Extending the Visualization of Classification Interaction with Semantic Associations

Richard P. Smiraglia
Knowledge Organization Research Group
School of Information Studies
University of Wisconsin, Milwaukee
smiragli@uwm.edu

ABSTRACT
General classification schemes hold the potential for being applied to large quantities of information resources. Yet the underlying infrastructure requires empirical understanding of the interaction between classifications and their inherent characteristics, as well as the inherent characteristics of the resources they classify. An important step is described here based on an attempt to derive terms from subject vocabularies (subject headings, index terms, terms from thesauri) in relation to UDC strings extracted from a random sample of KU Leuven MARC records and OCLC WorldCat MARC records. Results show see the clear presence of semantic clusters, which in future research might be generated from UDC strings and associated with other statistically-significant correlations to develop a navigable classificatory infrastructure for data-mining and information-sharing.

Keywords
Classification interaction, cultural warrant, Universal Decimal Classification, navigable classificatory infrastructure

1.0 REVEALING CONTOURS OF CLASSIFICATORY INFRASTRUCTURE

General classification schemes have been applied to millions of bibliographic records, and hold the potential for being applied to still larger quantities of information resources. Yet the underlying infrastructure will require empirical understanding of the interaction between classifications and their inherent characteristics, on the one hand, and the inherent characteristics of the resources, on the other. Research to reveal these rich contours is still in its nascent stages, but holds promise for pointing toward "a dialogic and contrapuntal relationship" (ASIST SIG/CR 2014) that could provide a navigable classificatory infrastructure not only for data-mining and information-sharing but also for revealing heretofore undiscovered knowledge relationships.

One relevant research stream is based on an elementary theory of knowledge organization that combines empirical understanding of bibliographic characteristics in interaction with traditional concept-based classifications (Smiraglia and van den Heuvel 2013). Such a theory leads naturally to the concept of an interactive faceted approach to classification in which facets allow switching among empirical dimensions from conceptual to instantiation structures as a means of teasing out interactions between elementary structures of knowledge (Smiraglia, van den Heuvel and Dousa 2011). An ongoing analysis of the evolution of the underlying network of knowledge in Wikipedia used as a control a parallel analysis of the Universal Decimal Classification (UDC) and its underlying network structure (Scharnhorst et al. 2012). To tease out the implications of classification interaction the team analyzed the use of UDC as represented in nine million UDC numbers extracted from the OCLC WorldCat (Akdag Salah et al. 2012), and ninety-five thousand extracted from the online catalog of Catholic University of Leuven (KU Leuven) (Smiraglia et al. 2013).

To discover whether predictable co-occurrences existed among elements of faceted UDC numbers and elements of MARC-tagged bibliographic records, Smiraglia (2014a and 2014b) mapped components of each. Statistically significant correlations occurred among most of the deconstructed components of the UDC numbers as well as among the MARC-designated elements of the respective bibliographic records. Interestingly, statistically significant correlations between the elements of classification and the bibliographic elements in each locus (OCLC WorldCat and KU Leuven) yielded an underlying network structure. Green (2014) and Green and Panzer (2014) demonstrated similar results by using bibliographic elements to generate classification

Advances in Classification Research, 2014, November 1, 2014, Seattle, WA, USA.

infrastructure for the *Dewey Decimal Classification* (*DDC*).

These studies represent beginning stages in empirical understanding of the potential for navigating relationships between classifications and the resources they describe, as well as among classifications and potentially applicable faceted structures. So far, results suggest classified associations as one means of data-mining for information and resource discovery. Next steps for research include the analysis of data values alongside the analyses of data structures to map statistically significant semantic and conceptual associations, along with those already uncovered. This paper takes that next step by presenting analyses of terms in subject vocabularies captured in Smiraglia (2014a and 2014b). Recent research (Ridenour 2014) has demonstrated a network among shared semantic boundary objects in discrete domains.

2.0 METHODOLOGY

Subject vocabularies (subject headings, index terms, terms from thesauri) extracted from a random sample of KU Leuven MARC records and OCLC WorldCat MARC records were subjected to analysis to map conceptual similarities, and then analyzed for statistically-significant co-occurrence across the decomposed UDC and bibliographic entities. Essentially, the samples used in Smiraglia (2014a and b) were subjected to extended analysis. Sample size was calculated for the two prior studies (Smiraglia 2014a and b) using estimates of correlation among classification and bibliographic characteristics from Smiraglia 1992. Estimating confidence at 95% with an acceptable confidence interval of ±5%, 329 elements were necessary.

