Richard P. Smiraglia, Rick Szostak

Converting UDC to BCC: comparative approaches to interdisciplinarity

Abstract

The knowledge organization domain has been turning its attention increasingly to problems of interdisciplinarity. Recently we have attempted to explore the approaches to interdisciplinarity represented by the synthetic and faceted Universal Decimal Classification (UDC) and that of the phenomenon-based Basic Concepts Classification (BCC). The questions for research are: how do both classifications express the same sets of concepts, what are the specific advantages or disadvantages of disciplinary versus phenomenon-based classification in the gathering of concepts, and how can these classifications be used to generate interdisciplinary ontologies for the Semantic Web? The study reported here takes an empirical approach to the comparison of the UDC and the BCC assigned to a set of documents found in the OCLC WorldCat. The present study demonstrates both the greater economy and greater conceptual precision in the phenomenon-based BCC. The network analysis suggests that there is great navigational strength in both approaches.

1. Interdisciplinarity, phenomena, concepts and KOSs

The knowledge organization (KO) domain has been turning its attention increasingly to problems of interdisciplinarity, both in the sense that the domain is working to create knowledge organization systems (KOSs) that bridge disciplinary boundaries, and in the sense that the domain has begun to seek truly interdisciplinary knowledge organization solutions. One powerful approach to interdisciplinarity increasingly demonstrated is the structure of classifications around individual phenomena rather than in disciplinary groupings. The obvious advantage of phenomenon-based systems for interdisciplinarity is gathering by phenomenon despite disciplinary or epistemic stance.

Recently we have attempted to explore the approaches to interdisciplinarity represented by the synthetic and faceted Universal Decimal Classification (UDC) and that of the phenomenon-based Basic Concepts Classification. Both are large-scale general classifications capable of expressing complex concepts with precision and subtlety. The questions for research are: how do both classifications express the same sets of concepts, what are the specific advantages or disadvantages of disciplinary versus phenomenon-based classification in the gathering of concepts, and how can these classifications be used to generate interdisciplinary ontologies for the Semantic Web? The first two questions have been examined briefly in an exploratory study that compared classification strings assigned by both classifications to the same set of documents (Szostak and Smiraglia 2017) and mapped the network relationships among facets in both sets of strings (Smiraglia and Szostak 2017). The third question is one trajectory of a large-scale research project connecting KOSs and the LOD Cloud

(Szostak, Scharnhorst, Beek and Smiraglia 2017) [1]. The study reported here takes an empirical approach to the comparison of the UDC and the BCC assigned to a set of documents found in the OCLC WorldCat.

We do not need to rehearse the origins of the Universal Decimal Classification here (see Szostak and Smiraglia 2017), except to recall that its founder, Paul Otlet, had as his goal the specific ordering of concepts extracted from deconstructed texts. Otlet generated, then, a classification of knowledge that is commonly used to assign many UDC strings to any particular document, in order to precisely identify topical phenomena. Empirical research has demonstrated the power of this approach to classification using the UDC (Smiraglia 2016; Scharnhorst *et al.* 2016), one aspect of which is its generation of a network linking phenomena within the classified set of documents represented by UDC strings.

The Basic Concepts Classification (BCC)[2] was created by Rick Szostak with the explicit goal of providing a means to classify documents (and objects and ideas) with respect to the phenomena they study. The BCC has evolved through the growth of schedules of (mostly verb-like) relators and adjectival/adverbial properties added to the original schedule of phenomena. Documents (objects, ideas, concepts) can be classified with combinations of phenomena, relators and properties. Szostak (2016; 2017) suggests subject classifications should follow basic grammatical structures in combining these three types of term; such subject classifications will thus appeal to the linguistic facility of both classifiers and users.

2. Methodology

We compiled a set of documents classified using the UDC taken from a random sample drawn from the OCLC WorldCat. Nine million UDC strings from the WorldCat, representing essentially a dump of the entire population of UDC in the WorldCat at one point in time, were originally downloaded by the OCLC Office of Research (Scharnhorst *et al.* 2016). These are representative of classification provided by mostly European UDC libraries using the WorldCat and assigning UDC to mostly scientific and technical late twentieth century works. Through pilot studies and analysis of earlier empirical studies it was determined that a sample drawn at random would need 381 cases to provide results generalizable at 95% confidence ±5%.[3] We deciphered each

[1] Digging Into the Knowledge Graph, 2016 Digging Into Data Challenge. Available at: https://diggingintodata.org/awards/2016/project/digging-knowledge-graph/

[2] https://sites.google.com/a/ualberta.ca/rick-szostak/research/basic-concepts-classification-web-version-2013.

