (TRS 1.1), the user will not retrieve #4. If a search is conducted on "Polychlorinated Biphenyls," the user will not retrieve #1 or #2 (though #2 would direct him to the phrase term). Thus, this does not fulfill a central function of TRS, to bring like terms together. Consider the user who searches "PCB," garners results and stops. He is not motivated to conduct a search on "Polychlorinated Biphenyls" and remains unaware that there are other definitions.

In a glossary, Example nos. 2 and 4 would be linked by a see reference to control the representation of the concept. This cannot be applied to TRS because of the principle of verbatim representation. In keeping with this principle, none of the above examples can be modified.

The solution requires alerting the user to the existence of both acronym and its full form in the database. One possible solution is to create a separate page containing all acronyms, an "acronym list." This page should include both acronym and the full form and allow the user to easily access each in the term list.

This acronym list might be represented in the following manner:

Acronym List

PCB
Polychlorinated biphenyls

If a term is underscored, it is linked to a definition in the TRS database. If no definition existed for Polychlorinated biphenyls, then the term would not be underscored. The link for "PCB" is necessary for Example no. 3 in the above list.

The last category of terms that could possibly be excluded from TRS is administrative terms, or perhaps, heavily referenced administrative terms. The basis for this exclusion is the observed high quantity, low quality characteristic of these terms. *The Environmental Dictionary* (1995) provides terms and definitions from Title 40 CFR and updates to the Federal Register for Title 40 spanning 1987-1994. The following are three examples of administrative terms found in *The Environmental Dictionary* along with the number of the definitions found for each term.

Act: 102
Administrator: 72
Person: 65

A large number of "definitions" for "Act" found in *The Environmental Dictionary* are actually denotations, such as:

Act means the Federal Water Pollution Control Act, 33 USC, 1151, et seq.
Act means the Clean Water Act (33 U.S.C. 1251 et seq.).

or

Administrator means the Administrator of the Environmental Protection Agency (EPA)
Administration means the EPA Administrator or an authorized representative.

A user who wants a definition for "Act" such as the following definition from *Black's Law Dictionary* is not likely to find the above of much use:

Act, legislative act: An alternative name for statutory law. A bill which has been enacted by legislature into law.

TRS as a Gateway

Some terms contained in TRS 1.1 were culled from state regulations. Although access to these terms in TRS 1.1 is a relatively simple process, once TRS becomes more populated, requiring more terms and sources to browse, finding the state terminology from regulations will probably take more time.

If it is determined that there is much greater need and use of the terms from state regulations, as opposed to for example, non-official or non-regulatory documents, a possible accommodation would be the installation of a legal resource page. This page might list states and the state code sections that are sources in TRS. Both the states and sources could be linked for quick access to their terms. This page might also act as a gateway to related law material via exit links. Possibilities include exit links to the Federal Register (**http://www.epa.gov/fedrgstr/**) the GPO site (**http://www.gpo.gov/**) or to the state regulatory locator (**http://www.nmfrc.org/srt/**).

Localizing the GEMET Thesaurus

During the process of writing definitions for GEMET, it was occasionally found that the supplied term, the GEMET preferred term, was not used in the United States. In such cases, the Team would attempt to determine the counterpart term (or terms); that is, the American equivalent. If such a term was found, the Team wrote an explanatory note and these notes were included in the definition sets. The following is an example of a note created for the GEMET preferred term, "waste category."

> "Waste category must not be confused with contaminant category
> which is defined not as a general classifying name for contaminants,
> but instead, refers to a system with which the EPA rates carcinogenicity.
> For American English, "waste category" might be presented as waste type."

This process of finding the American English term may be likened to the mathematical function of "mapping," that is associating one term to another term or terms based on similar characteristics. It must be emphasized that mapping the GEMET term arose from the understanding that the term was not recognized in American English usage. Therefore, if the GEMET term was not a predominantly used term, but had reasonable appearance in documents, term mapping was generally not done. The time allotted for the weekly definitions set did not allow term mapping to be routinely undertaken.

If term mapping was a routine part of the definition writing process, an independent product resembling an American English/British dictionary of environmental terms may have emerged. However, this dictionary would have had a single definition representing both terms. In identifying the utility of such a product, one might consider a future scenario in which European produced research or jointly written U.S.-European documents on the environment need to be indexed for retrieval.

A summarization of term mapping follows in consideration that this may be of interest to the EPA as a future effort in developing TRS.

Initially, it must be determined whether the term is recognized in the home environment. If the term is found in documents, those documents should be researched to identify alternative terms. If the term was not found in EPA documents, the term equivalents would have to be identified by using the supplied definition; this would require the sometimes very arduous task of determining the conceptual equivalent.