The OCLC Office of Research had provided the KSL team with 9,055,623 UDC numbers bibliographic records using the MARC 080 field in the WorldCat. Excel was used to generate 400 number pairs. A sample of 398 records was located using the OCLC Connexion platform, and MARC text records were downloaded for all of them.

The KU Leuven sample was more complicated. The KSL team had also received 95,544 UDC call numbers from KU Leuven. This was output from the KU Leuven authority file, in which UDC "call" numbers were matched with verbal terms. Each number/term pair was pasted into a spreadsheet, and the spreadsheet line numbers were used to generate a random sample. The KU Leuven online catalog was searched for each UDC number the record paired with the terms in our output was selected. There were no duplicates. It was not possible to locate bibliographic records for 22 of the number pairs. This yielded a random sample of 378 elements.

In both cases the samples were large enough to provide 95% confidence for generalizing our results within ±5% to the populations of bibliographic records at the time of the original data output. As it happened, results such as date of publication and UDC population conformed precisely to the KSL's analysis of the full set, demonstrating this generalizability.

3.0 RESULTS

3.1 Sample characteristics

The basic bibliographic shape of the sample from the WorldCat was reported in Smiraglia (2014a, 2-3):

> Slightly more than half of the works have ISBN standard numbers, less than a quarter have edition statements and about a third have series statements. Slightly more than a quarter have bibliographies noted, and only 2.5% have linked electronic texts.

Similarly, the KU Leuven sample was reported in Smiraglia (2014b, 178):

> For example, 63.2% of the works had series statements. Slightly less than a third are in English and slightly less than half are in French or Dutch, with a smattering of other languages. 39.2% have ISBNs, 10.6% have edition statements, and 38.6% have series statements.

Dates of publication associated with the bibliographic records in the samples provide a very interesting analytical lens. The UDC was created originally in 1905 and has been used continuously since. But the dates in the samples show a much different profile. Dates of publication in the WorldCat ranged from 1606 to 2009 (Figure 1), but the majority of works are dated after 1979.

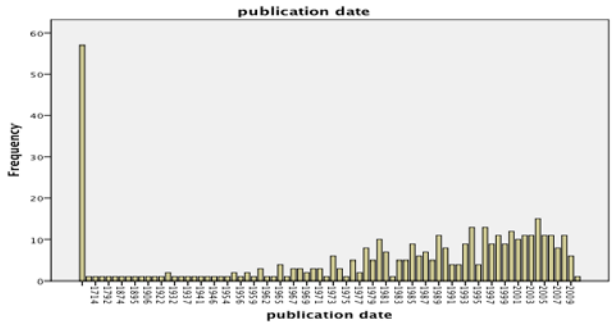

Figure 1. Dates of publication in the WorldCat sample

This likely is an artifact of the OCLC WorldCat; 34.2% of the works in the distribution have no date or pre-date 1979 and these likely represent works for which cataloging has been converted, but the majority of the cataloging is for works cataloged using the WorldCat in the last quarter century. The mean age of work was 23 years, and the median age was 15 years.

At KU Leuven dates of publication ranged from 1599 to 2011. The mean age of work was 44 years; the chronological midpoint of the distribution was 1805 but the median age of work was 30 years which made 1981 the midpoint of the population, and the mode was an age

of 22 years, meaning most works dated from 1989-2011 (Table 2).

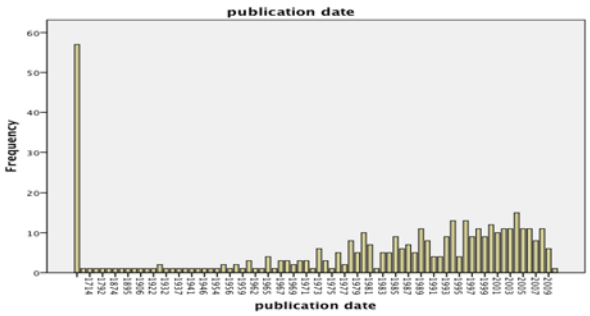

Figure 2. Dates of publication in the KU Leuven sample

This distribution likely is also an artifact of the online era, although the longer, flatter distribution reflects the characteristics of a university library versus those of the WorldCat. In both cases the majority of the works post-date online cataloging (for instance, OCLC was created in 1967). Thus our view of the usage of UDC is limited to cataloging mostly from the online era, enhanced of course by selective retrospective conversion.