[3] This sample size formula based on a proportion with a known population size is:

$$n = \frac{z^2Np(1-p)}{NE^2 + z^2p(1-p)}$$

where

n = sample size; z = curve value for the confidence interval (.95); N = items in the sampling

UDC string, deconstructing each element of it and citing an example from the WorldCat of a resource to which the particular string had been assigned. We then provided BCC classification to match each UDC string. In Szostak and Smiraglia (2017) we included cases from Portuguese sources (see Scharnhorst *et al.* 2016) to demonstrate the ability of BCC to capture the entire content and context in a single string. For each case one BCC string was assigned, but the range of number of UDC strings was 1-7; the ratio was 1 to a mean of 2.2, for a multiplier of .45. In the present study we are comparing UDC strings drawn from the WorldCat rather than classification assigned to particular resources. Thus, the data reported here do not further demonstrate this aspect of the relationship between the expressivity of the two classifications. Our emphasis here is on conceptual expression and network relationships.

3. Results

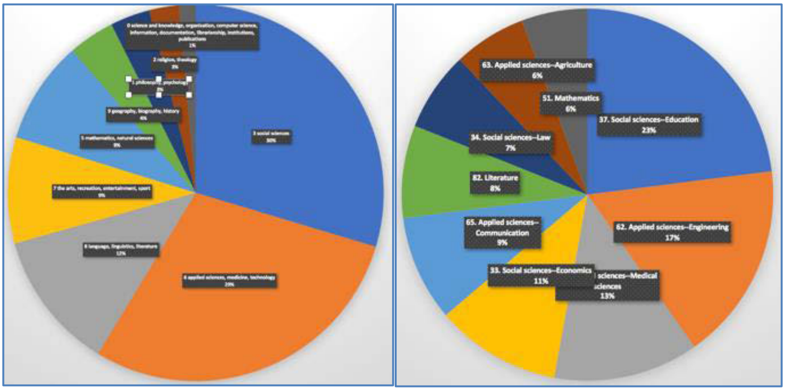
We were able to decipher 382 UDC strings using UDC Online (http://udc-hub.com/) and Attila Piri's UDC-parser (http://piros.udc-interpreter.hu/#). Thus our data yielded results that describe well the UDC-classified content of the WorldCat. We recorded data on ontogenetic questions to be used in a future paper. Our analysis here visits a) the population of both UDC and BCC concepts; b) string length and number of terms; c) comparative conceptual precision; and d) network analysis.

3.1. Population of the UDC and BCC

One informative metric about a classification can be the degree to which its classes are populated in any given classified environment. Classifications like the UDC are considered general classifications to the extent that they are thought to be hospitable to all major disciplines of knowledge. By visualizing the population of the UDC main classes in our study we create a contextual picture of the degree to which various disciplines are heavily used or, alternatively, little used. Figure 1 shows the population of the UDC main classes and two-digit disciplines derived from the sample used in this study.

frame(9,000,000): p = expected proportion (.45) ratio of BCC string size to UDC string size in Szostak and Smiraglia (2017a); E = tolerable error (.05).

Figure 1: Population of UDC main classes; most populous disciplines



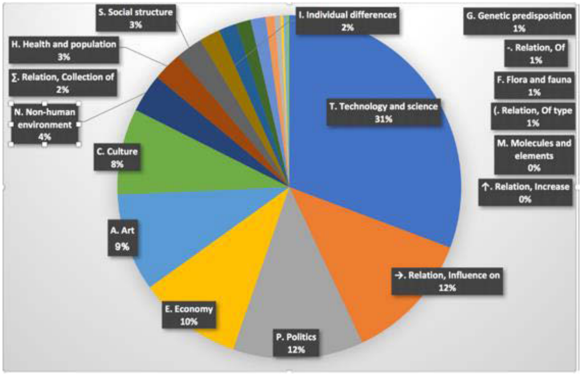
The distribution is consistent with earlier visualizations of the population of the main classes in the 2009 WorldCat UDC dump as summarized in Scharnhorst *et al.* (2016) and Smiraglia (2016), and this consistency helps to demonstrate the generalizability of the data in our sample. The largest disciplinary clusters and hence the majority of the works (approximately 30% each) fall in the social sciences and the applied sciences, and another 30% comprises natural sciences, literature and the arts. There were sixty-four disciplinary combinations of two digits in the sample. From this visualization we can see that the social sciences are predominantly populated by education, economics and law, that the applied sciences comprise medicine, engineering, communication and agriculture, and the other large disciplines represented are mathematics and literature.

