

# The Importance of Knowledge Organization

by Rick Szostak

## Economics of Knowledge Organization Systems

### EDITOR'S SUMMARY

Knowledge organization systems (KOS) are ubiquitous and helpful, but from an economist's perspective their value and contribution to economic productivity are hard to quantify. Though estimating the aggregate impact of KOS is a challenge, it is possible to enhance that impact through the use of a shared or interoperable controlled vocabulary with rules for interpreting text. Like RDF triplets underlying the Semantic Web, elements can be identified to represent phenomena, relationships and properties, and then combined using a synthetic approach to generate complex concepts. Such a vocabulary would be enhanced with related concept information, expanding a user's discovery through a web of hierarchical and associative relationships. The vocabulary would be interdisciplinary and applicable to any database, geared to building upon simple expressions for generic phenomena. The KOS could serve as a universal thesaurus complementing the combinable elements used to classify content with the potential to be a landmark classification system.

### KEYWORDS

knowledge organization systems	controlled vocabularies
interdisciplinarity	classification
interoperability	postcoordinate indexing
semantic web	economics of information

Rick Szostak is professor and associate chair in the Department of Economics, University of Alberta. He can be reached at [rick.szostak@ualberta.ca](mailto:rick.szostak@ualberta.ca).

**K**nowledge organization systems (KOS) are ubiquitous in the world. Yet we have no precise idea of just how important they are. Without some sort of KOS, it has usually been difficult to locate any object – book, article, artifact, document, work of art – in any collection. The advent of full-text searching has allowed some users to circumvent KOS in their searches for certain kinds of object. But almost all databases still develop some sort of KOS.

Economists try to estimate the contribution to economic productivity (our collective ability to turn inputs into valued outputs), of various forms of human activity. The next section of this paper draws some lessons from the experience of economists in trying to estimate the productivity impact of computers. The discussion is not meant to deter anyone from seeking to estimate the aggregate productivity contribution of KOS, but does suggest that such an exercise would be fraught with difficulty.

The subsequent section reflects in a non-quantitative manner on the importance of KOS. The third section then explores the possibility that certain types of improvements to KOS will significantly enhance productivity. The author confesses that he has migrated to the KO field because of a belief that at this point in time it is the most important field in the entire scholarly enterprise.

### Productivity and Computers

Nobel Laureate Robert Solow famously opined in 1987 that “[y]ou can see the computer age everywhere but in the productivity statistics.” [1, P.36] That observation was often repeated over the next decades. Though some economists can now detect statistically some impact of computers on economic productivity, for decades this was not the case.

Naturally, economists wondered about this productivity paradox. It seemed inconceivable that computers could be as unimportant as statistical analysis suggested. Several possible explanations were posited [2]. First, computers likely have their biggest impact within the service sector, but we measure output in services less well than we measure output in agriculture or manufacturing. A second and related problem is that computers may have their biggest effect in increasing the quality of output rather than the amount of output. Economists struggle to capture the value of quality improvements. If these are under-estimated then the causes of quality improvement will be under-appreciated.

Third, computers are used in concert with costly human inputs. This limits their productivity impact. But it also increases the possibility that we incorrectly estimate this impact by attributing the output to the people rather than the machine. We will return to this point below, for one possibility in the world of KOS is that the Semantic Web may allow computers to accomplish a lot without much human intervention.

Fourth, it was observed that it takes a while for companies to figure out how best to employ computers. There might thus be a considerable time lag between when computers are first purchased and when they have a positive impact on productivity. There might even be a significant learning period during which the net impact on productivity is negative. Few economists go so far as to argue that this time lag was decades in length and thus could explain why no positive productivity impact was found for so long. But the possibility of long and variable time lags complicates the estimation procedure. A fifth and related problem is that computers did not replace the use of paper-based filing systems to the extent that early prognostications of paperless offices had suggested. This limits the productivity impact, but it also raises questions regarding the quality advantage of having both digital and paper records.

Similar problems afflict any effort to calculate the aggregate productivity impact of KOS. These likely also disproportionately affect the service sector. If the main effect of a KOS is to enhance the speed with which a user finds information, it is (theoretically at least) fairly straightforward to calculate the value of the time saved. But if a KOS guides users to better information, then it becomes much harder to calculate the value of this qualitative improvement.

Like computers, KOS are only valuable when applied by some human actor. As with computers themselves, the Semantic Web holds out the possibility that a KOS can accomplish much with limited human input

Time lags affect KOS use in multiple ways. We know that individual searches generally involve multiple interactions with a KOS over a period of time: the benefit to the user may occur some time after the initial query. And most users interact with a particular KOS for multiple queries and become more adept with practice. As noted above, the ubiquity of KOS does not prevent alternative search strategies from occurring with some frequency.

