Tools and Technology Search

(as used in O*NET OnLine)

Prepared for: National Center for O*NET Development
Post Office Box 27625
Raleigh, North Carolina 27611

Author: Jeremiah Morris
jm@whpress.com

Date: June 1, 2015

Overview

The Tools and Technology (T2) search in O*NET OnLine (www.onetonline.org) enables customers to explore careers through 60,000+ examples of machines, equipment, tools, and software used by workers on the job. Developers may use O*NET Web Services (services.onetcenter.org) to include the search in their own applications. The underlying data may be downloaded from the Developer's Corner at the O*NET Resource Center (www.onetcenter.org).

Because T2 examples are concrete and specific, workers can easily relate to the daily activities on the job. The specific examples are also classified within a multi-level hierarchy, which bridges together occupations with similar but not identical examples. This combination of specificity and broader connectivity makes T2 data well suited for a career exploration search.

T2 database structure

Examples of tools and technology objects used in an occupation are compiled by analysts through internet-based research and feedback from professional organizations and workers. Up to 300 specific examples, such as "Torx screwdrivers" or "Adobe Photoshop," are collected for each O*NET-SOC occupation. For more information on the data collection process, see O*NET Tools and Technology: A Synopsis of Data Development Procedures.

Every T2 example is classified into the United Nations Standard Products and Services Code (UNSPSC). The UNSPSC is a four-level taxonomy for the classification of products and services, provided by the United Nations Development Programme (www.unspsc.org). In the taxonomy, the Segment is the most general element and the Commodity is the most specific. One example is listed below:

Segment:	**41000000**	Laboratory and Measuring and Observing and Testing Equipment
Family:	**41110000**	Measuring and observing and testing instruments
Class:	**41116200**	Patient point of care testing supplies and equipment
Commodity:	**41116201**	Glucose monitors or meters

Examples such as "Glucometers" and "Capillary glucose monitors" are classified to this commodity. In some cases, the wording of the example matches the commodity title exactly: the example "Compasses" is classified under UNSPSC 4411803 "Compasses." Examples which match the commodity title are ignored within the search.

In O*NET OnLine, the commodity classification is called the "T2 category." The goal of the T2 search algorithm is to produce a list of T2 categories relevant to a user's query. OnLine displays each matching category, along with a selection of examples and occupations associated with that category.

Weighted search: rings, tiers, frequencies

The T2 keyword search is based on the algorithm used for occupation searches in O*NET OnLine (A Weighted O*NET Keyword Search, National Center for O*NET Development, 2013). To provide useful results across a variety of queries, several content areas and word variants are searched and weighted separately:

- **Rings** indicate each of the five textual items associated with a T2 category: the four levels of the UNSPSC hierarchy, and the individual T2 examples.
- **Tiers** classify how well a word in a user's query matches the actual content. For example, "science," "scientist," and "scienses" all match the word "sciences" at different tiers.
- **Frequencies** measure how common a word is in the database. For instance, the word "power" matches many items (power saws, power drills, etc.) but the word "inkjet" is less common.

Each word in a search query is matched against the database, and assigned a score based on ring, tier, and frequency factors. An "exact match" phase is also applied to ensure that specific title searches appear at the top of the list. The process of calculating and scoring matches is detailed in the next section.

The T2 search algorithm

From the initial user input, periods are removed, and any characters besides letters and numbers are treated as word separators. All searching is case insensitive, so the normalized query can be seen as a list of words containing only lowercase letters or numbers.

For each unique word in the search query, a "word score" is calculated on a per-category basis as follows:

- The word is exactly matched against the content items of the database. The matches in each of the five rings for each T2 category are tallied. The count of example matches is limited to 3, and examples which repeat the commodity title wording are not counted. The final counts are then multiplied by the ring value (Table 1) and tier value (Table 2). For exact matches, the tier value is 8.
- The word is then stemmed, using the Paice/Husk stemming algorithm, and compared to the stems of the database content. The matches are tallied and multiplied as above; the stemmed tier value is 3.
- The word is matched as a substring; any word in the database beginning with the letters of the current word is considered a match. The matches are tallied and multiplied as above; the substring tier value is 2.

After these steps are applied, the implementation has a set of matching categories and a word score for each category. The number of matching categories is used to calculate a word frequency factor; this factor varies between 64 and 1. Table 3 shows the factors used. The word

score is multiplied by the frequency factor to get a weighted category score for that word. The final raw score of a category is the sum of these weighted word scores.

If any words are misspelled, the process above is repeated for each unique spelling suggestion. The spelling suggestions have their own tier values, as shown in the chart below; note that the spelling substring tier value is zero, so that step may be skipped as it does not affect the score.

Once all words and spelling suggestions have been processed, an "exact match" phase is processed. The entire normalized query is checked against a normalized set of complete T2 examples and UNSPSC titles. If the query exactly matches any item, the score for the corresponding category is adjusted as follows: the category's raw score is divided by 10, and the maximum raw score from the previous steps is added. Thus, each exact match's raw score is slightly higher than any previous score. When the scores are sorted, the exact matches will rank above all non-exact matches.

Finally, the scores can be normalized and the results sorted for display to the user. The raw scores are divided by the largest raw score and then multiplied by 100 to obtain the normalized score. The normalized scores are not shown, but the results are sorted with highest normalized scores first, and ties broken by using the 8-digit UNSPSC commodity ID of each category.

Table 1 - Ring weights and filters

Thematic ring	Ring weight	Maximum count
Category title	16	1
Examples	12	3
Class title	3	1
Family title	2	1
Segment title	1	1

Table 2 - Tier weights

Search tier	Tier weight
exact word	8
stemmed word	3
substring	2
exact word (spelling suggestion)	2
stemmed word (spelling suggestion)	1
substring (spelling suggestion)	0

Table 3 - Word frequency factors

Minimum number of matching occupations	Maximum number of matching occupations	Frequency factor
1	4	64
5	9	32

10	24	16
25	49	8
50	99	4
100	399	2
400	n/a	1

Search results display

O*NET OnLine displays the top 20 matching categories, based on the scoring above, with the option to see all matches on another page. Customers may also view the specific examples and occupations associated with each matching category. Examples are displayed in two alphabetically-sorted lists: those which match the search query (at any tier) are displayed before the less relevant examples. Occupations are sorted alphabetically by title. The main result page shows the first 4 examples and 6 occupations for a category, with links to see all.