



Pulling Environmental Information Together across the Internet:

European/United States Environment Agencies Cooperation in the Development of a Common Terminology System

By: Domingo Jiménez-Beltrán and Alvin M. Pesachowitz

1. Introduction

Environmental protection requires broad understandings that cross subject disciplines and national boundaries. Potentially useful information is increasingly available on the Internet. However, the quality of that information varies widely and the quantity increases at a tremendous rate. As a result, finding information that is reliable and targeted to a specific query is extremely difficult.

Differences in terminology and language are major barriers to effective information access and dissemination. Those differences do not have to be as great as Polish and English. Even the nuances between British English and American English or the use of different terminology by subject area disciplines can be enough to thwart information access. If I call a certain industrial plant a "site" and you call it a "facility" we will never be able to aggregate our information about that place using standard web search methodologies. Information services to the public, industry and environmental program staff can be greatly enhanced by deploying emerging Internet technologies that draw upon new semantic management techniques and tools.

At the third semiannual meeting in Brussels on March 24 and 25, 1999, the European Environment Agency (EEA) and the U.S. Environmental Protection Agency (EPA) strengthened their commitment to provide ready access to accurate, timely and reliable information utilizing the Internet as a major mechanism for information sharing and dissemination. They expanded on-going efforts to improve information access and sharing via the Internet.

1.3 Strategy for success

The Agencies of the European Union and the United States are involved in several interrelated efforts:

- use of structured descriptions of data, text, images and other information elements disseminated on the Internet through collaboration between EEA Catalogue of Data Sources (CDS) and EPA Environmental Data Registry (EDR);
- exploration of the use of agent technology to retrieve and aggregate relevant data from

disparate databases through Internet capable browsers piloted in the Environmental Data Exchange Network project (EDEN);

- collaboration on efforts to create and interrelate warehouses of environmental information such as EEA's Environmental Information and Observation Network (EIONET) and EPA's Envirofacts.

Unfortunately, there is no magic technology which will easily pull together all the electronic information needed to solve an environmental problem regardless of language, format, and database configuration. Internet technologies are rapidly advancing, but they rely on retrieval pulled from references that employ varying terminology. Therefore the Agencies have chosen to expand and refine the General European Multilingual Environmental Thesaurus (GEMET) with its links to national/sectoral/local environmental terminology systems. The Agencies will deploy GEMET as an accepted reference terminology in their information dissemination endeavors.

2. GEMET: the vision

2.1. Need for a common set of environmental concepts

Interrelation of national thesauri and other types of word lists via a multilingual thesaurus such as GEMET creates the possibility of deploying the terminology in several ways that improve information collection, management, retrieval and dissemination. Since each specialized field has a terminology of its own and each country works in its own language, distinct thesauri are necessary. However, in order to exchange information among different fields or countries, it is advantageous to map to terms within a common thesaurus, which is general in scope. The initial step toward dispersed information systems is to interrelate thesauri/word lists in a manner that allows them to retain their autonomy yet permits integration.

2.2. Deployment of terminology

Well-defined concepts associated with the appropriate terms (whatever the language or scientific jargon) will supply technologies with needed references or hooks to bring together information on a subject, whatever its structure or form. The terminology will be deployed in

- indexing (manual or automated) heterogeneous information resources in a uniform way,
 - formulation of data element definitions and values for database construction,
 - content driven searches using search engine and web crawler retrieval mechanisms that employ thesaurus functions and topic sets,
 - design of EDI message sets,
 - cataloging information resources, and
 - intelligent agent assisted queries that cross disparate databases.
-

3. GEMET: the starting point

3.1 Developers

GEMET was developed for the European Environment Agency (EEA) by the Consiglio Nazionale delle Ricerche (CNR) and Umwelt-Bundesamt (UBA), as a reference indexing and retrieval tool for the EEA Catalogue of Data Sources and other EEA databases. It has been adopted by the United States Environmental Protection Agency (US EPA) as the reference thesaurus for their Terminology Reference System (TRS). The US EPA has collaborated in the development of definitions and the incorporation of American equivalent terms.