The BCC is phenomenon-based rather than discipline-based, so that analysis of the first character of each string would point to classes of phenomena. Eighteen characters were found in first position. The list, shown in Table 1, is remarkably different from the basic disciplinary classes we saw in UDC.

Table 1: BCC lead characters

T. Technology and science	110	N. Non-human environment	13	F. Flora and fauna	5
→. Relation, Influence on	44	H. Health and population	10	G. Genetic predisposition	3
P. Politics	44	S. Social structure	9	-. Relation, Of	2
E. Economy	35	I. Individual differences	7	↑. Relation, Increase	1
A. Art	33	∑. Relation, Collection of	6	M. Molecules and elements	1
C. Culture	29	(. Relation, Of type	5	X. Mathematical concepts	1

Figure 2: Population of BCC lead terms



Population of these classes is shown proportionally in Figure 2 for comparison with the population of the UDC classes. (Note: One reason for the large number of entries beginning with T is that different types of document, such as dictionary or textbook, are captured within class T.) This visualization gives us a remarkably different, and perhaps fuller picture of the conceptual content of the classified collection. The largest cluster of 31% for Technology and science mirrors the largest cluster in the UDC. Politics, economy, art, culture and non-human environments are all clearly articulated in clusters ranging from 4% to 12% of the total. But causal relators, which are always affixed as auxiliaries in UDC lead in many cases in BCC, with “influence on” constituting the second largest cluster equal in size to that for politics. The vague disciplinary clustering is replaced with a more vivid description of the classified collection.

3.2. UDC string length and number of terms

The mean UDC string length is 7.7 characters and 1.5 terms; the median is 6 characters and 1 term; the mode is 3 characters and 1 term. Sorted by combination of string length and number of terms we discovered there were 53 different combinations. One-character strings (*e.g.*, “3” Social sciences) occur only 3 times and tend to be assigned to nonbook materials. Three-character single-term strings occur 59 times (17%), which is the largest frequency of occurrence. Five-character one-term strings occur 42 times, the second highest frequency (12%). Longer strings having 14-22 characters always have two or more terms. Strings with 23 characters always have three or more terms, and the longest strings (more than 24 characters) always have four or more terms. Strings combining two terms are fairly common; more complex combinations are rare.

3.3. BCC string length and number of terms

The mean BCC string length is 10.5 characters; the median is 9 and the mode is 7 characters. The mean number of terms per string is 3.5; the median and mode are 3; the number of terms ranges from 1 to 9. (Where the classification indicates “Cutter number” we have counted it as one term with two characters, *e.g.*, an alphanumeric “B3.”) There is greater disparity among string lengths; the largest occurring frequency is 7 characters, which occurs 42 times (20%); the second most frequently occurring is 12 characters, which occurs 34 times (16%). The range in character length is from 2 to 29 characters. The length hovers around the mean; extremes (25 or more characters, but oddly also 2 or 5 characters) are rare. The most frequently occurring combination is 7 characters and 3 terms (7%), followed by 3 characters and 1 term (5%). Thirty-five strings (36%) ranging from 3 to 29 characters in length, have only 1 term, which is a further reflection of the specificity of the phenomenon-based classification.

3.4. Comparative conceptual precision in UDC and BCC

Analysis of the text of the deconstructed strings reveals several large conceptual clusters in the data. For the purpose of this paper we isolated two, “military affairs” and “economic outputs” in order to compare the classified strings. Table 2 shows the comparative results.

Table 2: Concept clusters “Military Affairs” and “Economic Outputs”