The initial question of whether an environmental term or phrase is used in the United States may be a straightforward effort handled by EPA web site searches. For example a phrase search conducted on the EPA web site for "waste classification" generated 128 hits. As a routine, a number of the documents are viewed, but a results set of this size usually will confirm that the term is recognized here. The question of what other alternative terms are used is far more difficult to determine and may require an extensive effort. When "waste classification" was researched, the phrase "waste type" was identified as an alternative term since it represented the same concept in consulted sources. "Waste type" was then entered as a phrase search on the EPA web site and generated four times as many hits as "waste classification." This would support the conclusion that "waste type" is the more popular term, as least within the EPA.

The reverse process might also be included in term mapping. For example, the terms "waste stabilization" and "waste sludge" generate 203 and 252 hits, respectively, on the EPA web site. However, neither of these two terms appears in the GEMET 2.0 hierarchy as preferred or admitted terms (although the concepts may very well be represented). The additional effort might be undertaken of determining whether "waste stabilization" or "waste sludge" can, in turn, be mapped to a GEMET preferred term.

The above was meant to introduce the effort of term mapping. The process is somewhat simplified here but the purpose was to convey the key actions involved. No discussion was devoted to the inevitable situation where terms do not have a 1:1 correspondence or concerning the degree of equivalence between terms. If there is interest on the part of the EPA Management, a pilot effort might be undertaken and the conclusions of this effort submitted as a white paper.

The following is an evaluation of the inclusion of GEMET in TRS 1.1. It was written to be informative with respect to the functionality of TRS.

The Inclusion of GEMET

As of January 2000, 16,556 terms from the GEMET Thesaurus were transferred into TRS 1.1. (For more information on GEMET history, See sections 2.0 and 2.1). This formidable terminology set will allow users in the United States to access environmental terms as they were selected and organized by European environmental specialists. TRS Users can access GEMET terms and definitions as well as term instances in the hierarchy via a "hierarchical display" option.

The GEMET thesaurus in TRS 1.1 ("GEMET-TRS") only partially resembles the most recent release of the European version ("GEMET 2.0"). GEMET-TRS is not multilingual since only the terms/definitions in English were transported and the GEMET thesaurus is one source in a container of what will eventually be a large number of sources. Given these differences, it may be concluded that the status of GEMET in TRS 1.1 has been reduced from the original.

In light of this conclusion, it may be of interest to examine whether the functionality of the GEMET thesaurus in TRS 1.1 was at all compromised. This summary view on the functions of multilingual thesauri, by Michele Hudon, a consultant in the area of bilingual thesaurus construction, can be used as a basis for examination:

> "Multilingual thesauri serve mainly as indexing and retrieval aids in multilingual information systems. When a multilingual thesaurus is available, documents can be indexed in one or more of several languages. Searches can be conducted in a different language… The thesaurus then plays the role of switching language, and facilitates interlinguistic communication."

Ms. Hudon's statement can be applied to thesauri in general. A thesaurus facilitates document organization and control by identifying preferred terms to assign to documents, or more simply, indexing the documents. Thus typically, the indexer or cataloger selects one or more terms from the thesaurus perceived to most adequately represent those concepts reflected in the documents. The transfer of the GEMET thesaurus into TRS required it to be recast to fit the TRS term repository framework. In this new schema, its function as a thesaurus is somewhat obscured. The user is not alerted to the existence of a thesaurus, nor of its characteristics. Considering that there is usually some basis for selecting one source from a listing of sources, unless the user is knowledgeable of the GEMET and knows that it is available within TRS, they will have to "discover" the GEMET thesaurus via the search process. Thus it can be said that the TRS framework does not "promote" the GEMET thesaurus. Given this lack of promotion, a perception of GEMET as a tool for cataloging or indexing documents is probably weakened. It is not likely that an "embedded thesaurus" will generate the same degree of confidence of stand-alone thesauri. Rather, users will likely perceive the role of GEMET, if at all, to be the provision of environmental terms/concepts and quality definitions.

Appendices

I. --"Activities on EEA GEMET Thesaurus and Development of the GEMET-based U.S. EPA Thesaurus: Final Report on the CNR Activities. October 9, 1999-Subvention period 1998-99" (**format**: Word; **size**: 4 pages)
-- "Activities on EEA GEMET Thesaurus and Development of the American Version of GEMET—Subvention Periods: April 30th, 1998 (with Funds), April 30th, 1999 2000 (Extension without Additional Funds)" (**format**: Word; **size**: 4 pages)

II. EPA Press Release, "Common Global Environmental Vocabulary Being Developed" (2/03/00) (**format**: HTML; **size**: 1 page--4 KB)

III. CNR/GCI Correspondences Table (**format**: Microsoft Word; **size**: 195 KB, 43 pages)

IV. GEMET Definition Sets (Defwork00 to Defwork56) (**format**: Microsoft Word; **size**: 187 pages--613 KB)

V. GEMET Notes (Defwork00 to Defwork56) (**format**: Microsoft Word; **size**: 54 pages--202 KB)

VI. GEMET Sources Document (Defwork56) (**format**: Microsoft Word; **size**: 25 pages--200 KB)

VII. Table 1: EPA's National Systems in EPA's "Reinventing Environmental Information (REI) Action Plan" (**format**: GIF; **size**: 1 page)