Smiraglia, R. (2014). Big Classification: Using the Empirical Power of Classification Interaction.
Advances In Classification Research Online, 24(1). doi:10.7152/acro.v24i1.14673

Big Classification: Using the Empirical Power of Classification Interaction

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

ABSTRACT

Classification as a cultural artifact serves an epistemological role as disseminator of the culture it embodies. A theory of classification interaction has been proposed that would combine empirical iterations of bibliographic characteristics as factors interacting with traditional conceptual elements in classifications. Nine million UDC numbers extracted from the OCLC WorldCat are sampled and deconstructed, to look for correlations with content-designated components of the associated bibliographic records. Chi-squared is used to locate statistically-significant correlations among nominal-level variables. Results demonstrate a series of footprints of predictable associations. A complex network of associations is revealed and visualized. The results are promising and point to a series of more complex investigations.

Keywords

Classification interaction, cultural warrant, big classification, Universal Decimal Classification, empiricism.

I1.0 CLASSIFICATION AS ARTIFACT, CLASSIFICATION INTERACTION

Beghtol pointed out the cultural essence of what she called “knowledge organization classification systems,” which because of their origin, structure and use become, themselves, “cultural artifacts that directly reflect ... cultural concerns and contexts (2010, 1045). We often think of classifications as means for ordering either artifacts (i.e., documents) or their intellectual content for retrieval. But classification as a cultural artifact serves an epistemological role itself as disseminator of the culture it embodies. This happens in an inferential mode, or perhaps we should say in an as yet unrevealed mode, as given classifications and their faceted combinations are associated with the creative cultural characteristics of the artifacts they represent.

Smiraglia (1992) discovered weak associations between classes in the Library of Congress Classification and the tendency for classified works to be members of large instantiation networks. Table 1, extracted from the dissertation data, shows how the LCC class could be used both as a disciplinary icon and a weak predictor of instantiation (191, 197, 198).

LCC	**% Sample Works**	**Progenitor Works**	**Derived**	**Unique**	
			Frequency	Frequency	
?	1.2	5			
A			1	4	
B	10.5	43	21	10.2	n.s.
C	0.5	2	0	0	0.2
D	12.2	50	29	14.1	n.s.
E	2.9	12	7	3.4	n.s.
F	3.9	16	9	4.4	n.s.
G	1.2	5	3	1.5	n.s.
H	15.1	62	20	9.8	0.01
J	3.9	16	7	3.4	n.s.
K	1.5	6	4	2	n.s.
L	1.7	7	3	1.5	n.s.
N	14	3.4	6	2.9	n.s.
P	28.5	117	70	34.1	0.02
Q	6.1	25	12	5.9	n.s.
R	0.7	3	1	0.5	n.s.
S	0.2	1	0	0	n.s.
T	2.2	9	4	2	n.s.
U	2.2	9	4	2	n.s.
Z	2.2	9	4	2	n.s.

ISSN: 2324-9773

Table 1. LCC as discipline and predictor of instantiation

In this simple example, we see the distribution of the Georgetown University Library collection at a specific point in time (up to 1990 when the sample was drawn) across the LCC, as “titles” and as “works,” and the statistical significance of the class as predictor of derivation (or as we now would say, instantiation). Not shown here are further tests of predictive ability utilizing medium, form, genre, place and date of publication, language and LCC division.

Recently Smiraglia and van den Heuvel (2013) proposed a theory of knowledge organization that would combines empirical iterations of bibliographic characteristics as factors interacting with traditional conceptual elements in classifications. Smiraglia, van den Heuvel and Dousa (2011) demonstrated a faceted classification that would allow switching from conceptual to instantiation structures as a means of teasing out interactions between elementary structures of knowledge. In 2011 the Knowledge Space Lab or KSL (Scharnhorst et al. 2012) conducted a network analysis of the evolution of knowledge in Wikipedia. As a control mechanism, the Universal Decimal Classification was analyzed across its century of history to understand the evolution of academic knowledge as it is represented in the classification (Akdag Salah et al. 2012). The OCLC Office of Research provided the KSL team with a dataset of 9,055,623 UDC numbers extracted from 214,596,487 bibliographic records using the MARC 080 field in the WorldCat. A forthcoming paper (Smiraglia et al.) uses these UDC numbers to demonstrate the population of knowledge areas in the WorldCat as represented by the UDC.

