A Taxonomy of Energy Subjects briefly explained.

Preprint by David Wojick, PhD, & William Watson, PhD
DOE Office of Scientific and Technical Information <http://www.osti.gov>
(send correspondence to wojickd@osti.gov)

June 4, 2004

Introduction

This paper describes an experimental taxonomy of energy related subjects. The taxonomy was developed at the DOE Office of Scientific and Technical Information, by William Watson. It is presently undergoing development and testing as an OSTI search tool, but it is offered here for general consideration by the taxonomy community. A spreadsheet version of the complete taxonomy is available on request, as are samples. Contact William Watson (watsonw@osti.gov) for details.

I. Structure of the Taxonomy.

This prototype Taxonomy of Energy Subjects is not a simple, classical hierarchy like many taxonomies. Rather, it is based on a 3-dimensional matrix. The three sides of the matrix are three different and independent category systems. These categories are, respectively, (1) scientific and technical fields, (2) an ontology of entity and property types, and (3) a collection of central scientific concepts.

The hierarchical form of the taxonomy is derived as follows. First, science and technology is categorized into a downward branching tree of fields or disciplines, sub-fields, etc., such as physics, acoustics, and so on. Thus there are branching paths down into the tree from the top, each path ending at some most specific sub-field.

Second, the most specific sub-field on each path is further categorized by types and sub-types of entities and properties. Examples include natural objects, devices and equipment, processes, and characteristics of things, as well as mental entities and mathematical objects like equations, etc.

Thus the tree at this point consists first of layers of fields, below which are layers of entity and property types. Each path passes first through a categorization into fields and sub-fields, then into types and sub-types. The types are repeated frequently for different fields or sub-fields, thus making them independent of the fields. If one were to make a matrix of the lowest sub-fields and sub-types one would find that each path corresponds to a cell in that matrix. It is important to note that not all of the cells in this matrix are represented by paths in the taxonomy. Rather, the paths correspond to just those cells that are deemed to be most important, or most active, in the relevant scientific or technical field.

Finally, to complete the taxonomy the lowest sub-types are linked to concepts. These are presently a set of descriptor terms taken from the ETDE/INIS Joint International Energy Subject Thesaurus. This Thesaurus has its own internal hierarchical structures. There are about 30,000 descriptors, grouped onto sets of narrower and broader terms. Each set has a broadest term, and there are about 3500 such sets. It is these broadest terms that are linked to the narrowest sub-types in the taxonomy. A version of the Thesaurus is available at <http://www.etde.org/edb/etdesuth.pdf>.

Thus the taxonomy consists first of fields and sub-fields, then entity and property types and sub-types, and finally of Thesaurus descriptors. As was the case with the types, the descriptors are independent of the other two categories and so may be repeated frequently, appearing along different paths. Descriptors that represent concepts of several types in many different fields appear many times; for example, the descriptor "Resonance" appears on 97 different paths. Other descriptors, for concepts of very few types mainly dealt with in only a few fields, appear fewer times; for example, the descriptor "Infrared Divergences" appears on 3 paths: under the 3 "Physics" sub-fields "Electromagnetism", "Quantum Physics", and "Particle Physics", and for each of these under the single type "Characteristics" and sub-type "Mathematical".

Thus the addition of the descriptor set to the set of fields and the set of entity and property types in effect creates a 3-dimensional matrix. Each path that includes descriptors corresponds to a cell in this larger matrix. In the present version of the taxonomy, which includes only the broadest descriptor terms, there are about 45,000 paths. (See Examples of Paths in the Taxonomy, below.) But the cells in the 3-D matrix, being all the possible combinations of sub-fields, sub-types and descriptors, number in the millions. Of these, only about 45,000 are filled in the taxonomy. This shows that the taxonomy is a very precise and narrow selection of just the most important combinations.

II. Using the Taxonomy of Energy Subjects

There are many ways to view science or technology, and every view supports a potential taxonomy. Conversely, any given taxonomy presents just one view, so its best uses will be based on that view. This taxonomy can facilitate finding information on single concepts within any field. It may also be useful for comprehensive searches for all concepts within a field for those fields the taxonomy represents. For example, a comprehensive search for descriptors of natural objects and entities related to acoustics would be easy, since acoustics is one of the sub-fields listed under physics. On the other hand, a research area like solar energy involves many of the different technical fields represented in the taxonomy, such as physics, chemistry and electrical engineering. With this taxonomy, then, a comprehensive list of descriptors about any aspect of solar energy per se would be less easy to find.

Likewise, this taxonomy is not likely to facilitate finding information about entities and properties other than those most central to the technical fields the taxonomy represents. The set of concepts at the bottom level of the taxonomy was developed to describe the main subject matter of technical reports. So, for example, finding individual Nobel prize winners would not be very easy in this taxonomy, since the closest descriptor to "Nobel prize" is the generic descriptor "Awards". On the other hand, if one wants to explore the research area specific to, say, a given Nobel prize, that is the kind of information that fits the taxonomy.

III. Examples of Paths in the Taxonomy.

Below are 17 examples of taxonomy paths, for the field "Engineering and Technology", with sub-field "Electrical and Electronic Engineering". Note how divergence can occur at different levels, as well as how the same final descriptor can occur on different paths. Note too how the descriptors range over both physical and mental objects and properties.

