EPC Exhibit 137-28.2

February 28, 2014

 THE LIBRARY OF CONGRESS

 Dewey Section

To: Jonathan Furner, Chair

 Decimal Classification Editorial Policy Committee

Cc: Members of the Decimal Classification Editorial Policy Committee

 Karl E. Debus-López, Chief, U.S. General Division

From: Rebecca Green, Assistant Editor

Dewey Decimal Classification

OCLC Online Computer Library Center, Inc.

Via: Michael Panzer, Editor in Chief

 Dewey Decimal Classification

 OCLC Online Computer Library Center, Inc

Re: Facet detection

Attached is a paper to be presented at the 13th International Conference of the International Society for Knowledge Organization (ISKO) on the semi-automatic detection of the facets of a subject, based on the literature’s metadata (more specifically, based on the nouns and noun phrases in the literature’s titles). The theme of the conference is “Knowledge Organization in the 21st Century: Between Historical Patterns and Future Prospects.”

**Rebecca Green**

**Facet Detection Using WorldCat and WordNet**

**Abstract**

Because procedures for establishing facets tend toward subjectivity, this pilot project investigates whether the facet structure of a subject literature can be discerned automatically on the basis of its own metadata. Nouns found in the titles of works retrieved from the WorldCat bibliographic database based on Dewey number are mapped against the nodes of the WordNet noun network. Density measures are computed for these nodes to identify nodes best summarizing the title noun data / best corresponding to facets of the subject. Results of the work to date are promising enough to warrant further investigation.

**1 Introduction**

The Classification Research Group (1955) called for facet analysis—arguably the most important contribution to knowledge organization theory and practice of the twentieth-century—to become “the basis of all methods of information retrieval.” Among advantages of facet analysis, Broughton (2006, 50–51) and Kwasnik (1999, 40–41) note the following: expressivity, especially with respect to complex subjects; consistency in the linear order imposed by citation order; hospitability in accommodating new subjects; and flexibility through (1) variation of citation order to highlight different facets for different user needs, (2) the capacity of a faceted KOS to handle arbitrary combinations of subject aspects, and (3) the capability of expressing a faceted structure in different kinds of knowledge organization systems (KOSs) (e.g., classification schemes, thesauri).

Kwasnik (1999, 41), however, identifies the “difficulty of establishing appropriate facets” as a limitation of facet analysis. This difficulty is hinted at in Foskett’s (1958, 869) observation that Ranganathan, rather than giving an adequate explanation of his fundamental categories, “adopted them more or less intuitively.” Although facets are circumscribed by the topics needing to be described, their identification is nonetheless subjective to a significant degree. To the extent that the choice of facets turns out to be an ongoing decision, their identification is also intense and laborious.

The pilot project presented here aims to overcome the subjective and laborious nature of facet analysis by detecting the facet structure of a subject literature automatically on the basis of its own metadata. Specifically, we observe that the titles of the monographic literature of a subject typically include noun phrases that correspond to the facets of the subject. This correspondence may be one of equivalence, or the noun phrases may be instantiations (“manifestations” in the words of Ranganathan [1985/1962, 88]) of facets. The intent of the project is to learn how to generalize appropriately from such title data to facets.

As existing classifications such as the Dewey Decimal Classification (DDC) system adopt more faceted structures (Mitchell, 1996, *xx*), the procedures explored here may help guide the identification of facets appropriate to a subject, not only at the discipline level, but also at the level of subclasses within the discipline.

**2 Methodology**

Developing a faceted KOS is generally a bottom-up process. For a bibliographically-oriented scheme, the process begins by identifying topics, as expressed in the metadata (e.g., subject headings, titles) of a literature. Broughton (2013, 745) speaks of the “symbiotic relationship” in this process between the text of a literature and the categories or facets identified to organize that literature: on the one hand, “the text may be examined to determine categories,” while, on the other hand, the categories/facets are applied to the literature to organize it “into a structured domain (as is the case for the building of faceted KOSs).” The complex part of this process is how best to generalize from bibliographic metadata to the identification of facets.