There are further challenges. While economists might compare a pre-computer and post-computer world, KOS have existed as long as there have been collections of objects. Even more so than computers, perhaps, KOS have much of their impact outside of the realm of economic production. Helping people find information on gardening or child care would not enhance economic productivity as usually defined. Even if time is saved during search, that time is only valuable (in a productivity calculation) if devoted to economic production.

Crucially, if someone is guided by a KOS to economically useful information, that information would get much/all of the credit for any resultant increase in output. In other words, a KOS by its nature provides an intermediate service: it is one step in a chain of events. It is possible, but not easy, to disentangle the effects of each step in such a chain of events.

### Productivity and Knowledge Organization Systems

As with computers, the fact that it is challenging to estimate the productivity impact of KOS does not necessarily mean that this impact is not huge. The very ubiquity of KOS, given that they are costly to develop and manage, suggests that they are important. Indeed, if we assume that economic agents are generally rational (as economists have tended to do), then we might surmise that the costs of developing and managing KOS are exceeded by their benefits. If so, then an estimate of these costs would provide a lower bound estimate of the benefits of KOS. But economists never take such a shortcut in their studies of computers; indeed, they sometimes speculate that the investment might have been ineffective at least in the short term.

Billions of people use classifications directly or indirectly (through browsing shelves) to find information that they consider useful. They do this repeatedly and must thus perceive some benefit in the practice. If we save them time and or help them find better information, we enhance their productivity (as individuals if not always as economic agents).

For an organization developing a KOS one key question involves interoperability. If the organization wishes only its own employees to have access, then a unique KOS may be advantageous (though a hacker may hardly be dissuaded by needing to master a new KOS). The organization might still worry about the training costs of introducing employees to a novel system. If it is hoped that users from beyond the organization will access data, then employing or adapting a familiar KOS may be advantageous. This may also be the case if development or training costs are judged to be high. But how easy is it to borrow or adapt existing KOS? We will argue in the next section that it could be easier.

### Enhancing KOS Productivity

While we have not estimated the aggregate productivity impact of KOS, our discussion does allow us to identify strategies for enhancing that productivity impact. We mentioned one of these above. The premise of the Semantic Web is that computers can draw inferences across databases if these are each coded in terms of a shared (or at least interoperable) controlled vocabulary and set of syntactic rules. But after decades of work the Semantic Web community is far from achieving consensus on a KOS. Indeed research has lately turned away from the formal ontologies that it was hoped would provide the necessary shared vocabulary and rules [3]. Databases are thus coded in terms of a range of different KOS grounded in different assumptions. The potential advantages of the Semantic Web will only be realized if consensus on a particular KOS is achieved. I have argued elsewhere that the solution may lie in a bottom-up strategy where we first classify the phenomena (things), relationships and properties required by the Semantic Web and then add syntactic rules as necessary [4]. It is noteworthy that coding for the Semantic Web necessarily takes a synthetic approach: the RDF triplets employed take the form (object)(predicate or

property)(object), and thus it is assumed that we will in coding (fairly) freely combine phenomena, relationships and properties.

Even if we cannot yet facilitate computer navigation across databases, we might make it easier for humans to navigate across databases [5]. There is a strongly felt need for standardization of the classification systems employed by different online databases. As databases multiply but employ incompatible KOS, the potential productivity impact of developing a unifying KOS grows exponentially. A system that will be widely used will have to be easier to master than the KOS employed today for libraries. But such a system must nevertheless be capable of coping with the complexity of contemporary information. These requirements suggest (to me at least) a synthetic approach where simple terms can be combined to generate complex understandings.

A synthetic approach grounded in what I have elsewhere termed *basic concepts* will likewise facilitate search; users can then combine simple search terms to clarify what they are looking for [6]. This should make it easier both to master the KOS and to pursue individual queries. (A universal thesaurus would aid such searches; see below.)

Our analysis above guides us to worry not just about search speed but also about quality of information obtained. Two considerations seem particularly apposite here. The first involves those literatures of “undiscovered public knowledge,” “literature based discovery” and “serendipity” [7]. These literatures each extol the advantages of juxtaposing related but separate pieces of information. The history of science likewise tells us that the greatest discoveries generally involve combining previously separate pieces of understanding. Similar arguments are found in the history of technology and art history. Users cannot know what novel combinations to search for (or they would not be a novel combination). We need then a KOS that not only guides users to information they know to look for but to related information for which they might not have searched.