Language of text was easily extracted for the Leuven sample; a frequency distribution is shown in Figure 3.

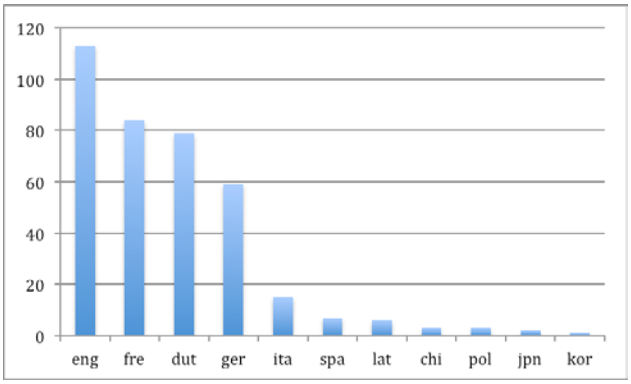

Figure 3. Languages of texts in Leuven sample

Although the range of languages is relatively global, the majority of texts in the sample are in English, French, Dutch or German, which is consistent with cultural norms in Belgium.

3.2 Sample population of the UDC

The UDC population in the WorldCat sample is shown in Figure 4.

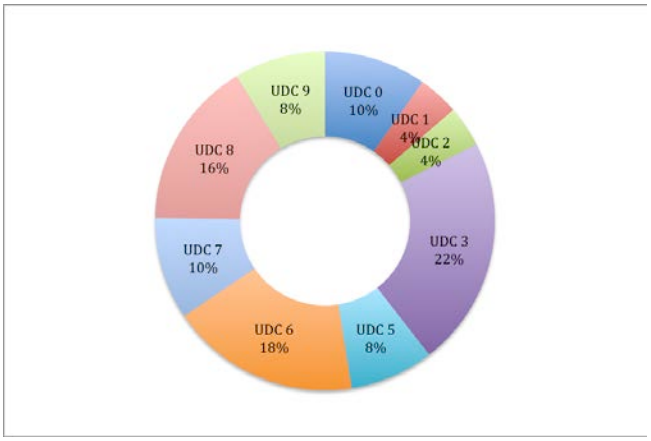

Figure 4. Population of the main UDC classes in WorldCat

This distribution matches the KSL team's analysis of the full population of numbers. The visualization shows how well distributed the population is in most main classes except 1 "Philosophy. Psychology" and 2 "Religion. Theology." The majority of the records are classed in 3 "Social Sciences," 6 "Applied Sciences," and 8 "Language and Literature." The auxiliary connecting devices were identified as well (but are not shown in the figure). "+" Addition (e.g., France and Spain, or Mining and Metallurgy, etc.), ":" Simple relations (e.g., ethics in relation to art, influence of politics on education, etc.), and "/" Consecutive extension (connects the first and last of a series of numbers to denote a range, or a broad subject) are the most used. "+" occurred 5 times, ":" 33 times, and "/" 31 times. The three operators were cross-tabulated; there were no statistically significant correlations among them.

The population of the UDC in the KU Leuven sample is shown in Figures 5-7, including main classes, and linkages between main classes and auxiliary signs.

