Submitted on: 14.11.2017

2016 Satellite meeting - *Subject Access: Unlimited Opportunities*
11 – 12 August 2016
State Library of Ohio, Columbus, Ohio, USA

The Simplest Approach to Subject Classification

Rick Szostak
Department of Economics, University of Alberta, Edmonton, Canada
rszostak@ualberta.ca

Abstract:

Museums, galleries, archives, and many libraries – as well as a host of online databases of various types – seek a simpler approach to subject classification. This paper poses the question: "What is the simplest approach to subject classification?" It suggests a synthetic approach which pursues a sentence-like approach to classification. Such an approach not coincidentally has the added advantage of best capturing the nature of a work.

Keywords: Classification, Synthetic, Facets, Poly-coordination, Interoperability

The Challenge

Some public libraries in the United States have switched from the Dewey Decimal System (DDC) to BISAC, the classification system employed in bookstores, claiming that their users find DDC too challenging. Museums, art galleries, and archives increasingly have an online presence, and aspire to varying degrees to provide users with online access to the items in their collections, but have not found existing library classification systems amenable (Menard, Mas, and Alberts 2010). There are a host of online databases of various types, and almost as many different classification systems as there are databases. And then there is the Semantic Web, which aspires to allow computers to draw connections across databases by coding these in terms of a shared controlled vocabulary and set of syntactic rules, and thus needs some sort of Knowledge Organization System (KOS; perhaps a formal ontology, but arguably a classification system can suffice; Szostak 2014). In most/all of these cases there are understandable concerns that both classifiers and users cannot be expected to master a complex KOS.

For a variety of reasons, then, it would be wonderful if we could develop a system of subject classification that was much simpler for both classifiers and users than present systems, but yet also provided exhaustive coverage. Such a system would allow scholars and general users to readily find what they were looking for across a range of databases.

It seems useful then to ask a simple question: *What is the simplest feasible approach to subject classification?* The rest of this paper will explore this question and suggest an answer.

The Proposed Solution

Since the days of Ranganathan, some sort of "faceted" approach to classification has been widely recommended within the field of Knowledge Organization. Such an approach provides for synthetic subjects that combine simpler elements into a more complex subject heading. The enumerative classifications – which instead delineate a set of complex subject headings – that dominate the library world have to varying degrees adopted some elements of the faceted approach. Most obviously, a fixed set of geographical and temporal identifiers can be amended to a host of subject headings, as in "Poetry – Nineteenth century – France."

Yet facet analysis as generally advocated will likely not meet our goal of "simplicity." Ranganathan (1967) suggested 5 key facets; later applications of facet analysis such as the Bliss2 Classification and the Integrative Levels Classification expand the number of facets to over a dozen. Classifiers must thus learn to distinguish these facets, and also master rules as to in what order they are to be captured in the subject classification of a particular work. Users may have to understand facets also, unless some sort of computer interface can do the work of translating their query into the terminology of facets.

While formal facet analysis is necessarily somewhat complicated, it may be possible nevertheless to pursue a simpler type of facet analysis. At the heart of the faceted approach is the idea of a synthetic approach to classification whereby simpler terms are combined in order to produce the subject classification of a particular work. It is quite possible to pursue a synthetic approach that does not require the formal identification of a dozen facets. Rather we can simply mimic the basic sentence structure with which we are all intimately familiar. If a book is about dogs biting mail carriers we can give it a subject classification of (dogs)(biting)(mail carriers). A classifier should have little trouble developing this subject string, and a user need not master anything more than common sentence structure in order to retrieve what they are looking for. Both can be aided by a visual interface and/or a comprehensive thesaurus (see below). And note that it is abundantly clear, but need not be stated, that the first term captures the active agent in facet analysis, the second the process, and the third the agent that is acted upon. Nor should this be a surprise, for we regularly communicate information about facets to each other in everyday conversation without pausing to specify which facets we are referring to.

The literature in classification theory would guide us to answer the question of "What is the simplest approach to subject classification?" with a further question: "Enumerative or faceted?" The answer given here is "No." But we can employ the core idea at the heart of facet analysis in order to provide a simpler approach.

Simpler is Better Too

It is a cornerstone of classification theory that a subject classification should as closely as possible capture the essence of a work. While there is debate in the field about the "nature of a work," it seems clear that a/the key element of the nature of a work is the main arguments that the work makes. A book that is about why dogs bite mail carriers is thus best classified as (dogs)(biting)(mail carriers). This synthetic classification most closely represents what the work is about.