Again a synthetic approach to KOS is indicated, so that users can easily explore possible connections to their original subjects of interest. Users start out interested in whether A influences B. They should be able to quickly move to investigating influences on A, how these might be related to B, what other

phenomena seem associated with B and so on. They should be facilitated, that is, in exploring any combination that includes their original phenomena of interest. Likewise, they may posit a particular relationship between A and B and should be facilitated in exploring other examples of that relationship occurring between quite different phenomena. And users may also on occasion find value in exploring works that address the same property (say, *beautiful*). The synthetic approach thus instantiates a “web-of-relations” [8] in which users can readily follow their curiosity to related items.

The second consideration is interdisciplinarity. I came to the field of KO as an interdisciplinary scholar frustrated by the barriers that extant KOS place in the way of interdisciplinarity [9]. Works on a particular phenomenon or relationship can be strewn across several disciplines, each classified in different ways with different terminology. I have worked to develop a universal classification grounded in phenomena (and relationships among these and properties of these) rather than disciplines [10]. Again, a synthetic approach seems best, allowing works to be classified in terms of any combination of phenomena, relationships and properties. This sort of approach seems well suited to the Semantic Web (see above). And the three literatures cited in the previous paragraph all stress the particular value of linking related information from different disciplines: this will obviously be easier in a KOS grounded in phenomena and relationships than one grounded in disciplines.

The vast bulk of scholarly research examines how one or more phenomena affect in a particular way(s) one or more others. The same may well be true of general works: (dogs)(bite)(mail carriers); (gardeners)(growing)(flowers). The best way to capture the unique contribution of such works is by synthetically capturing (phenomenon)(relationship)(phenomenon), as in (particular drug)(reduces)(blood pressure). Importantly the elements connected synthetically will generally represent basic concepts: the things and relationships that we regularly observe in the world. These are likely the terms for which there is the greatest agreement on meaning across groups, individuals and disciplines. The minority of works that capture the properties of a phenomenon or relationship can also best be captured synthetically: (steel)(is)(strong). The synthetic strategy advocated in the

interest of interdisciplinarity will thus, by better capturing the unique contribution of a work, be useful for specialized disciplinary scholars as well. Other elements of uniqueness will be captured by classifying works in terms also of theories, methods and perhaps authorial perspectives applied; these various elements were stressed in the León Manifesto of 2007 [11]

One further advantage of the synthetic approach is that it reduces/eliminates the temptation to abuse hierarchy. *Recycling* is treated as a subclass of *garbage* because there is no other obvious place to put it [12]. But recycling is something we do to garbage, not a kind of garbage. A synthetic approach allows *recycling* to be treated as a relationship. While the human user may (or may not) be unfazed by a classification treating recycling as a subclass of garbage, a computer attempting to navigate a classification will certainly be challenged. We can aspire to a classification in which subclasses are strictly logical, and different types of subclass (“type of,” “part of”) are strictly delineated. With respect to relationships, a combinatory approach may prove superior to hierarchy: thousands of verb-like relationship terms can be generated by combining some 100 basic verb terms (which can themselves be grouped into a handful of logical classes) with each other or with phenomena or properties [13].

Information scientists have devoted much attention to one particular sort of cross-database interoperability: how can we facilitate searches across libraries, archives, museums and art galleries? [14] It is worth noting that a synthetic approach utilizing basic concepts is applicable to all. Museum curators may not be inclined to master library classifications, but might find classifying their artifacts using combinations of basic concepts attractive: (axe)(for)(fighting). Museums collect artifacts that are either representative or unique: a synthetic approach allows both of these to be indicated, and specific elements of uniqueness to be represented: *broken, used by Julius Caesar, gold-plated*. Archival documents can also be represented in terms of the processes or phenomena implicated in a document. And works of art can be classified not just by artist but by the objects and actions captured in a work: (girl)(riding)(horse); as art galleries and art historians explore a thematic approach to art, this sort of thematic classification becomes increasingly important [15].



We have focused here on classification systems. But complementary advances in other KOS are also desirable. In particular, a universal thesaurus – or at least interoperable thesauri – would complement a universal phenomenon-based classification. Users could then enter any search terms and be guided to controlled vocabulary. If we wish to fully avail ourselves of digital possibilities, and especially the Semantic Web, then greater clarification of hierarchical (distinguishing “kind of” narrower terms from “part of” narrower terms), equivalence (distinguishing different degrees of similarity) and especially related term relationships (there are several distinct kinds of related terms recognized in ISO standards, but these are rarely distinguished in thesauri) are called for [16].