The present project is an attempt to combine the big data analytical methods used by the KSL with those generated by Smiraglia (1992) for analyzing instantiation. The result of this preliminary research will show how data-mining might be used in classified datasets to discover structurally-related knowledge sets..

2.0 METHODOLOGY

A random sample of the OCLC records in the UDC number dataset was required to generate data about conceptual classification, instantiation, and bibliographic demographic characteristics. The 9,055,623 UDC numbers were recorded as paired with OCLC record numbers, thus:

3200	a517.53
9911	a519.24
19677	a7 CAST SAC
26863	a172
27725	a284.1(43)
29003	a551.510
33327	a(*3)
33789	a519.2:681.32]:622.1](082)
33789	a681.32:519.2]:622.1](082)

Table 2. Sample of WorldCat UDC OCLC record pairs

Sample size was calculated using the estimates of correlation among classification, bibliographic characteristics, and probability of instantiation contained in Smiraglia 1992. Estimating confidence of 95% and a confidence interval of ±5% the various calculations returned sample sizes ranging from 316 to 334. To allow for the potential that some of the selected OCLC records might be unavailable a sample of 400 number pairs was generated using Excel’s random number generator. Two records were no longer extant in the WorldCat. A sample of 398 records was searched in OCLC Connexion, and MARC text records were downloaded. A set of bibliographic characteristics was used to generate a spreadsheet for analysis using IBM-SPSS Statistics™.

Analyses ranged from a simple survey of bibliographic characteristics of the sampled records through cross-tabulation of all bibliographic characteristics with the elements of the deconstructed UDC numbers to look for statistically significant correlations. This research is preliminary, but the results are sufficiently promising to move on to more complex analyses. This will be discussed again in the conclusions of the paper.

3.0 PRELIMINARY RESULTS

3.1 The bibliographic population

The basic bibliographic shape of the sample is revealed with some simple metrics, shown in Table 3. Characteristics that could be expected to occur in every record (such as titles proper) were not recorded. We can state with 95% confidence that these proportions represent the population represented by our nine million plus UDC numbers from the WorldCat within a margin of errror of ±5%.

Characteristic	Yes	No
ISBN	47.7	51.8
Edition	22.9	77.1
Series	33.2	66.8
Bibliography	27.1	72.9
Subject name	8.3	91.7
Subject topic	42.7	57.3
Subject place	10.6	89.4
Index term	23.4	76.6
Genre form	15.1	84.9
Linked e-text	2.5	97.5

Table 3. Bibliographic characteristics of the sample

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. The subject indexing is split among various MARC fields; a summary shows that very few (8.3%) have named persons as subjects, less than half have topical subject headings, 10.6% have places named as subjects, about a quarter have uncontrolled index terms, and 15.1% have genre or form terms assigned.

The publication dates of the works represented are shown in Figure 1.

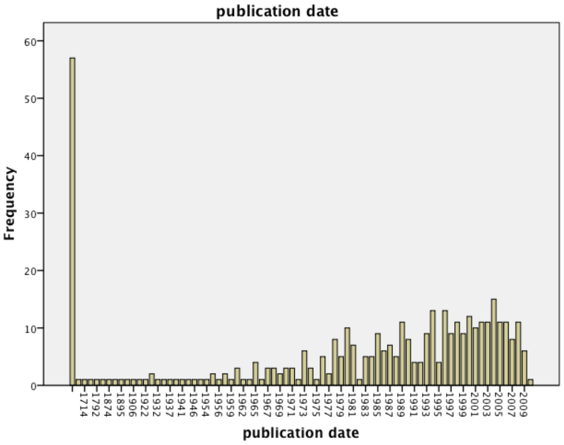

Figure 1. Dates of publication in the sample

Dates of publication ranged from 1606 to 2009, but 57 records had no date of publication. It is clear from the visualization that the majority of the works are dated after 1979. This is likely 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