Engineering and Technology
Electrical and Electronic Engineering
Mathematical Objects and Entities
Attractors

Engineering and Technology
Electrical and Electronic Engineering
Mathematical Objects and Entities
Equations

Engineering and Technology
Electrical and Electronic Engineering
Mathematical Objects and Entities
Hypothesis

Engineering and Technology
Electrical and Electronic Engineering
Natural Objects and Entities
Electric Fields

Engineering and Technology
Electrical and Electronic Engineering
Natural Objects and Entities
Electromagnetic Fields

Engineering and Technology
Electrical and Electronic Engineering
Electric and Magnetic Entities
Electric Fields

Engineering and Technology
Electrical and Electronic Engineering
Electric and Magnetic Entities
Electromagnetic Fields

Engineering and Technology
Electrical and Electronic Engineering
Methods, Techniques
Discharge Quenching

Engineering and Technology
Electrical and Electronic Engineering
Methods, Techniques
Electron Scanning

Engineering and Technology
Electrical and Electronic Engineering
Methods, Techniques
Many-Dimensional Calculations

Engineering and Technology
Electrical and Electronic Engineering
Characteristics
Disturbances

Engineering and Technology
Electrical and Electronic Engineering
Characteristics
Benchmarks

Engineering and Technology
Electrical and Electronic Engineering
Characteristics
Information Needs

Engineering and Technology
Electrical and Electronic Engineering
Components, Devices, Equipment, Systems
Alarm Systems

Engineering and Technology

Electrical and Electronic Engineering
Components, Devices, Equipment, Systems
Clean Rooms

Engineering and Technology
Electrical and Electronic Engineering
Components, Devices, Equipment, Systems
Electrochemical Cells

Appendix 1

Origin of the Taxonomy of Energy Subjects

The taxonomy was originally designed to help technical abstractors find terms in the controlled vocabulary defined by the 1997 ETDE (Energy Technology Data Exchange) International Energy Subject Thesaurus <http://www.etde.org/edb/etdesuth.pdf>. It is now applicable to the successor document, the ETDE/INIS Joint International Energy Subject Thesaurus.

The Thesaurus itself constitutes one dimension; the categorization of its descriptors into the other two dimensions—fields and sub-fields, and types and sub-types—was initiated by William Watson at the DOE Office of Scientific and Technical Information. Here is Watson's account of the development of the categorization.

With over 30,000 terms, the ETDE/INIS Thesaurus is so large that someone trying to find a thesaurus descriptor for a concept they seldom have to index may spend a long time searching before either finding the most suitable descriptor, or eventually determining that no suitable descriptor is included. To solve this problem, the thesaurus has been segmented into overlapping sections—first by scientific or technical field, then by types of entities that the descriptors represent (e.g., natural objects, devices and equipment, processes, characteristics of things). Such a categorization of descriptors may be useful to information seekers as well as people indexing reports.

To begin the categorization, the Thesaurus was checked to find all the "top-level" descriptors — those that had no broader terms. There were about 3500 of these. This list was examined to see which scientific and technical fields each descriptor pertained to. Many of the descriptors related to more than one field. This was no problem, since the purpose was not to segment the thesaurus into mutually exclusive areas, but to help people avoid looking through irrelevant parts of the thesaurus for descriptors. Well before the entire list was finished with, it became apparent what most of the scientific and technical fields would be.

After the main scientific and technical fields were determined, new lists were made that included the top-level descriptors pertaining to different fields. Physics was treated first, as this field related to more top-level descriptors than any other. The physics list was printed and examined as the original list was, to determine the main subfields of physics represented and which top-level descriptors pertained to them. The next step was to make lists for each subfield and categorize the descriptors in them by the types of entities each represented. Once again, most of the kinds of entities could be determined well before finishing with the list. Whenever a large number of descriptors was found for any one type of entity (the usual case with descriptors for characteristics of things and for processes, among others), lists of these were made and similarly subsegmented. Descriptors for the other scientific and technical fields were treated similarly. The process would stop when the lowest-level categories either had 40 descriptors or less, or else (rarely) when no reasonable-looking subcategorization of a larger set was apparent.

This treatment resulted in lower-level categories whose names and implicit definitions were sometimes similar across fields but not exactly alike. Since it seemed likely that the categorization would be easier to find descriptors with if the categorization were more uniform, considerable effort was put into revising the different scientific and technical fields' lower-level category names and descriptors, to make them as uniform as possible.

Appendix 2

About the Authors

William Watson is a physicist with OSTI specializing in the design of scientific information systems. He has a Ph.D. and MS in physics from the University of Texas at Dallas, and a BS in physics from the University of Arkansas at Little Rock. He has taught physics at UT Dallas and at Eastfield College, and was on the staff of the federal project office of the Superconducting Super Collider. (WatsonW@osti.gov)

David Wojick is a consultant with OSTI in the area of information and communication science. He has a Ph.D. in philosophy of science and mathematical logic from the University of Pittsburgh, and a BS in civil engineering from Carnegie Tech. He has been on the faculty of Carnegie Mellon and the staffs of the US Office of Naval Research and the Naval Research Lab. (WojickW@osti.gov or dwojick@direcway.com)