The sort of approach to classification suggested above is much easier in a digital world. It would have been extremely difficult in an age of card catalogues to provide multiple subject entry points for a single document. And facilitating user curiosity in multiple directions would have been even more challenging. It is possible that the sort of approach recommended here might complement classificatory approaches developed during an age of card catalogues. Or perhaps the digital imperative will lead us to the most significant change in classificatory practice for well over a century. Cutter and Dewey may not have foreseen the staying power of their classifications;

we may well be at another historical turning point for KOS; how well we handle it will reverberate throughout society.

### Concluding Remarks

We began by considering the productivity impact of KOS. This discussion in turn suggested several ways in which productivity might be enhanced. All of these point toward a synthetic approach grounded in basic concepts (rather than disciplines). This would allow users of all types – disciplinary and interdisciplinary, general and scholarly – to better find what they are looking for in any one database. It would also enhance their ability to search across databases and, especially, across libraries, archives, museums and art galleries. Importantly it will aid users in finding not only what they know to look for but in identifying related information from other fields of which they were previously unaware. This will enhance the rate not only of scholarly discovery but also of technological innovation, entrepreneurial activity, artistic exploration and even public policy improvement.

I suspect that the gains in each of these last respects will be large. But even if the field of knowledge organization can enhance the rate of human discovery and creativity just a little it will still be the most important field in the entire scholarly enterprise. ■

### Resources Mentioned in the Article

- [1] Solow, R. (July 12, 1987). We'd better watch out. *New York Times Book Review*.
- [2] Brynjolfsson, E. (1993). The productivity paradox of information technology. *Communications of the ACM*, 36(12), 66–77.
- [3] Hart, G., & Dolbear, C. (2013). *Linked data: A geographic perspective*. Boca Raton, FL: CRC Press.
- [4] Szostak, R. (2013). Advances in classification research online 2013: Classification, ontology and the Semantic Web. *ASIST SIG/CR Classification Workshop*, 24, 30–37. Retrieved February 26, 2014, from <http://journals.lib.washington.edu/index.php/acro/article/view/14674>
- [5] Clarke, S., & Dextre, G. (2011). ISO25964: A standard in support of KOS interoperability. In A. Gilchrist & J. Vernau (Eds.), *Facets of knowledge organization: Proceedings of the ISKO UK Second Biennial Conference, July 4th–5th, 2011, London* (pp. 129–33). Bingley, UK: Emerald.
- [6] Szostak, R. (2011). Complex concepts into basic concepts. *Journal of the American Society for Information Science and Technology*, 62, 2247–65.
- [7] Swanson, D. R., Smalheiser, N. R., & Bookstein, A. (2001). Information discovery from complementary literatures: Categorizing viruses as potential weapons. *Journal of the American Society for Information Science and Technology*, 52, 797–812.

*Continued on following page*

SZOSTAK, continued

### Resources Mentioned in the Article, cont.

- [8] Olson, H. (2007). How we construct subjects: A feminist analysis. *Library Trends*, 56(2), 509-41.
- [9] Palmer, C. L. (1996). Information work at the boundaries of science: Linking library services to research practices. *Library Trends*, 45(2), 165-191.
- [10] Szostak, R. (2013). *Basic concepts classification*. [Web version 2013]. Retrieved February 1, 2014, from <https://sites.google.com/a/ualberta.ca/rick-szostak/research/basic-concepts-classification-web-version-2013>
- [11] León Manifesto. (2007). *Knowledge Organization*, 34(1), 6-8. Retrieved February 1, 2014, from [www.iskoi.org/ilc/leon.php](http://www.iskoi.org/ilc/leon.php)
- [12] Mazzocchi, F., Tiberi, M., De Santis, B., & Plini, P. (2007). Relational semantics in thesauri: Some remarks at theoretical and practical levels. *Knowledge Organization*, 34(4), 197-214.
- [13] Szostak, R. (2012). Classifying relationships. *Knowledge Organization*, 39(3), 165-78.
- [14] DCMI International Conference on Dublin Core and Metadata Applications. (n.d.) Metadata intersections: Bridging the archipelago of cultural memory: Call for papers. Retrieved February 1, 2014, from <http://dcevents.dublincore.org/IntConf/dc-2014/schedConf/cfp>
- [15] Ørom, A. (2003). Knowledge organization in the domain of art studies: History, transition and conceptual changes. *Knowledge Organization*, 30(3-4), 128-143.
- [16] Shiri, A. (2012). *Powering search: The role of thesauri in new information environments*. (ASIS&T Monograph Series). Medford, NJ: Published on behalf of the American Society for Information Science and Technology by Information Today, Inc.