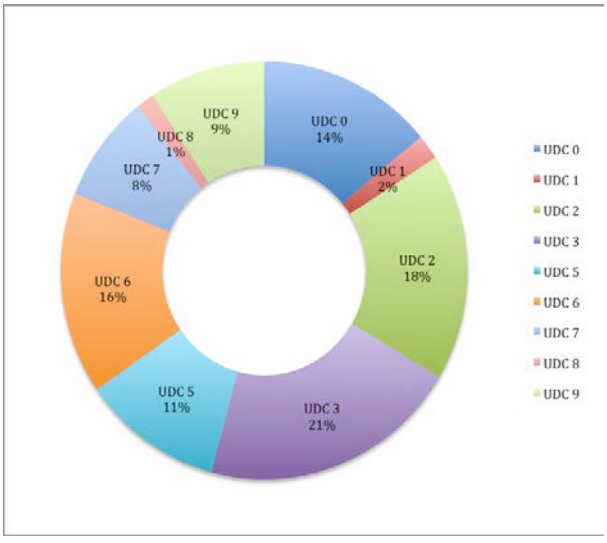

Figure 5. Main UDC classes in the sample

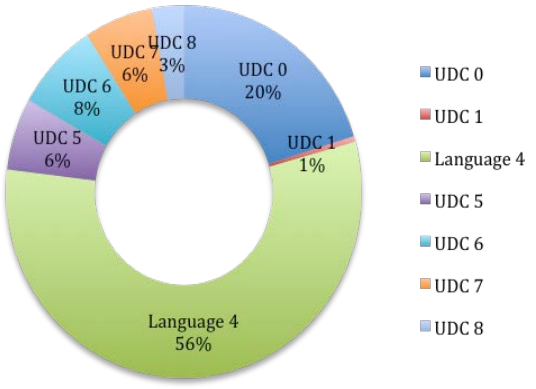

Figure 6. UDC classes linked with common auxiliary signs in the sample

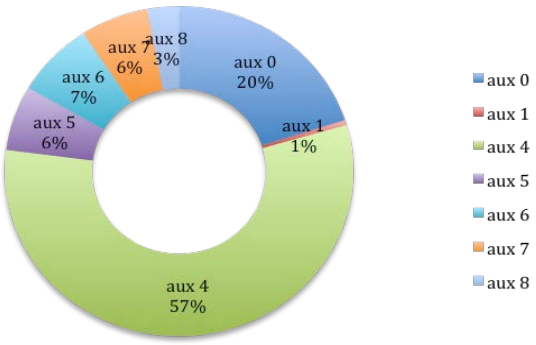

Figure 7. Common auxiliaries linked to main classes in the sample

Figure 4 shows a fairly reasonable disciplinary distribution for a major research university with an emphasis on humanities and social sciences. The largest proportions fall in 2 "Religion. Theology," 3 "Social sciences," and 6 "Applied Sciences. Medicine. Technology." Relatively little falls in philosophy, literature or the arts. Generalities is a large class likely reflecting a population of works assigned to reference use. Figures 6 and 7 show us that most use of the auxiliaries is made for linking place and language, and the largest cluster of those linkages falls in works classed in generalities, although every main class shows linkage with auxiliaries to some extent.

3.3 Sample places of publication and publishers

Figures 8 and 9 show the distribution of places of publication and publishers in the WorldCat.

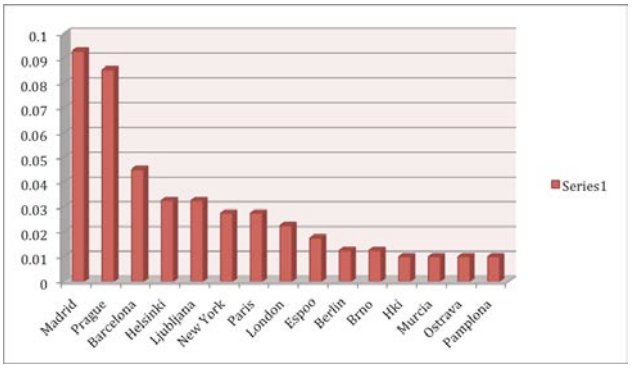

Figure 8. Places of publication in WorldCat sample

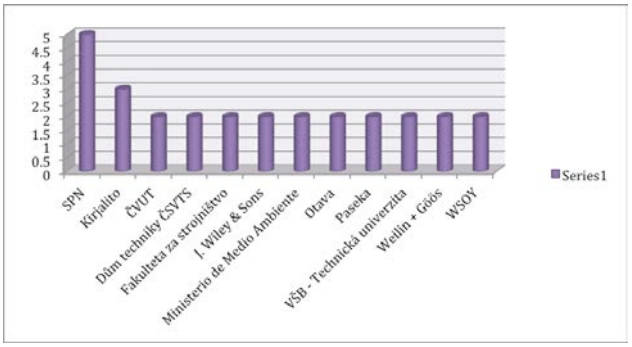

Figure 9. Publishers in the WorldCat sample

In the WorldCat, 41 places of publication occurred more than once; Madrid (37 times or 9%). Prague 85, Barcelona 4%. Etc. 91 place names occurred only once (23%). 12 publishers occurred more than once; one with 5 occurrences, which still is only 1%. 306 occurred only once or 76%.

Figures 10 and 11 show places of publication and publisher names from the KU Leuven sample.