Military Affairs	Economic Outputs
UN peacekeeping doctrine 355.4 3 Social science—35 Public administration—355 Military affairs—355.4 **War operations** TF(P15a)^→gcrx TF professional field—TF(P15a) **military field—^→gcrx associated with ending conflict**	**marine pollution and sea life** 591.9(26) 5 Natural sciences—59 Zoology—591 General zoology—591.9 Geographic distribution of animals—(26) common auxiliary of place 2 Physiographic designation—26 **Oceans, seas and interconnections** (EO →*ga* QN) → F>NT3o (**Pollution** = economic output does not facilitate the quality of the environment)(affects)(life)(in)(oceans)
Atlantic convoys 355/359 3 Social science—35 Public administration—355 Military affairs—/359 consecutive extension to **Armed forces** TF(P15a) ∑EO925111^PI5n>NT3oa TF(PI5a) Military science—∑EO925111 > NT3oa **Convoy** (collection of ships—NT3oa is **Atlantic Ocean**—navy PI5n: ∑EO925111^PI5n > NT3oa	**Mary Cassatt prints** 76 7 Arts—76 Graphic art, **printmaking** EO9821215 ^ A (economic output of **prints**)(associated with)(**art**)
Missile defense 355.45(075.8) 3 Social science—35 Public administration—355 Military affairs—355.45 **Defence** of the realm—common auxiliary of form: Texts for university, higher education PI5a→ga E094612 **(military)(prevents)(missiles)**	**Confucianism impact on modernization** 22 2 Religion—21/29 Religious systems—22 **Religions originating in the Far East** CR1d→EO↑ **(Confucianism)(influences)(economic output)(growth)**

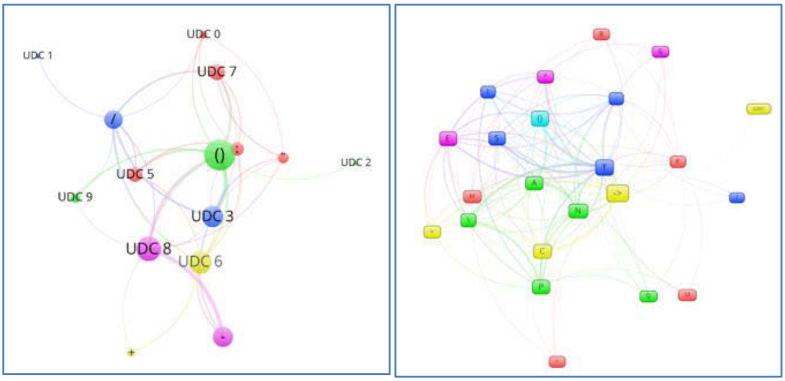
We have used bold type as a simple mechanism to illustrate the presence or proximity of the concept in each case. In most cases the concept itself is not explicit in the UDC string, which is used to create discipline-based gatherings. In all cases the concept is explicit in the BCC string, and in all but one case it is the first element in the string. Note especially the clarity of the sentence-like BCC strings. The phenomenon-based gatherings, then, are more precisely conceptual.

4. Network analysis

Network analysis allows us to discover and visualize nodes representing main classes, auxiliaries, and connectors in UDC classified strings, and nodes representing classes and relations in BCC. This technique of analysis, which is explained fully in Smiraglia *et al.* (2013), involves construction of matrices using the quantity of co-occurring connectors; the matrices can then be used to create both multi-dimensionally-scaled plots to demonstrate co-occurrence and network diagrams to show navigable pathways and their relative strengths.

Among the UDC classified strings, 209 strings (or 54%) represented single term expressions using no auxiliaries or connectors. The majority of these strings occurred in class 3 “social sciences” (35%) and class 6 “applied sciences” (32%). Class 5, 7, 1 and 2 occupied another 27%. Analysis of two-digit disciplinary classes shows the majority of strings occurred in 37 “education,” 62 “engineering,” 33 “economics” and 61 “medical sciences.” Of course, underlying these strings is the hierarchical network of disciplines divided by class, division and subdivision, creating a tree-like structure. The navigability of a classification’s syndetic structure is well-understood, and thus we have not looked further into that in the present study. On the other hand, 144 of the UDC classified strings (37%) were combinatorial strings using main classes, auxiliaries and connectors. These strings fell mostly in classes 8 “language, linguistics, literature” (25%), 6 “applied sciences, medicine, technology” (24%). 3 “social sciences” (22%) and 7 “the arts, recreation, entertainment, sport” (11%). Analysis by two-character disciplinary codes showed that the largest clusters were in 38 “education” (23%), 82 “literature” (18%), 62 “engineering” (13%), 83 “German literature cancelled number” (13%) and 61 “medical sciences” (11%), with smaller clusters in chemistry, sport and philology. Thus, there is a small but important difference between the two groups of classified strings such that those using UDC’s faceted synthetic structure tend to be more associated with technology, education and literature. For example, in the UDC string “808.2(075.3),” main class “8 Language. Linguistics. Literature” is linked to “(0) Special auxiliaries of form.” A network map is a way of visualizing these linkages as they occur in a set of classified strings, such as the sample in our study. For these matrices were constructed and used to create a visualization of the networks underlying the combinatorial structure, shown in Figure 3.