3.2 CORRELATIONS: BIBLIOGRAPHIC CHARACTERISTICS

In the analyses that follow, nominal level data were cross-tabulated to look for statistically significant correlations. The Chi-squared (or X2) test of independence measures whether there is a statistically significant difference between observed and expected frequencies. It is used with nominal-level data (as we have here, most elements are either present “1” or not “0”). If there is no relationship between two variables the observed and expected frequencies should be about the same. When they are not the same the X2 computation yields a value of significance. The test requires large expected frequencies, when there are few data points in a cell in a contingency table, the statistic cannot be computed. In the present study, IBM-SPSS™ was used to construct contingency tables for all variables and to compute X2 for each cell. For each table a probability statistic is calculated; when that statistic is less than our stated confidence level of .05 then it suggests the co-occurrence of these variables is not due to chance. If one variable is present, the other is likely also to be present. There is no assumption of causation in these analyses, only confirmation of the likelihood of co-occurence. X2 is a very simple test for this reason but it is a good starting point.

	ISBN	Edition	Series	Bibliog.	Linked record
ISBN		■	■	■	
Edition	■		■		
Series	■	■			■
Bibliog.	■				
Linked record			■		

Table 4. Cross-tabulated bibliographic characteristics

characteristics were cross-tabulated and Table 4 contains the results of those tests. In this and the tables that follow, shaded cells represent those where co-occurrence was statistically significant.

The table shows that there are a few predictable associations. For instance, records with ISBNs are likely to have edition statements and bibliography notes, records with series statements are likely to have edition statements and be linked to an ebook. The same tests were performed on the subject indicators recorded. These are recorded in Table 5.

	name	topic	place	Index term	Genre Form
name					
topic			■	■	■
place		■			
Index term		■			■
Genre Form		■		■	

Table 5. Cross-tabulated subject indicators

Again there were a few predictable associations. Topical subject headings were associated with the presence of places, and genre terms as well as with the use of uncontrolled index terms. Uncontrolled index terms and genre terms were associated, but not names or places. The two sets were then cross-tabulated.

	ISBN	Edition	Series	Bibliog	Linked record
name					
topic	■			■	
place	■		■		
Index term	■	■	■	■	
Genre Form	■	■			

Table 6. Cross-tabulated subjects with bibliographic characteristics

There were few statistically significant associations. Topical subject headings occurred in conjunction with bibliography statements, uncontrolled index terms occurred in conjunction with edition and series statements, ISBNs were associated with topical subject headings, places, uncontrolled index terms and genre terms but not names. Edition statements were associated with uncontrolled index terms and genre terms, and series statements were associated with place names and uncontrolled index terms.

The next step was to convert the dates of publication to a variable called "age" by subtracting the date of publication from 2011 (the year the population was extracted from WorldCat). The mean age of work was 23 years, and the median age was 15 years, once again indicating the relative currency of the works in the population. Linear regression statistics were calculated. There was a weak influence of age on the presence of an edition statement, but no statistically significant influence on any other variable. (Interestingly, this result is consistent with studies of instantiation.)

3.3 DISTRIBUTION OR POPULATION OF THE UDC AND CORRELATIONS AMONG OPERATORS

The UDC numbers were parsed next. The population of the main classes was evaluated, and the results appear in Figure 2.

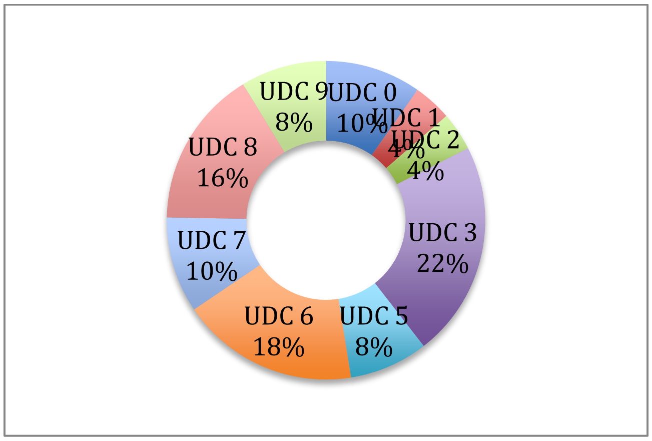

Figure 2. Population of the main UDC classes

Not surprisingly the distribution matches the KSL team's earlier analysis of the full population of numbers. 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.