**2.1 Identifying topics**

We begin our investigation by issuing searches against the WorldCat bibliographic database to identify relevant works in the subject areas of education, medicine, and music. Specifically, we search for English-language monographs categorized by a span of DDC numbers representing the subject area, but not including biographical or juvenile literature. The restriction to monographs and the elimination of biographical and juvenile literature targets a more academic literature whose titles are more likely to include noun and noun phrases reflecting the facets of the subject. The searches retrieved 811,914 records for the 370s (education), 1,075,535 records for the 610s (medicine), and 201,783 records for the 780s (music), the restriction to monographs having eliminated scores and recordings.

From the Title ($a) and Remainder of title ($b) subfields of the 245 field of these MARC bibliographic records, we adopt a set of heuristics to identify potential longest noun phrases, as recognized by WordNet (2010), a lexical database for English. For example, in the title *Music in the baroque era: from Monteverdi to Bach*, the process correctly identifies *music; baroque era; Monteverdi;* and *Bach*, as longest noun phrase candidates. For the title *How to read, write, and understand music*, the process identifies *read* and *music*, although *read* functions here as a verb.

The academic literature retrieved by our searches also includes titles like *Nineteenth-century music: selected proceedings of the tenth International Conference, Music for everybody: a report and pictorial review,* and *Musical growth: a process of involvement*. Nouns like *proceedings, conference, report, review*, and *process* are not music terms; they are what we might call academic terms. In the same way that natural language processing often filters out words appearing on a general stopword list, thus bypassing the further processing of frequently used terms (chiefly those in closed categories [e.g., articles, prepositions, pronouns, modals]), so too we would want to filter out academic terminology that is not specific to any subject area. To identify such words, we took the lists of words identified as potential nouns for ten subject areas (computer science, psychology, education, communication, linguistics, astronomy, medicine, architecture, music, literature [limited to works classed under DDC 808 or 809]). Words found among the 250 most frequently occurring words for eight or more of the ten subject areas are thought to be too non-subject-specific to be helpful in identifying facets and are treated as academic stopwords.

As a final step in the process of identifying the topics of a subject area, we want, if possible, to remove from the list of potential nouns those likely to be other parts of speech. For example, while both *read* and *musical* can be nouns, the former is more likely to be a verb and the latter more likely to be an adjective. Use of data from WordNet proves useful in performing this filtering process. WordNet organizes the nouns, verbs, adjectives, and adverbs of English into four distinct networks. These networks consist of nodes, each of which corresponds to a group of synonymous word senses (called *synsets*), and pointers between nodes, each of which indicates a (mostly commonly, semantic) relationship. For example, (unique) senses of the nouns *bass horn, sousaphone*, and *tuba* comprise a synset. The node for this synset is subordinate to two other nodes, one for a synset consisting of senses of *brass* and *brass instrument* and the other for a synset consisting of a sense of *bass*; the node is also superordinate to two nodes, one for a synset consisting of a sense of *euphonium* and the other for a synset consisting of senses of *helicon* and *bombardon*. An overall index to the data files indicates which synsets in which part-of-speech network a given word form appears in. This index also indicates the number of times, if any, the particular word sense has been identified/tagged in a specific text corpus. From this data, we learn how many senses a word has of various parts-of-speech in WordNet and may get a sense of how frequently the specific senses occur in naturally-occurring text. For example, from the overall index we find that *read* has 12 senses in WordNet, only one of which is a noun; all the other senses are verbs. All 169 of the tagged occurrences of *read* are verb senses. Similarly, WordNet recognizes five senses of *musical*, one, a noun sense and the other four, adjective senses. Only two of 25 tagged occurrences of *musical* are the noun sense. We retain for further processing only those potential nouns for which the percentage of WordNet senses that are noun senses or the percentage of tagged occurrences that are noun senses meet or exceed a threshold (presently set at 60%).