Smiraglia (2001) identified the key element of the "nature of a work" as the ideas that the work contains. He did not expand in detail on the nature of these ideas. But humans express ideas generally in the form of sentences. And most works – and especially scholarly works – discuss how one or more things exert a particular influence on one or more other things: (broken)(windows)(encourage)(crime) or (statins)(reduce)(blood pressure) or indeed (why)(dogs)(bite)(mail carriers). The remainder generally investigate the nature of particular things: (steel)(is)(strong). In all of these cases, the nature of the work is captured by a sentence fragment containing at least one noun-like term, at least one verb-like term, and perhaps adjectives or adverbs; these terms are arranged in an order dictated by grammar. We need, then, separate classifications of "things," verb-like relators between things, and adjectivial/adverbial qualifiers; and agreement that these will be arranged and searched for in grammatical order. We need, in other words, to agree to organize our subject classifications in a sentence-like structure that mimics everyday grammar (that we all use all the time, even if we struggled to identify grammatical rules in elementary school), and then we need a controlled vocabulary of (only) the most basic components of human sentences. [Note that we can generally skip prepositions: we thus generally capture sentence fragments rather than full sentences.]

It is not a fluke that the simplest approach also has representational advantages**.** *It is precisely because the recommended approach to subject classification captures in sentence (fragment) format the key arguments of a work that it is so simple for classifier and user to employ.* When a user approaches a reference librarian with a query, they phrase it as a sentence. The reference librarian then helps them turn their sentence into a complex subject heading that roughly approximates their query. That complex subject heading then guides them to many works they do not want. If the work is about a sentence (fragment), and the user query is phrased as a sentence (fragment), the subject classification that connects them should ideally be in the form of a sentence (fragment).

Why haven't we always done it this way? Because a synthetic approach would have been supremely difficult in the age of card catalogues. We have now devoted well over a century to the development of classifications well suited to card catalogues. Digitization creates both a demand for a new approach (see above) while facilitating this.

Documents, Ideas, and Objects

It has often been argued that a subject classification scheme should be able to classify both documents and ideas (Gnoli 2007). This will be especially important if scholarship moves toward a system whereby scholars contribute "nuggets" of information to a network of understandings (Börner 2006). The approach recommended above classifies documents in terms of the ideas they contain and thus is admirably suited to the classification of both documents and ideas.

Galleries, museums, and a host of commercial and government databases wish to classify objects rather than ideas or documents. Happily the recommended approach has a distinct classification of things that can be used for this purpose. Note that we will often want to qualify these things with adjectives: (golden)(axe). Moreover, we will often want to capture the purpose or method of manufacture or composition of objects, and this will also best be achieved with sentence fragments: (axe)(for)(war); (mass)(produced); (wooden)(shaft)(steel)(head). Happily, then, the simplest approach to classifying documents and ideas is also the simplest approach to classifying objects.

A Host of Objections

Is the proposed system feasible? We can sketch here the most prominent objections and the potential answers to these.

Pre-and post-coordination each have advantages and disadvantages.

This argument is repeated often in the literature (see Sauperl 2009). Post-coordination (in which the classifier synthesizes simpler terms to form a subject classification) is mainly accused of lacking precision: The user looking for works on philosophy of history is shown many works on history of philosophy. But this is not a problem with post-coordination but rather with search interfaces. It is blindingly easy (I have had computer science students do this for me) to generate search algorithms that prioritize the order of search terms (see Szostak 2016). Our hypothetical user above need not be told about works on mail carriers biting dogs unless they wish it.

There is too much terminological ambiguity in the world.

This is, of course, an argument that threatens the pursuit of any classification that seeks general coverage. Yet of course classificationists have been grappling – with some degree of success – with ambiguity for millennia. The answer here is that ambiguity can be lessened by focusing on "basic" concepts. Users disagree a lot about what "globalization" means, but have very similar understandings of "dog" and "mail carrier." The recommended approach can generally utilize (combinations of) basic concepts. Moreover the shared understandings associated with basic concepts can be further clarified by placing these terms within logical hierarchies (Szostak 2011). This is hard to do in enumerative classifications: Recycling is treated illogically as a subclass of garbage because there is no other place to put it. In the recommended approach, recycling is captured synthetically and logically as something that is done to garbage.

It is not as easy as it looks.

In utilizing any KOS, classifiers and users must be able to readily identify controlled vocabulary. But the classifier or user might input "mailman" instead of "mail carrier." A logical classification can help: classifiers and users can look in the classification of things for the subclass of occupations and fairly readily identify appropriate terminology. A comprehensive thesaurus could even more quickly guide users to the correct term: they input "mailman" and are instantaneously told to employ "mail carrier." It could also tell them that "dog" is a narrower term than "canine." We will be aided in developing such a thesaurus by the fact that we can employ flat and logical hierarchies for the terms in our controlled vocabulary (since we are not trying to capture combinations of nouns and adjectives and sometimes verbs in a single pre-coordinated subject heading: Again, we need not treat "recycling" as a subclass of "garbage" simply because there is no other place to put it).