When the three connectors were cross-tabulated with UDC main classes. The results show a few statistically-significant correlations, shown in table 7.

	"+"	":"	"/"
0		🟦	
1			
2			
3			🟦
5			
6	🟦	🟦	🟦
7			🟦
8			
9	🟦		🟦

Table 7. Common auxiliary signs cross-tabulated with UDC main classes

The colon operator (simple relations) co-occurred with classes 0 and 6, the plus operator (addition) with classes 6 and 9, and the slash operator (consecutive extension) with classes 3, 6, 7 and 9. Class 6 is the most likely to co-occur with all three operators.

The linking classes also were recorded (those found to the right of auxiliary operators). These appear in Table 8.

	Percent
0	.08
1	.09
2	.04
3	.08
5	.03
6	.04
7	.22
8	.17
9	.22
total	.97

Table 8. UDC classes linked with common auxiliary signs

84% of the UDC numbers have no common auxiliary associated with them, indicating the usage is relatively rare in the dataset. However, all of the main classes appear following linking auxiliaries, most in classes 7 "Arts,

UDC Operator	ISBN	Edition	Series	Bibliog.	Linked record	name	topic	place	Index term	Genre Form
0-9 main UDC class	🟦	🟦				🟦		🟦	🟦	🟦
“+”							🟦			
“:”									🟦	
“/”								🟦		
Linked class										

Table 9. Cross-tabulation of UDC operators with bibliographic characteristics

UDC class	ISBN	Edition	Series	Bibliog.	Linked record	name	topic	place	Index term	Genre Form	“+”	“:”	“/”	Linked class
0	🟦								🟦	🟦		🟦		🟦
1						🟦								
2														
3							🟦			🟦				
5														
6	🟦					🟦		🟦			🟦			🟦
7			🟦			🟦							🟦	🟦
8	🟦	🟦		🟦			🟦	🟦	🟦				🟦	
9								🟦						

Table 10. Cross-tabulation of UDC main classes with operators and bibliographic characteristics

UDC class	ISBN	Edition	Series	Bibliog.	Linked record	name	topic	place	Index term	Genre Form	“+”	“:”	“/”	Linked class
0	🟦								🟦	🟦		🟦	🟪	🟦
1						🟦								
2														
3							🟦			🟦				
5	🟪													
6	🟦					🟦		🟦			🟦		🟪	🟦
7	🟪		🟦			🟦							🟦	🟦
8	🟦	🟦		🟦			🟦	🟦	🟦				🟦	
9								🟦	🟪					

Table 11. Cells that would have been statistically significant at 90% confidence.

Entertainment, Sports,” 8 “Language, Literature,” and 9 “Geography, Biography, History.” These results are consistent with network visualizations in earlier analyses. Data on usage and linkage of common auxiliary linkages were also collected but are not analyzed here. A cross-tabulation of UDC main classes with linked classes was also undertaken. Results were not statistically significant.

3.4 CORRELATIONS: BIBLIOGRAPHIC CHARACTERISTICS WITH CLASSIFICATION OPERATORS

The next step was to cross-tabulate all of the bibliographic characteristics with all of the UDC operators. A few statistically significant correlations were identified (see Table 9).

Individual UDC main classes correlate with several of the bibliographic characteristics, especially with ISBNs, edition statements and uncontrolled index terms. Topical subject headings were associated with the addition operator, uncontrolled index terms with the relations operator, and place names with the consecutive extension operator. None were associated with UDC numbers that included a linked class.

For further analysis, individual variables were created for each UDC main class 0-9 and each was cross-tabulated with all characteristics and operators. Results are shown in Table 10.