**2.2 Identifying facets**

 How can we turn a list of nouns and noun phrases into a list of facets? To learn how best to do this, we start with several specific facets of the discipline of music (Maples, 1995)—composer, performance medium / instrument, and historical period—and consider how they are manifest as nouns and noun phrases in the titles of bibliographic works:

1993 *songwriter*'s market: where & how to market your songs

Adding *percussion* to *medieval* and *renaissance* music

*Bach* and the *pedal clavichord*: an *organ*ist's guide

*Campion*, *Dowland* and the *lute*nist *songwriters*

Dutch *woodwind instruments* and their makers, *1660–1760*

Guide to teaching *woodwinds*; *flute*, *oboe, clarinet, bassoon, saxophone.*

*Late Renaissance* and *baroque* music (*c. 1525–c. 1750*)

Music in the *baroque era*: from *Monteverdi* to *Bach*

Music theory and its sources: *antiquity* and the *Middle Ages*

*Scott Joplin* and the *ragtime era*

The best of *Bacharach* for *saxophone*

The breath of the *symphonist*: *Shostakovich*'s tenth

The great German *composers*, *Bach* to *Dvôrak*

The world of *baroque* and *classical musical instruments*

In these titles the composer facet is referred to by *composers, songwriters, symphonist, Bach, Bacharach, Campion, Dowland, Dvôrak, Scott Joplin, Monteverdi,* and *Shostakovich*; the musical instrument facet by *musical instruments, percussion, woodwind instruments,* *woodwinds*, *flute*, *oboe, clarinet, bassoon, saxophone, pedal clavichord, lute(nist),* and *organ(ist)*; and the historical period facet by *antiquity,* *Middle Ages,* *medieval* and *renaissance* (music)*, 1660–1760, late Renaissance* and *baroque* (music) *(c. 1525–c. 1750)*, *baroque era*, *baroque* and *classical* (musical instruments), and *ragtime era.* Several patterns emerge: facets may be referred to by terms at the level of the facet (e.g., *composers, musical instruments*) or by terms for classes subordinate to the facet at several levels of specificity, from the general (e.g., *songwriters, woodwinds)* to the more specific (e.g., *flute, pedal harpsichord*), to the level of instances (e.g., *Bach, Scott Joplin*). Historical periods, which are, by nature, instances, may be designated by name (e.g., *Middle Ages,* *baroque era, ragtime era*) or by time range (e.g., *c. 1525–c. 1750, 1660–1760*). Words and phrases may name a facet directly (e.g., *oboe, Monteverdi*) or may combine reference to multiple facets (e.g., *lutenist, baroque* . . . *musical instruments*).

What we need is a means of noting the convergent use of terms at various levels within a single hierarchy. Although not perfect for the task—WordNet has never claimed to be an ontology—it is the best resource available. Borrowing from the conceptual density measure developed by Agirre and Rigau (1996), we propose to detect the convergent use of terms referring to the same facet by locating title nouns in the WordNet noun network as identified in section 2.1 and calculating the density of trees under individual nodes.

**3 Results**

 We start by considering two sets of data we make use of in our analysis. Data in the first set (see Table 1) characterize individual nodes in the WordNet noun network; data in the second set (see Table 2) summarize the behavior of our title nouns relative to the WordNet noun network.

**Table 1. Data characterizing nodes in WordNet noun network**

|  |  |
| --- | --- |
| **Measure** | **Definition** |
| level | How many levels removed this node is from its most distant “progeny.” Leaf nodes (nodes with no children) are defined as level 0; each other node has a level value that is 1 greater than the highest level of any of its children. Just over 79% of the nodes (64,958 of 82,115) are at level 0; the node at the top of the network is at level 18. |
| num\_words | How many words belong to this node |
| treesize\_nodes | How many nodes belong to the tree of which this node is the top node  |