CNR is willing to contribute to the GEMET expansion by acting as the reference center for:

- updating of the thesaurus internal structure and base of terms in British and North American English;
- adapting the thesaurus function to the application needs of the CDS, EDR and other meta databases;
- linking between the thesaurus and the national terminology reference systems (TRS);
- displaying of GEMET as a stand alone thesaurus, periodically updated, in a suitable Web page;
- handling/upkeeping of the reference master file, including the translations. The translations are expected to be performed under co-ordination of the Agencies in the different countries;
- establishing effective communication flows for sharing documents and organising events and meetings of the thesaurus Working Group at the global level utilizing an Extranet type of tool such as EnviroWindows (see 4.2).

3.2 Structure

GEMET is a vocabulary of more than 6,500 controlled terms (keywords), representing broad environmentally significant concepts. The terms are arranged in three schemes: 1) a hierarchical classification, 2) alphabetical list, and 3) clusters of 40 environmentally relevant themes.

The 2.0 version of GEMET will be edited in British and American English and will present the equivalents in 10 European languages. In order to ensure an exact equivalence between the various languages, terms are mapped to an English concept definition. A neutral numerical identifier is assigned to each GEMET concept. The clustering of terms which name the concept around the identifier permits the retrieval across synonyms and languages.

In the environmental information systems which use GEMET as a reference thesaurus, additional terminology is maintained in the form of specialized collections which are linked to GEMET through a common interface. All of these collections can be accessed through the GEMET List of Lists (GLoL).

4. GEMET Globalization: practical implementation

4.1 Asia and Pacific Economies Cooperation (APEC) Participation

Several of the APEC economies have chosen to participate in this initiative. Chinese Taipei is heading an effort to translate the GEMET core into Mandarin. Proposals have been launched to include translations in Arabic, Bahasa Indonesian, Thai, and Vietnamese. The inclusion of these language groups will expand the environmental concepts covered in GEMET to include those specific to Asian and Pacific entities, assist in the development and refinement of the GEMET structure, and expand the retrieval of information to include work done in these languages.

Each of the APEC partners will be responsible for translation of the GEMET terminology into the target language; compilation and linkage of national terminology collections unique to the needs of the partner, and; participation in the International GEMET Working Group to review proposals for changes of structure, nomination of terms, content, linguistic equivalence, and application requirements.

Each participating country will require approximately two years to be an active partner in GEMET. One year for translation and the development of national terminology collections and one year for refinement and harmonization with the GEMET system.

4.2 Coordination mechanism

Expanding the development of GEMET across such a geographically wide area has suggested the use of an Extranet-type of connection to allow various contributors in different countries to communicate, share documents, and organize events and meetings. The EEA began, in 1998, to offer this service under its EnviroWindows project: an effort to provide a WWW-based communication interface to external partners such as non-governmental organizations and the private sector. GEMET's EnviroWindows Interest Group is currently accessible at <http://eea.eionet.eu.int:8980/envirowindows/>.

New users can request access by filling out the subscription form available through the above Web site. EnviroWindows offers the following services:

- *information* pages available in HTML;
- *library access* to download documents in several formats and software;
- *newsgroups* in which discussions among members of the workgroup can take place;
- *central user directory* containing the names of all the IG users;
- *contacts service* containing addresses of persons and organizations, and;
- *meetings* service for announcing and preparing meetings.

Access to the Library and newsgroups is also possible by e-mail. A comprehensive *search facility* assists in the identification of documents and events. Finally, a comprehensive *On-line help* service is also provided with guidelines for the use of all the available services and their internal functions.

5. GEMET: projected results

5.1 Environmental Data Exchange Network (EDEN)

Utilizing emerging information technology, the US Department of Defense (DOD), the US Department of Energy (DOE), the US Environmental Protection Agency (EPA), the National Institute of Standards and Technology (NIST), and the European Environment Agency (EEA) are participating in a collaborative effort to develop and demonstrate a means for sharing and using environmental data. Intelligent software agents and Java applets operate over the Internet to access and retrieve information from disparate data sources.