The most highly populated classes (6 and 8) have the most correlations. There are a few predictable footprints, such as class 8 being associated with richer subject vocabulary, but class 0 being associated with uncontrolled index terms (i.e., more specific) and genre terms. Classes 0, 6 and 7 are associated with linked classes. Classes 1, 6, and 7 are associated with names as subject headings, but classes 8 and 9 are more associated with places. Class 0 is associated with the “:” simple relation operator, but classes 7 and 8 are associated with the “/” consecutive extension operator.

4.0 DISCUSSION

This research represents perhaps the simplest first step possible, by attempting to find statistically-significant correlations among easily content-designated components of the bibliographic universe and deconstructed elements of UDC classification numbers. The results are affirmative to the extent that statistically-significant associations exist at many levels. In many cases they make sense intuitively. For example, it seems logical that series statements and edition statements and bibliography statements occur together, as well as that they all are common on works with ISBNs. Books with linked electronic records are fairly new, so it makes sense that there are few of them and also that they might be associated with works published in series. It makes sense that genre terms should be more closely allied with uncontrolled index terms than with traditional subject headings, because of what we know about cataloging practices. It also seems logical that names as subjects are not associated with other kinds of subject headings; typically biographical works might not have many other headings assigned to them. Similarly, names as subject headings are not associated with the other detailed bibliographic characteristics. What we know about the associations of UDC classes with auxiliary operators or other classes also makes sense, given the age of the population of classified works and the population of the classification in general.

When we looked for associations between the UDC components and the bibliographic characteristics we also found logical associations occurring as statistically significant. Works in class 8 “Languages and literatures” seem to be associated with richer subject vocabulary, represented here by a more complex mix of subject headings as well as the simple relation operator (to express phase relations) and the consecutive extension operator (to express broader subject areas). Class 0 “Generalities” contains many reference works, thus it is associated with uncontrolled index terms that likely are more specific, genre terms, the simple relation operator and linked classes. Classes 6 and 7 are the sciences and the arts, and thus are associated with linked classes, as well as names as subject headings and the consecutive extension operator. Class 1 “Philosophy and psychology” is associated with names as subjects.

All of the associations enumerated here could be considered footprints of sorts, which is to say they represent ways in which associations among classified bibliographic entities are reliably present and predictable. The associations themselves could be classified. In any case, these associations demonstrate the power of classification for indicating nonlinear pathways through a collection of bibliographic entities.

An artifact of this study (or it could be termed a limitation) is the confidence level of 95%, which therefore rendered other likely associations not statistically significant. If we were to reprise table 10 here and this time shade the cells that would have been statistically significant at 90% confidence we get the result shown in table 11.

Now we have a slightly richer picture, with a few more sensible associations. This suggests that a slightly larger sample might have yielded this many more statistically-significant correlations, and therefore, predictable, navigable footprints. Also, it is important to consider what we learn from all of the blank cells where no associations were apparent. This tells us that those characteristics are ubiquitous and therefore predictable because they are always present across all classes to some degree.

5.0 CONCLUSIONS: EMPIRICAL POWER OF CLASSIFICATION INTERACTION

With this simple preliminary investigation we have seen a first glimpse of the empirical power of classification interaction. The shape of the correlations we observed aligns with the population of the UDC across the bibliographic domain represented. Predictable footprints demonstrated here could be used as pathways for navigation across the structural relationships shared by works in differing domains, in addition to traditional author and title navigation of known item searching or gatherings of like subjects in traditional classification. In this sense the UDC is shown to represent Beghtol’s notion of a classification as cultural artifact, because it both gathers works together and facilitates navigation across their bibliographic characteristics. In other words, the classification here is a footprint, of sorts, of the works it classifies.

At another level we can consider all of the results of this study to be hypotheses for further research. A limitation of this study is the post-1979 chronology of the UDC numbers assigned to works in the WorldCat. The majority of the nine million UDC numbers in the sample under analysis represent works cataloged in the last quarter century. This means that the UDC practices and the population of the classification revealed here does not represent the whole history of the UDC, but just a recent snapshot. The Knowledge Space Lab team also received output of all UDC numbers from the online catalog of Catholic University Leuven (KU Leuven) in order to compare the population of the UDC in a specific library to that found in the WorldCat. An obvious next step for this research is to repeat the analyses using the data from KU Leuven for comparison.