**Table 2. Data summarizing title noun behavior relative to WordNet noun network**

|  |  |
| --- | --- |
| **Measure** | **Definition** |
| num\_words\_attested | How many words belonging to this node are attested in the title data |
| dense\_node | 1, if (num\_words\_attested / num\_words) >= 0.5, else 0 |
| num\_tree\_nodes\_ attested | A count of how many nodes belonging to the tree of which this node is the top node are attested by words in the title data |
| ctree\_nodes\_attested | Like num\_tree\_nodes\_attested, except that each attested word contributes a normalized measure of its frequency of occurrence in the input (instead of 1)  |
| num\_tree\_dense\_ nodes\_attested | A count of how many nodes belonging to the tree of which this node is the top node are dense |

 We have considered the use of 10 measures of a node’s density, as reflected in Table 3. Several of the measures take into account treesize\_factor, a measure of how large the tree under a node is relative to other trees of the same level. Its use is based on the assumption that, for example, a high value for %\_treesize\_nodes—the proportion of nodes in the tree under a node that are attested in the title data—is more significant the larger the tree is.

**Table 3. Density measures for nodes**

|  |  |
| --- | --- |
| **Measure** | **Definition** |
| %\_treesize\_nodes | num\_tree\_nodes\_attested / treesize\_nodes |
| %\_treesize\_cnodes | ctree\_nodes\_attested / treesize\_nodes |
| %\_treesize\_dense\_nodes | num\_tree\_dense\_nodes\_attested / treesize\_nodes |
| TSFd\_%\_treesize\_nodes | (num\_tree\_nodes\_attested / treesize\_nodes) x treesize\_factor |
| TSFd\_%\_treesize\_cnodes | (ctree\_nodes\_attested / treesize\_nodes) x treesize\_factor |
| TSFd\_%\_treesize\_dense\_nodes | (num\_tree\_dense\_nodes\_attested / treesize\_nodes) x treesize\_factor |
| TSF\_sum | TSFd\_%\_treesize\_nodes + TSFd\_%\_treesize\_cnodes + TSFd\_%\_treesize\_dense\_nodes |
| TSF\_prod | TSFd\_%\_treesize\_nodes x TSFd\_%\_treesize\_cnodes x TSFd\_%\_treesize\_dense\_nodes |
| TSF\_hi2\_sum | Sum of highest 2 of TSFd\_%\_treesize\_nodes, TSFd\_%\_treesize\_cnodes, and TSFd\_%\_treesize\_dense\_nodes |
| TSF\_hi2\_prod | Product of highest 2 of TSFd\_%\_treesize\_nodes, TSFd\_%\_treesize\_cnodes, and TSFd\_%\_treesize\_dense\_nodes |

Given the investigation is in only a preliminary stage, we first analyze the values computed for these measures with reference to music, where a fair degree of consensus has been reached on its facet structure. The findings of this initial investigation are then applied to education and medicine to see how well they generalize.

**3.1 Discerning facets of music**

 In addition to the facets of composer, instrument, and historical period noted above, oft-suggested music facets also include form/genre of composition, space / cultural influence, and physical format (Maple, 1995); we have purposely omitted the latter facet from our searches and so will not consider it further. For each of the remaining facets, we first summarize its position in the WordNet noun network, after which we consider the behavior of its closest node with respect to the data measures in Table 3.

 Composers are captured in WordNet by a node to which belongs the single word *composer*. The node has 137 children, 4 of which are for composers of various types of music (e.g., symphonist) and the remainder of which are for specific composers. All of the children of the composer node are leaf nodes (that is, the *composer* node is at level 1); the typical child node gives 2 or 3 ways of referring to a specific composer (e.g., *Bach, Johann Sebastian Bach*).

 Musical instruments are captured in WordNet by a node to which belong *musical instrument* and *instrument*. The node has 10 children, several of which (e.g., nodes for *keyboard instrument, stringed instrument, wind instrument*) head trees going down multiple levels. The *musical instrument* node is at level 7 and sits atop a tree of 164 nodes, consisting altogether of 299 word senses.

