Rick Szostak, University of Alberta

Richard P.Smiraglia, University of Wisconsin, Milwaukee

Comparative Approaches to Interdisciplinary KOSs: Use Cases of Converting UDC to BCC

Abstract: We take a small sample of works and compare how these are classified within both the Universal Decimal Classification and the Basic concepts Classification. We examine notational length, expressivity, network effects, and the number of subject strings. One key finding is that BCC typically synthesizes many more terms than UDC in classifying a particular document – but the length of classificatory notations is roughly equivalent for the two KOSs. BCC captures documents with fewer subject strings (generally one) but these are more complex.

1.0 Interdisciplinarity, phenomena and two classifications

Interdisciplinarity is an important new approach to knowledge organization seeking to provide useful clustering of knowledge concerning particular phenomena that might otherwise be scattered by discipline. While gathering by discipline provides certain epistemic assurances concerning the treatment of phenomena, scattering by discipline can prevent phenomenon-based knowledge discovery. In this paper we report an exploratory study in which we seek to compare the approach to interdisciplinarity provided by the Universal Decimal Classification's synthesis and faceted auxiliaries to that provided by the Basic Concepts Classification, which is a phenomenon-based interdisciplinary general classification.

The origins of the Universal Decimal Classification (UDC) can be found in the vision of Belgian documentalist Paul Otlet, who was seeking a mechanism by which to index (and therefore non-semantically gather) specific topical content of documents. Rather than classifying an entire document by placing it in a summary disciplinary class, Otlet wanted to provide specific ordered indexing by concept. Otlet generated a classification utilizing a decimal system based on the basic structure of Melville Dewey's 1876 *Decimal Classification.* First published in 1905, the Universal Decimal Classification (International Federation for Documentation 1905) evolved such that it often is described as the only worldwide multilingual, multicultural, knowledge classification. Although the UDC is used for library classification, it is not used primarily to gather documents at a summary level for browsing, in the manner of the *DDC.* Rather, UDC is a classification of knowledge; it is commonplace, then for libraries to assign many UDC strings to the bibliographic record for each document, in order to precisely identify topical phenomena. Recent research has shown the facile capability represented by this usage of UDC (Smiraglia 2016a-b, Scharnhorst et al. 2016), demonstrating the presence of a network linking phenomena within the classified set of documents represented by bibliographic records bearing UDC strings.

The BCC has been developed by Rick Szostak over the last decade. It seeks explicitly to classify documents (and objects and ideas) with respect to the phenomena they study. As the BCC has been developed, Szostak has added schedules of (mostly verb-like) relators and adjectival/adverbial properties to the original schedule of phenomena. Documents can be classified with combinations of phenomena, relators and properties. In recent papers Szostak has advocated that 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.

2.0 Methodology

The purpose of this exploratory descriptive study was to discover similarities and differences in the approaches taken by the two classifications when applied to a small set of documents. A central research question is, if a network of phenomena underlies the assigned elements of UDC strings in a collection of documents, can a more direct interdisciplinary classification provide a shorter path between any two points?

To explore answers to this question we created a set of use cases by selecting 25 cases from among UDC classified bibliographic records derived from the research by Scharnhorst et al. (2016). Five cases were selected from bibliographic records made available from the National Library of Portugal because of their complexity and use of multiple UDC strings; in the earlier study it was discovered the mean number of UDC strings per record was 2.8, with a range from 1-11. To this was added a small random sample drawn from the OCLC WorldCat. These latter are representative of mostly European UDC libraries using the WorldCat and assigning UDC to mostly scientific and technical late twentieth century works. Our sample is too small to allow generalization; nevertheless because our study is exploratory we believe the results are indicative.