In this study only content-designated elements were analyzed for correlation, and not any of the values of any of those elements. In other words, another obvious next step is to look at individual names, works, subject terms, and so forth, to see what more detailed network of associations might be present in the classified bibliographic universe, and whether those associations are predictable. The technique of identifying statistically-significant correlations among nominal level variables also has the limitation that no direction or causality can be demonstrated. All we can say at this point is that we see certain elements occurring in conjunction with each other in this data set.

However, it is already clear from this simple study that in a classified bibliographic dataset predictable pathways of association exist through the data. The raw data associated with the elements in table 11 were entered into a matrix to generate a network diagram of the associations uncovered in this research. Figure 3 shows the network.

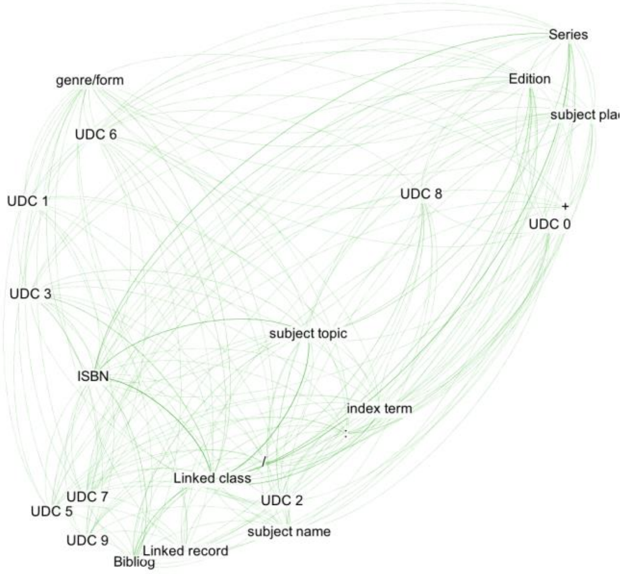

Figure 3. Network diagram of classification interactions visualized in Gephi 0.8.2

The visualization in figure 3 demonstrates the extent to which the classification is an integral cultural artifact of the environment it represents. The classes and auxiliary operators are not only interlinked with each other, but they also are interlinked with the bibliographic characteristics of the collection in diverse and predictable ways.

ACKNOWLEDGMENTS

The OCLC WorldCat UDC number extraction was provided by Ed O'Neill of the OCLC Office of Research. The distribution of the nine million number pairs across nineteen Excel spreadsheets was accomplished with a program created by Brian Thompson. The OCLC MARC records were downloaded and their characteristics transferred to SPSS for processing by Hyoungjoo Park and Ann Graf. Laura Ridenour was immensely helpful in constructing the network graph.

REFERENCES

Akdag Salah, A.,Gao, C.,Suchecki, K., Scharnhorst, A. * Smiraglia, R.P. (2012). The Evolution of Classification Systems: Ontogeny of the UDC. 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, 2012, pp. 51-57.

Beghtol, C. (2010). Classification Theory. *Encyclopedia of Library and Information Sciences*, Third Edition, 1: 1, 1045 — 1060

Scharnhorst, A., Akdag Salah, A.,Gao, C.,Suchecki, K. (2012). Design vs. emergence: visualization of knowledge orders. http://scimaps.org/maps/map/design_vs_emergence__127/

Smiraglia, R.P. (1992). Authority Control and the Extent of Derivative Bibliographic Relationships. Ph.D. dissertation. University of Chicago.

Smiraglia, R.P., van den Heuvel, C. & Dousa, T.M. (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, pp. 25-40.

Smiraglia, R.P. & van den Heuvel, C. (2013). Classifications and concepts: towards an elementary theory of knowledge interaction. Journal of documentation 69: 360-83.

Smiraglia, R.P., Scharnhorst, A.,Akdag Salah, A. & Gao, C. (2013). UDC in action. Forthcoming in the UDCC proceedings, October 2013, The Hague.