 Historical periods are captured in WordNet by a node to which belong *historic period* and *age.* The node has 19 children, all of which are leaf nodes (that is, the *historic period* node is at level 1), no more than half of which would be used for musical historical periods (e.g., *New Deal* would not). The tree of which it is the head has 20 nodes and 29 words. (On the one hand, we would not expect the musical historical periods tree to be sizable; on the other hand, WordNet does not provide for musical historic periods very well: for example, it has no node for the classical period, the romantic period, or even the modern age.)

 Musical forms/genres are captured in WordNet by a node to which belong *music genre, musical genre, musical style,* and g*enre*. The node has 5 children (e.g., nodes for classical music, religious music, and popular music), all of which are further developed. The music genre node is at level 4 and sits atop a tree of 67 nodes and 117 words. Most of the nodes within the tree express musical style, but within the classical music subtree one also finds nodes for musical forms, e.g., concerto grosso, fugue.

 Space / cultural influence is captured in WordNet by a node to which belong the words *country, state*, and *land*. The node has 59 children, a small number of which are kinds of countries (e.g., sultanate), a small number of which are groups of countries (e.g., North American country), but most of which are individual countries. The node, which is at level 3, sits atop a tree of 226 nodes and 576 word senses.

 The nouns in our title data from the literature for music are mapped to 6701 WordNet nodes, for which the ten density measures in Table 3 are computed. We eliminate 41 nodes from further consideration for being too specific (level = 0) or too broad (level >= 10) to function as facet correspondents, leaving 6660 nodes to be ordered by the density measures. We do not expect the first three measures to be very useful, because they take into account neither the size of the tree below the node nor the level of the node; we will mention them only when they perform better than the other measures.

 In the sorting of the 6660 nodes for each of the remaining measures, the composer node is found in positions 14–36 (median 29), while musical\_instrument appears in positions 121–4358 (median 242), music\_genre in positions 40–92 (median 76), historic\_period in positions 30–305 (median 200), and space in positions 28–116 (median 41). On the one hand, the nodes corresponding to facets are not surfacing to the very top of the sorted list of nodes, but, on the other hand, the median density scores are all within the top 5% of nodes, while some scores are within the top .2%. The premise that title nouns mapped to the WordNet noun network will generate high density scores in the nodes corresponding to facets is supported generally, but significant noise obscures them.

 It is worth noting why the musical\_instrument node fares poorly. Most of the density scores reward nodes that head trees with a greater number of subordinate nodes for the level, but the tree under musical\_instrument has a lesser number of subordinate nodes; the node at the top of the tree is accordingly penalized. The musical\_instrument node has its highest ordering when the density measure used is %\_treesize\_cnodes, that is, when the frequency of nouns in the title data input is taken into account, but the relative tree size under the node is not.

 It is also worth considering the nodes that score higher than music facet nodes when the density measure used produces the best results, which is TSF\_prod (for which the composer, music\_genre, and space facets are ranked 14, 40, and 28, respectively). The first 20 nodes in that output are for senses of *person, communication, location, region, activity, geographical area, message, creator,* musician (as composer or conductor), *action, performer, structure, artist, composer, musician* (as performer), *change, written communication, entertainer, measure, device*. Some of these are more general than what we seek as subject-specific facets (e.g., person, activity, action), while others are closely related to facet nodes (as, for example, *location, region,* and *geographical area* are akin to our space facet). Further avenues for improvement of the results are suggested: (1) general nodes may be pre-specified by enumeration rather than by level, and (2) results may be presented as sets of related nodes rather than as individual nodes.

**3.2 Generalizing to education and medicine**

Here we note only general findings for our investigations in education and medicine. The results for education are somewhat disappointing, with nodes for students and teaching methods too far down to be easily discoverable, and with nodes for course of study, educator, and educational institution or school ranking high on different measures. The results for medicine are more promising: disease/illness and drug nodes both score high on at least half of the measures; health professional and body system score relatively high on several measures; patients, however, are practically lost.