3.0 Results

Table 1. Comparative UDC and BCC assignments

Case#	**Identifier**	**UDC**	**Length (Terms)**	**BCC**	**Length (Terms)**
1	Health and ethics in Portugal poster	613.8(469)(084.5)	17 (3)	H+CV2b>N1cpt^AN7 [Health (H) and (+) Ethics (CV2b) in (>) Portugal (N1cpt) associated with (^) Poster (AN7)]	1 (7)
2	Photographic poster	77.03(084.5)	12 (2)	AR3>AN7	7 (3)
3	Contest in	7.092(469.121)"	27 (4)	CE7>N1g6786>N2g	23 (7)

	Porto 1984 poster	1984"(084.5)		1984^AN7	
4	Portuguese literature poster	821.134.3(084.5)	16 (2)	(AN3+AN4)-Nicpt^AN7	19 (7)
5	Poster of Portuguese commercial graphics from 1980s	766(=1:469)"198" (084.5)	23 (4)	EO960106105>N1cpt >N2i1980^AN7	30 (7)
6	Introducing the reconstruction of the economic mechanism in car repair organizations	629.119 334.4.001.73	7 (1) 12 (1)	→ga EO9\ →ivmfN2w EO925101503	26 (6)
7	The technology of heavy equipment I	621.313.022	11 (1)	EO923 (QC5QH4) – TF(SOe(→ne))	26 (6)
8	Resources for designing steel structures	624.014.2 624.07.001.63	9 (1) 13 (1)	→gc NB1(MEFe) is designing steel structures. We could specify a type of resource such as a textbook.	12 (3)
9	High School textbooks in Russian language	808.2 (075.3)	12 (2)	T4f > CLru \ PE3	12 (5)
10	University textbooks for Numerical methods	518 (075.8)	10 (2)	T4f (TF9f) \ PE1	13 (4)
11	Measurements on electric machines	621.313.083 (075.8)	18 (2)	(EO923 (→ne))(QT2) – TM02	22 (5)

12	One Autumn] Maxim Gorkij	882-321.1-821	13 (1)	AN3 (→gm I>N1cru ^ SOC5) [Note: AN3 is prose; We then capture the themes of travelling in Russia and associating with an underclass.]	20 (7)
13	Trip to Rheinsburg] Kurt Tucholsky	830-321.2-321.4-821+92 Tucholsky, K.	22 (2)	AN3 (IR4→gm /(S+P)>N1cde)	23 (9)

14	Anchor. The world's religions. teacher assistance	29(07)
		372.82(07)
		371.671.12 (07)
			6 (2)
			10 (2)
			14 (2)
				→rh (CR - SO1t)	12 (4)

15	Mariella and the Old Lady's treasure Marita Lindquist. Talking book	839.79-3 (024.7)	15 (2)	TF1(→rt) [This would be the notation for Talking Book]	8 (2)
16	Newspapers	07 917	2 (1) 3 (1)	EO9 55101504	11 (1)
17	Geography of North and Central America	913 (7)	6 (2)	N1bn – TF2 (Note: Central America would be southern North America)	8 (3)
18	(Devices for) reducing lubrication in Mobile machinery and tractors	629.1-42 629.1-43 629.114.2 631.372	8 (1) 8 (1) 9 (1) 7 (1)	→gt ↓→mr E09251 01901	18 (4)
19	Latin language	807.1 (075.8)	12 (2)	CL – TF5→ie T4	11 (5)
20	Sabino Álvarez	35 Álvarez-Gendín, Sabino	28? (2?)	PI2f - TF7d > N1cma (N1ces)	22 (6)

	Gendín: La Administración española en el Protectorado de Marruecos	(048)		[This is notation for Political science of bureaucracy in Spanish Morocco]	
21	Public administration. Urban	012Éhen Gy.	10 (1)	N1g is “Cities” See above for public admin.	9 (2)
22	El túnel de cristal / Maria Gripe. People with physical disabilities	82-31 839.7-3"19"(024.7)	5 (1) 18 (2)	I(IP) I is individual; IP is physical abilities; underline means opposite N2j20 is 20th century	5 (3)
23	Sandokan. El rey del mar / por Emilio Salgari (Juvenile literature)	850-3"18"(024.7)	16 (2)	SA5 is children in general; SA4 is teens….	7 (2)
24	The Island of numbers. Electronic resource. Primary education, first cycle.	51 371.38 (07) 681.31 371.694 372.4 372.851	2 (1) 6 (1) 4 (1) 6 (1) 7 (1) 5 (1) 7 (1)	EO9432115 \ (TF3^→ir) \ PE5 TF3^→ir Education associated with Rehearsing or Practicing EO9432115 \ TF3 computers for education PE5 elementary school	23 (7) 7 (3) 13 (3) 3 (1)
25	Voyages of discovery	910.4	5 (1)	→ip N2x –(→gm I>EO92511)	21 (7)