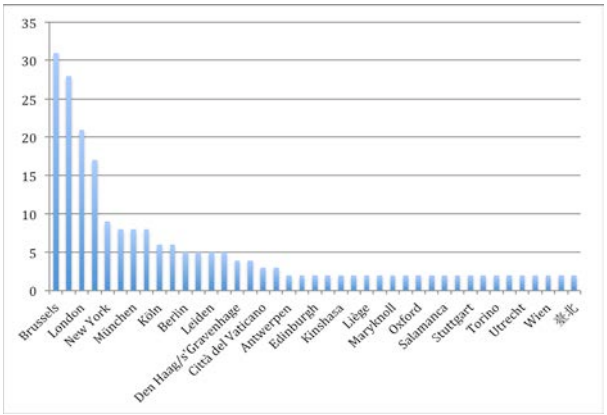

Figure 10. Places of publication in the KU Leuven sample

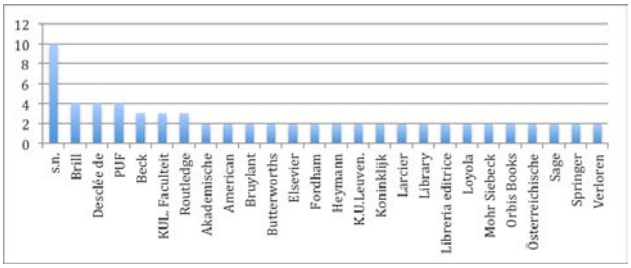

Figure 11. Publishers in the KU Leuven sample

Figures 10 and 11 show places of publication and publisher names from the KU Leuven sample. There were 168 place names in the KU Leuven sample; 41 names occurred more than once, and 18 more than twice. Not surprisingly, the long tail (not pictured) included mostly European place names, many from Belgium and The Netherlands. There were 308 publishers named, but only 26 appeared more than once, and only 6 appeared more than twice. "S.n." occupied the largest category in the distribution. An attempt was made to correlated language, place and publisher. Because of the flat frequency distributions of place and publisher, it was not possible to demonstrate statistically-significant correlations. A cross-tabulation shows that only Brussels and Leuven produced enough works in English, French, Dutch and German to potentially demonstrate a correlation. There is only one publisher, Duclée, with a high-enough frequency for cross-tabulation; no correlations can be demonstrated for these variables in these samples.

3.4 Correlation of place and publisher with UDC

Six place names in the WorldCat distribution had enough occurrences to cross-tabulate with UDC main classes. Only 2 publisher names occurred often enough to cross-tabulate. Tables 1 and 2 shows the result of these cross-tabulations.

	0	1	2	3	5	6	7	8	9	+	:	/
Madrid			■	■							■	
Prague										■		
Barcelona						■					■	
Helsinki						■						
Ljublijana						■						
New York												

Table 1. Places correlated with UDC in the WorldCat sample

	0	1	2	3	5	6	7	8	9	+	:	/
SPN										■		
Kirjalito								■				

Table 2. Publishers correlated with UDC in the WorldCat sample

These results were disappointing. There are two ways to look at this result. First, obviously, there were too few instances of place names or publisher names in the sample to generate statistically-significant correlations. But the second way to consider this result is to consider that place and publisher names are, in fact, poorly coordinated with intellectual indicators to provide other than secondary meaning in any data-mining using deconstructed classification strings. In the Leuven data only one publisher occurred 5 times, which prohibited cross-tabulation.

3.5 Subject values in 65x fields

There were 864 65x fields in the WorldCat sample; one single record had 19 65x fields. There were, however, only 99 headings that occurred more than once, only 20 that occurred 3 or more times. These are shown in Table 3.

Sborníky	**9**
Marruecos	**7**
Učebnice vysokých škol	**7**
Energía de la biomasa	**4**
Křesťanský život	**4**
Painting	**4**
Painting, Modern	**4**
Sborníky konferencí	**4**
Asociaciones	**3**
Brožury	**3**
Česko	**3**
Christian life	**3**
Español (lengua)	**3**
Katalogy	**3**
Literatura española	**3**
malířství	**3**

Populárně-naučné publikace	**3**
Příručky	**3**
Textbooks (higher)	**3**
turizem	**3**

Table 3. Frequency distribution of 65x terms from WorldCat sample

Although no specific strings occurred with high enough frequencies to generate correlations, it is plain to see how semantic clusters might emerge from the data. For example, "painting" occurred 8 times in 2 strings, shown in Table 3. By gathering all of the usages from the associated UDC strings, the semantic clustering shown in Figure 12 was developed.