Visual interfaces can usefully convey such thesaural information. Note that neither user nor classifier need try to guess what composite subject terminology might capture the whole idea of dogs biting mail carriers, but can enter three simple search terms and be quickly guided to appropriate vocabulary. And visual interfaces can guide them to a host of similar queries: cats biting mail carriers, dogs licking mail carriers, mail carriers biting dogs. This facility may be particularly important if the library does not possess documents that directly address the user's search query. The user may often need to identify different works that address different components of a complex query (Green 1995). Moreover by clicking on any of their search terms the user can be exposed to a host of other subject classifications that incorporate it: Their curiosity may then guide them to other aspects of the lives of dogs and mail carriers or to other instances of biting. It is far harder to identify a range of related queries within a world of pre-coordinated subject headings, but this sort of information will be of profound value to a user performing an exploratory search.

Verbs!!

Subject classification at present emphasizes nouns, and to a lesser extent adjectives. When verbs are used, they are generally translated into noun form: communication rather than communicate. But of course some sort of verb or verb-like term is at the heart of what most works are about. We can hardly guide our user to the work they seek if we eschew reference to "biting" in our subject classification.

It won't work for everything

The vast majority of scholarly books and articles, and most general works of non-fiction make some sort of argument. A minority instead talk about the characteristics of a place or thing. These can also be captured synthetically: (steel)(is)(strong); (Mexico)(is)(mountainous). Archival documents can also be captured synthetically: (hiring)(practices)(for)(hospitals). The subject matter of a work of art (bomb)(exploding) and the theme (horror)(of)(war) can also be captured synthetically. A subject classification of museum artefacts needs to grapple with the purpose of an item (battle)(axe), its method of manufacture and material (forged)(from)(steel) and perhaps place of manufacture (near)(lake)(south-eastern)(Europe). In all these cases, a synthetic approach which links nouns, verbs, adjectives and adverbs in the order they would normally appear grammatically will provide the ideal subject classification.

As for the Semantic Web it is to be coded in terms of RDF triples of the format (subject)(predicate or property)(object). It thus needs both a sentence-like structure and distinct controlled vocabulary for nouns, verbs, and adjectives/adverbs. The Semantic Web community has signally failed to agree on controlled vocabulary; this is a task best performed by the Knowledge Organization community. This we can only do if we embrace the sort of approach to classification urged above. That is, we need to recognize the value of synthesizing across separate classifications of things, relators, and adjectives/adverbs.

What about exploratory search?

The examples provided above have generally assumed that a user knows what they are looking for. What, though, of the user who is browsing with a more general curiosity? Such users can input one or more search terms. The sort of visualization techniques urged above can then alert them to a host of possibilities: these search terms can be placed within classifications, alerting them to narrower, broader, and similar nouns or verbs or adjectives. They can be shown possible combinations of their search term(s) with other terms. They input "dog" because they have just bought a dog and are alerted to the fact that they might read about “biting” among a variety of other practices. The visualization interface can draw upon all of the subject strings that involve “dog” within the database, while also alerting the user to the possibility that further connections are drawn with “canine” and “poodle.”

This facility for exploratory search is particularly important given the literatures on serendipity and literature-based discovery (McCay-Peet & Toms 2015, Workman, Fiszman & Rindflesch 2014). By following the links suggested by the search interface, users can discover connections between ideas in different documents that have not previously been connected.

Different languages employ different grammar

This is indeed a challenge. The approach recommended here will complicate the task of translating the classification into other languages where standard word order is different. French, for example, tends to place adjectives after nouns rather than before. Note, though, that each language has a standard word order. It is thus entirely possible to develop classifications that respect the grammar of each language. Translation across these will not be difficult if grammatical differences are general and clearly stated. In moving from English to French we could generally move adjectives behind nouns.

There are, it should be confessed, grammatical challenges within any one language. There are practices that violate grammatical norms. And there are often alternative ways of expressing an idea. Such challenges deserve further investigation. Appropriate search interfaces could suggest alternative word orders to users.

How do we know it will work?

Szostak (2013) has developed separate classifications of things, verbs, and adjectives/adverbs in his Basic Concepts Classification. The Integrative Levels Classifications takes a different approach to facet analysis but also develops separate classifications of things, relators, and properties. Extensive justification of this general approach to classification is provided in Szostak, Gnoli, and Lopez-Huertas (2016). As noted above, classification is pursued with respect to basic concepts, terms for which there is likely considerable agreement on meaning across individuals, disciplines, and cultural groups.