It may be useful to briefly explain the structure of BCC. In number 8 above, the term →gc for design comes from the schedules of relators, under the schedule g of general relators; the term NB1 for structures comes from the class N of non-human environment, subclass NB for built environment; the qualifier MEFe comes from the class of things M for "Molecules and elements," subclass ME "Chemical elements,": Iron is Fe. Most qualifiers come from schedule Q, such as QT4 "historic" or QI3 "secret"; sometimes these are combined as in QC5QH4 is (more)(mass) or heavy. Note that hierarchies are generally flat and thus notations are usually short. The United Nations Standard Products and Services Code is employed for individual goods within Subclass EO9, yielding lengthier notation. Some verbs formed via combination also have lengthier notation. "Repair" →ivmfN2w in #6 above combines →iv (achieve), →mf (function), and N2w (again).

4.0 Discussion:

We can compare these two columns of subject classifications in several ways.

4.1 The notational length

Brevity is preferable in notation, both for practical reasons and because brief notations are generally easier for users to comprehend. In the 4th and 6th columns of table 1 we indicate the notational length of the notations provided in UDC and BCC respectively. Letters, numbers, and punctuation marks (including periods and parentheses) were each counted, but spaces that might occur between notations were ignored. Since the UDC and BCC notations are not always equally precise, we do not calculate an average notational length here. But a glance at the Table establishes that they are roughly similar in length. In some cases BCC notation is longer; in other cases it is shorter. In the vast majority of cases, there is a rough equality in length.

4.2 The expressivity of the subject classification

The rough equivalence in length is achieved despite quite different approaches to subject classification. Though both UDC and BCC are synthetic – they allow notations from different schedules to be combined – BCC pursues a synthetic approach to a far greater extent. For example, in the first case, UDC achieves a notation for "Transport vehicle engineering" through hierarchical subdivision:

6 Applied technology
62 Engineering. Technology in general.
629 Transport vehicle engineering

BCC instead combines separate terms for "engineering" and "transport vehicles" from different schedules:

Engineering is TF(SOe) where TF indicates “fields” and SO is occupations. [Note that “Engineering” itself is thus a synthetic construct.]

Transport vehicles are E0925 where E09 is “Particular goods and services” and 25 is the general code for vehicles in the United Nations Standard Products and Services Code.

Though BCC also employs hierarchical schedules these are generally much flatter than those in UDC.

We calculate – in parentheses in columns 4 and 6 – the number of separate terms synthesized in each of our subject classifications. Here a stark difference does emerge between UDC and BCC. There are a couple of cases for which UDC and BCC combine the same number of terms. In the vast majority of cases, though, BCC combines more terms than UDC. In some cases this difference is quite large: 7 versus 3 in case 1, 7 versus 2 in case 4, and so on.

Does this difference in number of terms reflect a difference in expressivity? Care must be taken here. Logical subdivision within hierarchies can also be expressive – if the rules guiding subdivision are clear and logical. Yet it would seem that a classifier is more constrained within a hierarchical approach to choose among recognized subdivisions rather than synthesize across any terminology in the schedules. One further advantage of BCC is that letters often (but inevitably not always) reflect the term being signified: “S” signifies the category “Social structure,” “O” captures “Occupations,” and “e” signifies Engineering.

The freedom to synthesize across all schedules allows greater precision in at least some cases. In case 7 “.022” signifies “properties of magnitude” whereas QC5QH4 captures the more precise “heavy.” In case 18 we can specify precisely what is happening with respect to lubrication within BCC.