-41	Painting, Czech	Painting, Modern--20th century--Czech Republic	Painting, Modern--21st century--Czech Republic	
-437.313	Painting, Modern--19th century--Russia	Painting, Modern--20th century--Russia	Painting--Czech Republic--Náchod	Painting, Russian
7.03	Painting--19th-20th centuries			
72/76	Painting			
75-x(1/9)	Gothic painting--mural painting--Slovenia			
75-x(1/9)	painting--mural painting—frescoes--17th/18th cent.--Italy			

Figure 12. Semantic cluster "Painting"

In the Leuven sample 1,005 subject headings were assigned in 69x fields to 400 records. These are not exactly subject headings in the traditional sense; rather they are subject term strings developed from the UDC hierarchies (the technique is described in Schallier 2004). In most cases the final string is subdivided by a personal name. Interestingly most strings occur twice but not more than twice. Eight strings occurred three or more times. These are shown in Table 4.

Forests and forestry from the national point of view. Forestry policy. Afforestation policy. Social economics of forestry--Colombia	**5**
Handschriftenkunde. Handschriftencatalogi--Grieks	**6**
Handschriftenkunde. Handschriftencatalogi--Italië	**4**
Handschriftenkunde. Handschriftencatalogi--THEOPHANES NICAENUS	**4**
Kerkgeschiedenis--Frankrijk-- 18e en 19e eeuw. Periode 1700-1899.--(eveneens voor boeken over recht periode 1789-1799)	**4**
Missies. Evangelisaite. Zending--China	**3**
Missions. Evangelisation--China	**3**
Rijksoverheidsbibliotheken. Staatsbibliotheken. Nationale bibliotheken (Koninklijke Bibliotheek)--Portugal--LISBOA	**3**

Table 4. Frequently occurring subject strings in Leuven sample

The low frequency of occurrence makes it difficult to generate statistical analyses, such as cross-tabulations with other variables. But it is clear that there are frequently-occurring semantic clusters, such as those related to missionary evangelization of China, which occurs 6 times in two different strings. To illustrate, the list of 1005 strings were entered into Voyant software (http://voyant-tools.org/) to generate a term distribution. The result is shown in Figure 13.

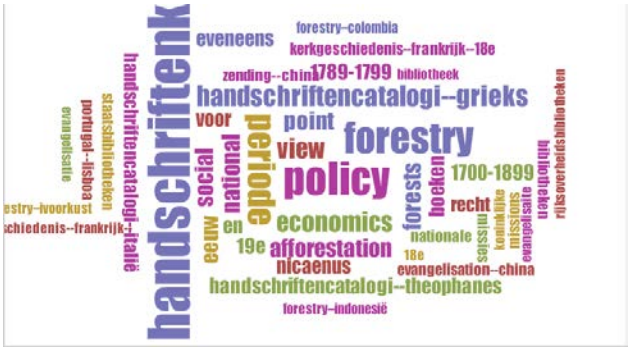

Figure 13. Frequently occurring subject terms in Leuven sample

In fact, about ten terms, those in largest type in the word cloud, occur more than 5 times in the sample. These could provide a core of semantic clusters for further future analysis.

4.0 Discussion and conclusions

In this third search for statistically-significant correlations between elements of classification and the resource descriptions to which they are assigned, further support for this research stream has emerged. Although stronger correlations were reported in the earlier papers for associations among elements of UDC strings and bibliographic characteristics, we still have seen a few correlations emerge here, such as those among specific places of publication and UDC main classes.

More importantly, the parallel analysis of samples from two populations of classified resources has been very revealing. We see, as has been shown in studies of instantation (Smiraglia 2001), much broader and flatter distributions in the WorldCat than in a specific library collection. We see clearly how contributions to the WorldCat from specific UDC libraries has influenced the structure of its contents. We see also how the specific

requirements of the KU Leuven are reflected in its collection. In both cases it seems apparent we are looking only at post-automation cataloging for the most part, rather than at retrospectively-converted cataloging for works from across time. These observations are very important for generating further research based on classified bodies of data.

Finally we see the clear presence of semantic clusters, which in future research might be generated from UDC strings and associated with the statistically-significant correlations observed earlier. In other words, we can develop a navigable classificatory infrastructure for data-mining and information-sharing that will allow us to reveal heretofore undiscovered knowledge relationships.