Figure 3: Network maps of nodes in UDC and BCC strings (created using VOSViewer 1.6.5)



The size of the nodes indicates the strength of the linkage. For example, classes "8" "6" "3" and "9" most often link to special auxiliaries of form or language, classes "3" "0" and "7" have most of the chronological form linkages, classes "8" and to a lesser extent "5" "6" and "7" link to simple relations. Classes "1" "7" "5" and "9" have most of the consecutive extensions, and classes "6" and "8" are most linked through addition to each other. Classes "2" and "1" have few linkages and thus appear at some distance from the core of the network.

Among the BCC strings, sixty represented single term expressions with no relators or synthesis. These fell mostly in classes T "technology and science," P "politics," C "culture," E "economy," and A "art." There were 301 combinatorial strings, for which a matrix was created. Figure 6 shows the network map of the combinatorial structure among the BCC strings in this study. We can see first that there is a much denser network, and also that many nodes are represented by relators; this is a reflection of the grammar-like structure of the BCC in implementation. Mirroring what we saw in Figure 4 above, classes T, P and E are the most influential (reflecting the content of the sample). The strength of the connections is reflected in the density of the curved lines; for example, classes E "economics" and T "technology and science" is heavily connected to each other as well as to class S "social structure," reflecting in this case the large number of resources relevant to kinds of workers in economics and the sciences.

5. Conclusions

This is the first large-scale empirical analysis of a phenomenon-based classification applied to a set of resources in parallel with the discipline-based UDC. Similar to the results reported in our preliminary study (Szostak and Smiraglia 2017, 13), this analysis demonstrates the greater economy of the BCC: "The present study demonstrates the greater economy provided by the phenomenon-based BCC classification, which combines conceptual semantic representations in precise relator-defined syntactic strings". The present study adds to that conclusion the demonstration of greater conceptual precision in the phenomenon-based BCC. The network analysis suggests that there is great navigational strength in both approaches.

Acknowledgments

We are grateful for the assistance of Hyoungjoo Park, Xin Cai and Vanessa Schlais in drawing the sample. We are in great debt to Aida Slavic, Editor-in-Chief of the UDC for assistance with cancelled sections of the UDC.

References

Scharnhorst, Andrea, Richard P. Smiraglia, Christophe Guéret and Alkim Almila Akdag Salah (2016). Knowledge Maps of the UDC: Uses and Use Cases. *Knowledge Organization* 43: 641-654.

Smiraglia, Richard P. (2016). Empirical Methods for Knowledge Evolution across Knowledge Organization Systems. *Knowledge Organization* 43: 351-357.

Richard P. Smiraglia and Rick Szostak (2017). Comparative Approaches to Facets in Interdisciplinary KOSs: UDC and BCC. Poster abstract in *Proceedings of the International UDC Seminar 2017, London (UK), 14-15 Sept.2017*, ed. Aida Slavic and Claudio Gnoli. Würzburg: Ergon Verlag, 279-284.

Smiraglia, Richard P., Andrea Scharnhorst, Almila Akdag Salah and Cheng Gao (2013). UDC in Action. In *Classification and Visualization: Interfaces to Knowledge, Proceedings of the International UDC Seminar, 24-25 October 2013, The Hague, The Netherlands*, ed. Aïda Slavic, Almila Akdag Salah and Sylvie Davies. Würzburg: Ergon-Verlag, 259-272.

Szostak, Rick (2016). The Simplest Approach to Subject Classification," *Proceedings of the IFLA satellite conference*, Columbus OH, Aug. 2016.

Szostak, Rick (2017). "Facet analysis without facet indicators" In *Dimensions of Knowledge: Facets for Knowledge Organization*, ed. Richard P. Smiraglia and Hur-li Lee. Würzburg: Ergon Verlag, 69-85.

Szostak, Rick and Richard P. Smiraglia (2017). Comparative Approaches to Interdisciplinary KOSs: Use Cases of Converting UDC to BCC. In *Proceedings of the Fifth North American Symposium on Knowledge Organization, Champaign-Urbana, Illinois, June 15-16, 2017.* Available at: http://www.iskocus.org/NASKO2017papers/NASKO2017_paper_3.pdf.

Szostak, Rick, Andrea Scharnhorst, Wouter Beek and Richard P. Smiraglia (2017). Connecting KOSs and the LOD Cloud. Submitted to ISKO 2018 Porto, Portugal.