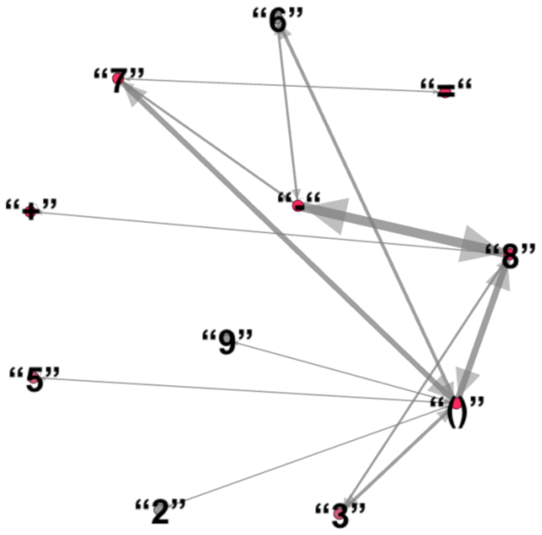
One potential advantage of a synthetic approach is that users can more readily search for related documents. Faced with a subject classification that synthesizes seven different terms, one can reflect on what combinations of these one might wish to pursue. In case 18 one might wonder about reducing lubrication in other devices, or alternatively might wonder about other behaviors involving mobile machinery. In case 12 one might seek out other works – perhaps fictional, perhaps not – that address travelling in Russia or associating with an underclass.

Though we can only be sure once user studies have been performed, it may also be the case that the more synthetic approach will be easier for both classifier and user to navigate. Classifiers may be able to move fairly directly from a sentence in a document description to a classification by simply identifying relevant controlled vocabulary; they will not need to engage as much with hierarchical subdivision. In case 8, for example, the classifier arrives at “steel buildings” by a simple synthesis of those two terms, while the classifier in UDC needs to find this combination deep within the engineering hierarchy. Users likewise may be able to move fairly directly from a query sentence to a string of controlled vocabulary terms (Szostak 2016, 2017). On the other hand, though, classifiers using BCC will need to identify a larger number of separate terms to synthesize.

4.3 Network Analysis

The twenty-five UDC examples are fairly simple. There are 71 points of connection among main classes and auxiliaries in the 40 strings describing the 25 cases (mean 1.77 per string). There are 20 instances of common auxiliaries of form or place, 13 of common auxiliaries of time, but only 1 instance of common auxiliary of language, and 1 instance of coordination of two main classes. Main classes 0 and 1 do not occur, only classes 3, 6, 7 and 8 are blended with common auxiliaries, and only classes 6 and 8 use multiple auxiliaries. Class 8 occurs 8 times in the sample, and has 18 of the total of 71 connecting nodes in the sample (or, approximately 25%). Thus literature, in this sample, has the most complexity in coordination of elements. This is easily contrasted with the five examples of class 6, which all have been assigned multiple strings. A network diagram of the connecting nodes (main classes and auxiliaries) produced using Gephi is shown in Figure 1.

Figure 1. Network map of UDC components from Table 1.



In contrast, the twenty-five BCC strings are fairly complex, but in every case the entire context is represented in a single string. There are 143 points of connection among main classes and relators in the 25 strings (mean 5.72 per string). Main classes could be said to supply semantic context:

Table 2: BCC Main classes

BCC Main Classes	**Frequency**
A “Art”	9
C “Culture”	10
E “Economy”	20
H “Health and Population”	6
I “Individual Differences”	4
N “Non-Human Environment”	10
P “Politics”	6
T “Technology and Science”	8

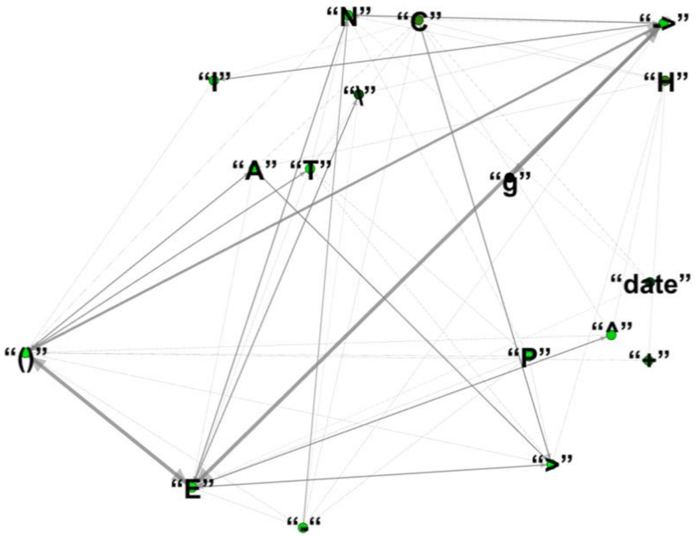
Whereas in the UDC application, the most populous main class is for literature, in the BCC application, the document semantic representations are spread over more specific phenomena. Relators are used in BCC to provide grammar and syntax:

Table 3: BCC Relators

BCC Relators	**Frequency**
->g “general” relators	3
-> “causation” relators	18
> “in” relators	10
^ “associated with” relators	5
+ “and” relators	2
\ “for” relators	4
– “of” relators	6
() “of type” relators	20
Dates “chronology” relators	2

A network diagram of the connecting nodes (main classes and auxiliaries) produced using Gephi is shown in Figure 2.

Figure 2. Network map of BCC components from Table 1.



4.4 The Number of Subject Strings

BCC provides only one subject string for each entry in our sample (though multiple strings might be called for in cases where a document description yields quite different descriptive sentences). UDC provides 2.2; that gives a multiplier or ratio of .45. Though our sample size is small, this seems to be a notable distinction. There is a potential rationale in the context of coextensivity. It may prove easier to search for combinations of terms if these are captured in the same string. Note that Smiraglia, in three previous comparative studies, found that breaking the content into separate strings disperses the probability of co-occurrence. Note also that one key reason for the development of PRECIS was to preserve context by keeping all of the elements in a coordinated string. This led to the notion of thesauro-facets, again a single string containing terms from each applicable facet.

5.0 Brevity, Clarity, Precision: Classification as Language

To the extent that a classification consists of symbols that represent concepts, and rules for combining them into meaningful statements, classification can be seen as a sort of language. Indeed, writers on knowledge organization in general, and classification in particular, often use the "language" metaphor to describe the structure and function of classifications. For example, Svenonius (2001, 54) refers to all components of the domain of knowledge organization as "bibliographic vocabularies" and in particular, refers to "classification language[s]." More recently, Smiraglia, van den Heuvel and Dousa (2011) used quotations from Paul Otlet to discuss concepts of precision in the context of classification as a documentary language:

> "A word […] not only evokes the object named in its concrete form, but also by logical association, all the characteristics and attributes of the object in the same way that the formula for a compound expresses its relationships and quickly makes its elements evident" (Otlet, 1891-1892: 19).

and

> 'Classification numbers will […] be complex numerical expressions made up of different factors whose respective meanings when juxtaposed will express a complex idea after the fashion of compound words in spoken languages" (Otlet,1895–1896: 52).

Smiraglia, van den Heuvel and Dousa (2011, x) wrote:

> Otlet's notion that the structure and the characteristics of the relationships between classes and the dynamics of interaction between them were somehow comparable to language and have implications for notation is important for our question of how syntax and semantics interact in various KOSs.

and:

> In the analogy of classifications as artificial language their grammars posses lexemes—i.e., terms representing concepts or classes—that are organized by means of paradigmatic and syntagmatic relations. The paradigmatic relations express the meaning of terms by establishing hierarchical relations among them, while syntagmanic relations provide the syntax for combining lexemes into more complex terms (Hutchins, 1975: 6–7, 33–55; Svenonius, 2000: 131).

Thus there is more to the evaluation of a classification as a knowledge organization system than simple judgment of its array of concepts or its ability to express complexity. Rather, another means of evaluation is the adjudication of its capability for brevity, clarity and precision in the expression of the content of works.

Indeed, we might here appeal to essential concepts of expressivity in language. It is for this very reason that the seventeenth century Flemish mathematician Simon Stevin explained the greater functionality of Dutch for scientific representation, based on the concept that Dutch contains more monosyllabic words than other classical languages. Van den Heuvel wrote (14):

> Stevin was convinced that Dutch was superior to other languages, such as Greek, Latin or French to explain scientific concepts because it contains far more monosyllabic words which could be combined to create clear compound words. To support his view, Stevin included in this introductory discourse a list with hundreds of monosyllabic words in Dutch of which their Latin and French translations needed more syllables to express the same concept.