ACKNOWLEDGMENTS

This work originated in the Knowledge Space Lab, Royal Netherlands Academy of the Sciences). The KU Leuven number pairs were donated to the Knowledge Space Lab by Johan Rademakers and Bart Peeters. The OCLC MARC records were donated by the OCLC Office of Research. MARC records were downloaded and their characteristics transferred to IBM® SPSS® Statistics for processing by Hyoungjoo Park.

REFERENCES

Akdag Salah et al. 2012. The Evolution of Classification Systems: Ontogeny of the UDC by Almilah Akdag Salah, Cneg Gao, Kryzstof Suchecki, Andrea Scharnhorst and Richard P. Smiraglia. In A. Neelameghan and K.S. Raghavan eds. *Categories, contexts, and relations in knowledge organization: Proceedings of the Twelfth International ISKO Conference, 6-9 August 2012, Mysore, India.* Advances in knowledge organization 13. Würzburg: Ergon Verlag, pp. 51-57.

ASIST SIG/CR. 2014. Call for proposals: Universal Classification in the 21st Century, SIG/CR Classification Research Workshop http://sigcr.wordpress.com/2014/07/12/149/.

Green, Rebecca. 2014. Facet Detection Using WorldCat and WordNet. In Wiesław Babik, ed. *Knowledge Organization in the 21st Century: Between Historical Patterns and Future Prospects: Proceedings of the Thirteenth International ISKO Conference 19-22 May 2014, Kraków, Poland*. Advances in knowledge organization 14. Würzburg: Ergon Verlag, pp. 168-75.

Green, Rebecca and Michael Panzer. 2014. The interplay of big data, WorldCat and Dewey. In *Advances in classification research online* 24(1), doi:10.7152/acro.v24i1.14677.

Ridenour, Laura. 2014. Bridging gaps between domains. Unpublished seminar paper. University of Wisconsin, Milwaukee.

Schallier, Wouter. 2004. On the razor's edge: between local and overall needs in knowledge organization. In McIlwaine, Ia, ed., *Knowledge Organization and the Global Information Society: Proceedings of the Eighth International ISKO Conference, 13-16 July 2004 London, UK.* Advances in knowledge organization 9. Würzburg: Ergon Verlag, pp. 269-73.

Scharnhorst, Andrea et al. 2012. "Design vs. emergence: visualization of knowledge orders." http://scimaps.org/mapdetail/design_vs_emergence__127.

Smiraglia, Richard P. 1992. *Authority control and the extent of derivative bibliographic relationships.* Ph.D. dissertation. University of Chicago, 1992.

Smiraglia, Richard P. 2001. *The nature of "a work": implications for the organization of knowledge.* Lanham, Md.: Scarecrow.

Smiraglia, Richard P. 2014a. "Big Classification: Using the Empirical Power of Classification Interaction." In *Advances in classification research online* 24(1),doi:10.7152/acro.v24i1.14673

Smiraglia, Richard P. 2014b. Classification Interaction Demonstrated Empirically. In Wiesław Babik, ed. *Knowledge Organization in the 21st Century: Between Historical Patterns and Future Prospects: Proceedings of the Thirteenth International ISKO Conference 19-22 May 2014, Kraków, Poland*. Advances in knowledge organization 14. Würzburg: Ergon Verlag, pp. 176-83.

Smiraglia, Richard P. and Charles van den Heuvel. 2013. "Classifications and Concepts: Towards an Elementary Theory of Knowledge Interaction." *Journal of documentation* 69: 360-83.

Smiraglia, Richard P., Andrea Scharnhorst, Almila Akdag Salah and Cheng Gao. 2013. "UDC in action." In Slavic, Aïda, Almila Akdag Slah and Sylvie Davies eds., *Classification and Visualization: Interfaces to Knowledge, Proceedings of the International UDC Seminar, 24-25 October 2013, The Hague, The Netherlands.* Würzburg: Ergon Verlag, pp. 259-72.

Smiraglia, Richard P., Charles van den Heuvel and Thomas M. Dousa. 2011. Interactions Between Elementary Structures in Universes of Knowledge. In Slavic, Aïda and Civallero, Edgardo eds., *Classification & Ontology: Formal Approaches and Access to Knowledge: Proceedings of the International UDC Seminar 19-20 September 2011, The Hague, Netherlands.* Würzburg: Ergon Verlag, 2011, pp. 25-40.