USGS ENTERPRISE WEB THESAURUS 9/23/02

Basic Design Concepts

The *USGS Enterprise Web Thesaurus* is a tool to provide better access to U.S. Geological Survey earth science information. This database gives a controlled list of subject terms to be used for tagging USGS web pages to allow searching from a formal index. Use of controlled terms insures that documents on the same subject are brought together despite variations in wording. The terms from the controlled list will be used, in addition to geographical coordinates, place name, feature name, and terms from specialized indexes, to allow faster and better retrieval of useful information from the USGS websites.

The thesaurus is limited in depth and specificity of coverage. It does not attempt to replicate existing controlled vocabularies, such as GeoRef. The intention is to provide a set of controlled vocabulary terms to be used as an additional searching tool along with free-text searching, framework organizations, and the use of other terminologies, taxonomies, and classifications.

The audience for the USGS Gateway includes three groups: professionals who work in the disciplines that are core to the scientific activities of the USGS; professionals who are educated in other disciplines; and the general public. Popular terminology, instead of highly specialized terms, is used (birds, not *aves)*, but terms used apply to specific scientific subjects and are not broad words with multiple or vague meanings.

Thesaurus terms and arrangement do not reflect the organizational structure of the USGS, but instead provide cross-cutting presentations of disciplines, methods, topics, etc. for the USGS.

Organization

The *USGS Enterprise Web Thesaurus* contains terms for the following categories or facets:

- Science (for the life sciences, Earth sciences, engineering sciences, information sciences, planetary sciences, and social sciences)
- Physical formats (to describe the distribution format, including digital formats and non-digital formats)
- Methods (computational methods; field methods, lab methods; management methods; photography; videography, and remote sensing methods)
- Object types (including audiovisual, datasets, documents, graphics, images, maps and atlases, models, terminologies and classifications, and software)
- Time period (geologic, historic, projected, real-time)
- Topics (including biological and physical processes, ecological processes, geologic and hydrologic processes, hazards, organisms, natural resources, and population and community ecology)
- USGS (including terms for USGS administrative and scientific activities, services, and facilities)

USGS Enterprise Web Thesaurus does NOT contain terms for:

- Types of named geographic features. (mountains, lakes)

- Geographic names (Rocky Mountains, Chesapeake Bay)
- Names of USGS product/publication series and formal programs.
- Biological taxonomy, mime types, geologic eras, and other detailed terminology. (The thesaurus should link to more detailed thesauri for specific purposes).
- Names of specific scientific instruments (Names may be used as non-preferred terms to link to associated terminology in the methods and topic facets.)

Each term in the *USGS Enterprise Web Thesaurus* has a listing of related terms, a definition and, where necessary, scope notes restricting the meaning of the term to that used in the thesaurus. Terms are listed in a shallow hierarchy (Broad Term-Narrow Term) and may also be viewed as an alphabetical listing.

Non-preferred terms are given which are synonyms (or near-synonyms for indexing purposes) of the preferred term. They are included in the *Thesaurus* to help catalogers and searchers locate the preferred terms.

The Thesaurus does not include every possible sub-category under a board term. Sub-categories that are especially important to the USGS or those for which numerous information resources exist will be entered. . Any sub-topics that are not specifically included will be represented at the higher level of the hierarchy.

Indexing Concepts

The purpose of indexing is to tag information objects so that a user's query retrieves all objects in the collections that are useful "answers" to the query. The indexing guidelines that enhance retrieval performance are:

1. Use of the controlled vocabulary to represent concepts; thus reducing the variety of words and phrases that can be used to express a concept.
2. Consistency in the assigning of the terms. That is, indexing similar objects with similar terms.
3. Using the most specific term from the thesaurus that is available for the concept.
4. Choice of index terms representing the major thoughts or themes in the document, not every possible idea or thing mentioned in the document. Use as many terms as is appropriate, but choose only the most evident. Do not over-index or put in every conceivable subject. Entries are not intended as an exhaustive list of subjects for each website.
5. Make informed choices of page level. In general, enter citations for web pages that are the first page of a set of pages that can stand alone without requiring the user to page back or go to another site for using or understanding of the information given.

Indexing Priorities

The sample websites chosen for indexing in the prototype will include:

1. Digital USGS publications in standard series, i.e. Open-File Reports, Bulletins
2. Major websites of USGS research programs
3. Websites of interest to the general public including some from each disciplines
4. Websites for reference use such as glossaries, classifications, etc.