The terminology is key to the success of this approach. The terminology will serve as a reference for the creation of an ontology to bridge the gap between similar concepts expressed by different terms in the participating databases. The ontology is a set of concepts, relationships and meta-information that describes and links data in one system in a useful functional fashion to data in other systems. The ontology enables querying across data systems without incurring the cost of restructuring existing data systems or building new ones.

The EDEN project is expected to support a dynamic environment in which databases can be added or removed without affecting the basic behaviour of the system. Thus the project will be developed from a small initial group of databases but will provide the platform for incorporation of additional databases. Because the unifying feature of EDEN is the ontology, each database and tool provider adding an information source to the project will need to support the integration by mapping their system into the common EDEN ontology. GEMET will be a reference for the mapping activity.

EDEN is an ambitious effort and represents a significant step forward in the potential to organize, access and analyse environmental information. Depending on the results of this pilot, it is anticipated that the InfoSleuth technology will become commercially available.

5.2 Catalogue of Data Sources and Environmental Data Registry

The Catalogue of Data Sources (CDS) maintains a description of authoritative EEA and the European Environment Information and Observation Network (EIONET) information, such as the EIONET organisations, EEA products, main EIONET databases and the sources of national environmental information in the EEA member countries. It also includes a description of European Union environmental legislation and other environmental reporting obligations of the EEA member countries. All entries in the CDS are indexed with GEMET terms, in order to assure access to this core EEA information across language barriers.

The Environmental Data Registry (EDR), developed by the US Environmental Protection Agency (EPA), is a comprehensive, authoritative source of reference information about environmental data. It is not the environmental data itself, but rather the information that helps describe the data and make it more meaningful. The EDR serves as the clearinghouse for information about the data. It provides information on the definition, origin, source, and location of environmental data. When used in conjunction with an environmental information database, the EDR enables users to better understand the information they are accessing. It also serves as a major tool to support a standard-setting process, to record and disseminate these standards, and ultimately to facilitate data sharing between organizations and users. The EDR is linked to GEMET through the EPA Terminology Reference System and utilizes GEMET as a source of data concept and data

element definitions.

5.3 Search engines

EEA is, as a part of its European Environmental Reference Centre, developing a multilingual search service. GEMET terms are used in this service to facilitate multilingual search in full-text indexes of Web sites generated by search robots.

US EPA has plans to utilize the GEMET terminology in the development of subject or theme-specific topic sets for search engine retrieval and to load the thesaurus function of the search engine tool replacing the general thesaurus in order to improve retrieval.

5.4 Other applications

The Agencies also see GEMET as a reference tool which will function as a

- map of environmental knowledge indicating how concepts relate to each other and a reference for organization of environmental terminology into classification schemes, ontologies, glossaries, etc;
- set of well-formed definitions to serve as a guide in mapping terms and synonyms to concepts and database fields;
- nesting structure for specialized national thesauri to be linked under the more generalized central thesaurus;
- resource for assistance in translation of environmental documentation; and, as
- an index for retrieval of environmental information across languages through a common identifier

6. Summary

Environmental protection is everyone's job. A major responsibility of governmental organizations on both sides of the Atlantic is to put information in the hands of those who have an interest in protecting the environment. The Internet has become the most effective method of information dissemination, but it needs development in order to achieve its full potential. Users need to be able to bring together information across disparate databases and they need to be able to aggregate information from a variety of formats. They also need to manage the information glut and to obtain retrieval more targeted to exact needs. EEA and EPA are cooperating in addressing these Internet dissemination problems. The Agencies see semantic management through a multilingual environmental thesaurus system, GEMET, as a key to more efficient dissemination of environmental information through the Internet. They are working on the extension and refinement of GEMET and the development of a general environmental querying ability based on this tool.