Szostak worked with a group of computer science students in 2015. They developed a search interface that allowed users to search a small database of some dozens of museum artefacts that were classified using sentence-like strings of terminology from the Basic Concepts Classification. As noted above, it proved quite easy to prioritize the order of search terms in the search interface. Given the small size of the database and the lack of a general thesaurus, it is not possible at this point to provide precise metrics on the ease of use of such a search interface. It would be highly desirable to perform user testing on a larger database.

The recommended approach links the fields of Knowledge Organization and Information Retrieval: We can only achieve the simplest subject classification if this is linked to an appropriate search interface. Cleverly and Burnett (2015) make the general point that we can combine automatic and manual Knowledge Organization methods to facilitate search, and note in particular that automatic query expansion software employs such KO tools as thesauri in order to suggest search terms. The approach recommended here can be seen as an extension of such strategies.

Conclusion

The two questions "What is the simplest approach to subject classification?," and "What type of subject classification best replicates the essence of works?," both point to a feasible synthetic approach which follows common sentence structure and employs simple and logical hierarchies of nouns, verbs, and adjective/adverbs. Such a subject classification has myriad uses in the contemporary world. It will also, importantly, encourage literature-based discovery.

References

Börner, K. (2006). Semantic association networks: Using semantic web technology to improve scholarly knowledge and expertise management. In V. Geroimenko and C. Chen (Eds.),*Visualizing the Semantic Web,* 2nd ed. (pp. 183-98). Berlin: Springer.

Cleverley, P. & Burnett, S. (2015) The best of both worlds: Highlighting the synergies of combining manual and automatic Knowledge Organization methods to improve information search and discovery. *Knowledge Organization* 42:6, 428-44.

Gnoli, C. (2008). Ten long-term research questions in knowledge organization. *Knowledge Organization* 35(2/3), 137-49.

Green, R. (1995). Topical relevance relationships. 1. Why topic matching fails. *Journal of the American Society for Information Science* 46(9), 646-53.

Integrative Levels Classification (ILC), http://www.iskoi.org/ilc/.

Martínez-Ávila, D., San Segundo, R., Olson, H.A. (2014). The Use of BISAC in Libraries as New Cases of Reader-Interest Classifications. *Cataloging & Classification Quarterly* 52(2), 137-55.

McCay-Peet, L.& Toms, E.G. (2015) Investigating serendipity: How it unfolds and what may influence it. *Journal of the Association for Information Science & Technology* 66(7), 1463-76.

Ménard, E., Mas, S., Alberts, I. (2010). Faceted classification for museum artefacts: A methodology to support web site development of large cultural organizations. *Aslib Proceedings* 62 (4/5), 523-532.

Ranganathan S.R. (1967). *Prolegomena to Library Classification*, 3rd ed. Bangalore: SRELS.

Šauperl, A. (2009) Precoordination or not?: A new view of the old question. *Journal of Documentation* 65:5, pp. 817 – 833.

Smiraglia, R.P. (2001). *The Nature of "a work": Implications for the Organization of Knowledge.* Lanham MD: Scarecrow Press.

Szostak, R. (2011). Complex concepts into basic concepts. *Journal of the American Society for Information Society and Technology* 62(11), 2247-65.

Szostak, R. (2013). *Basic Concepts Classification.* Available at: https://sites.google.com/a/ualberta.ca/rick-szostak/research/basic-concepts-classification-web-version-2013

Szostak, R. (2014). The Basic Concepts Classification as a Bottom-Up Strategy for the Semantic Web. *International Journal of Knowledge Content Development and Technology*, June 2014. http://dx.doi.org/10.5865/IJKCT.2014.4.1.039

Szostak, R. (2015). A pluralistic approach to the philosophy of classification. *Library Trends* 63(3), 591-614.

Szostak, R. (2016). Poly-coordination. *Proccedings of the Annual Conference of the Canadian Association for Information Science.* http://www.cais-acsi.ca/ojs/index.php/cais/issue/archive

Szostak, R., Gnoli, C. & Lopez-Huertas, M. (2016) *Interdisciplinary Knowledge Organization.* Berlin: Springer.

Workman, T.E., Fiszman, M., Rindflesch, T.C., & Nahl, D. (2014). Framing serendipitous information-seeking behavior for facilitating literature-based discovery: A proposed model. *Journal of the Association for Information Science and Technology* 65(3), 501-12.