**4 Future work**

 As a preliminary investigation, results to date are somewhat mixed: nodes in the WordNet noun network that correspond to facets tend to score high on some of the measures we are studying, but so do other nodes. Among the avenues we will consider in the future are different thresholds (e.g., for identifying academic stopwords, for identifying dense nodes), more targeted means of word sense disambiguation (e.g., at the title level), different measures (e.g., measures that take into account the behavior of a node’s children), pruning of output by enumerating nodes that are too broad and by presenting sets of related nodes, the use of subject-specific words in the presentation of nodes (e.g., using *virtuoso* instead of *ace* to capture the node glossed as “someone who is dazzlingly skilled in any field”), and the possibility of using different profiles at different levels of the network (e.g., facet-like nodes whose children are instances vs. those whose children are subclasses). We have treated WordNet throughout as if it were a tree, but over 2200 nodes (2.7%) have two parents, which may affect results; further investigation is warranted. Further study is also needed in a broader range of subject areas and at different levels of specificity of subject areas.

**5 Conclusion**

The goal of detecting the facet structure of a subject in the title nouns of its literature as mapped against the WordNet noun network makes two significant assumptions. The first assumption, that nouns corresponding to the facets of a subject will occur frequently in the titles of its academic literature, appears to be true to a significant degree. (But instances of academic terminology in these titles need to be filtered out.) The truth of the second assumption, that when title nouns are mapped against the WordNet noun network, the nodes in that network with the highest degree of attestation will correspond to facets of the subject, depends on significant interpretation of both input and output. Work to date has been promising enough to warrant further investigation.

**References**

Agirre, Eneko & Rigau, German. 1996. Word sense disambiguation using conceptual density. In Proceedings of the 16th conference on Computational linguistics (COLING '96), Vol. 1. 16-22. [https://aclweb.org/anthology/C/C96/C96-1005.pdf]. Accessed 29 January 2014.

Broughton, Vanda. 2006. The need for a faceted classification as the basis of all methods of information retrieval. *Aslib Proceedings.* 58 (1): 49–72.

Broughton, Vanda. 2013. Faceted classification as a general theory for knowledge organization. *SRELS Journal of Information Management,* 50/6, 735-750.

Classification Research Group. 1955. The need for a faceted classification as the basis of all methods of information retrieval. *Library Association Record*, 57(7): 262-8.

Foskett, D.J. 1958. The construction of a faceted classification. In*Proceedings of the International Conference on Scientific Information, 1959 (*Washington, DC: National Academy of Sciences), 867–888.

Kwasnik, Barbara H. 1999. The role of classification in knowledge representation and discovery. Library Trends 48 (1): 22–47.

Maple, Amanda. 1995. Faceted Access: A Review of the Literature. [http://library.music.indiana.edu/tech\_s/mla/facacc.rev]. Accessed 29 January 2014.

Mitchell, Joan S. 1996. New features in Edition 21. In *Dewey Decimal Classification and Relative Index*, devised by Melvil Dewey; Edition 21, edited by Joan S. Mitchell, Julianne Beall, Winton E. Matthews, Jr., & Gregory R. New (Albany, New York: Forest Press).

Ranganathan, S. R. 1985/1962. Facet analysis: Fundamental categories. In *Theory of subject analysis: a sourcebook*, edited by Lois Mai Chan, Phyllis A. Richmond, and Elaine Svenonius.( Littleton, Colo: Libraries Unlimited), 88–93. Reprinted from his *Elements of Library Classification*, 3rd ed. (Bombay, New York: Asia Publishing House).

WorldCat. OCLC Online Computer Library Center, Inc. 2013. A public-accessible interface to the database is accessible at http://worldcat.org. Accessed 29 January 2014.

WordNet. Princeton University. 2010. [http://wordnet.princeton.edu]. Accessed 29 January 2014.