Strunk and White famously suggest that a writer should prefer specificity and concreteness (30), avoid unnecessary wordiness (32), and use parallel syntax (35), all of which are relevant to the traditional KO point of view concerning coextensivity and expressivity—two critical aspects of the implementation of a classification.

Bliss (1929) made this very point with regard to criticism of the bibliographic classifications of his day, many of which are still the most important such systems today. With regard to analysis and synthesis he wrote (407, emphasis original):

> Knowledge is both analytic and synthetic. In analysis we pass from the more general to the more special, from the more comprehensive to the more definite. In synthesis, the antithetic process, we pass from the more specific to the more general and comprehensive. A system of knowledge should function in both these ways; it should be both analytic and synthetic …. Analysis is analogous to the branching of a tree. Synthesis is analogous to the confluence of streams in a widening valley, or to the unitary relation of twigs to branches and of the branches to the tree. In this analogy we are wont to validify the metaphor of *the tree of knowledge.*

Thus, brevity, clarity and precision are critical to the dynamic synergy of the relationship between analysis and synthesis. He went on to criticize bibliographic classifications as (412) “structurally wrong,” “below *maximal efficiency*” in their ability to collocate subjects, and ultimately, “uneconomical,” by which he means they are lacking precision.

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. Notably, BCC subject strings pursue a grammatical construction (Szostak 2017). The flexible and multi-faceted UDC, because of its disciplinary base, must instead resort to multiple, overlapping and therefore uneconomical use of multiple strings to achieve coextensivity in the expression of a works' knowledge content.

References

Heuvel, Charles van den. Forthcoming, 2017. "As the author intended". Transformations of the unpublished writings and drawings of Simon Stevin (1548-1620)." In: *Translating Early Modern Science*, eds. Sietske Fransen, Niall Hodson and Karl A.E. Enenkel. Leiden: Brill.

International Federation for Documentation. 1905. *Manuel du Repertoire Bibliographique Universel*. Bruxelle: Institut International de Bibliographie.

Otlet, P. 1891–1892. Something about Bibliography. In: Rayward 1990, pp. 11–24. [Originally published as "Un peu de bibliographie", *Palais*, 1891–1892, pp. 254–271.]

Otlet, P. 1895–1896. On the structure of classification numbers. In Rayward 1990, pp. 51–62. [Originally published as "Sur la structure des nombres classificateurs", *IIB Bulletin* 1 (1895–1896), 244–249].

Rayward, W.B. (ed. & trans.) 1990. *International organization and dissemination of knowledge: Selected essays of Paul Otlet* (FID 684). Elsevier: Amsterdam.

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. 2013. "Big Classification: Using the Empirical Power of Classification Interaction." In *Proceedings of the ASIST SIG/CR Classification Workshop, Montréal, 2 November 2013*, ed. D. Grant Campbell, p. 21-29. doi: 10.7152/acro.v24i1.14673

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

Smiraglia, Richard P. 2016b. "Extending Classification Interaction: Portuguese Data Case Studies." In *Knowledge Organization for a Sustainable World: Challenges and Perspectives for Cultural, Scientific and Technological Sharing in a Connected Society, Proceedings of the Fourteenth International ISKO Conference 27-29 September 2016 Rio de Janeiro Brazil*, eds. José Augusto Chaves Guimarães, Suellen Oliveira Milani and Vera Dodebei. Advances in Knowledge Organization 15. Würzburg: Ergon Verlag, 97-104.

Smiraglia, Richard P. Charles van den Heuvel and Thomas 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, pp. 25-40.

Strunk, Oliver and E. B. White. 2000. *Elements of Style*, 4th ed. New York: Allyn and Bacon.

Svenonius, Elaine. 2000. The intellectual foundation of information organization. Cambridge, MA.: MIT Press.

Szostak, Rick. 2013, regularly updated since. *Basic Concepts Classification*. https://sites.google.com/a/ualberta.ca/rick-szostak/research/basic-concepts-classification-web-version-2013

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

Szostak, Rick. 2017, forthcoming. “Facet analysis without facet indicators” In *Dimensions of Knowledge: Facets for Knowledge Organization* (Richard Smiraglia and Hur-li Lee, eds.). Berlin: Springer.