Lambe, Patrick, "Changing the Conduct of Science in the Information Age: Discussion Points"

Changing the Conduct of Science in the Information Age

Discussion points by Patrick Lambe, Adjunct Professor Hong Kong Polytechnic University and Principal of Straits Knowledge, Singapore.

The Role of Knowledge Organisation Systems in the Conduct and Advancement of Science

To understand – and influence – how science grows and develops, it is also necessary to:

- have consistent ways of describing science,
- maintain a conspectus of the relationships between different areas of scientific knowledge, and
- maintain continuity between past (science memory), current (science activity) and emerging ways (new knowledge creation) of describing science.

Taxonomies and formal knowledge organization systems play a sophisticated role in delivering these capabilities, but this role is often poorly or partially understood.

When people think about taxonomies, they often think of them as subject vocabularies or as fixed hierarchical structures that show how a subject should be organised. In fact, taxonomies are only one element in what are called Knowledge Organisation Systems (KOS), and these turn out to also be critical to the growth and development of scientific knowledge.

A KOS performs three critical functions which are relevant to the development and progress of science.

- It standardizes language, which enables coordination and knowledge-building around shared language and the entities described by that language
- It identifies connections or relationships between different areas of knowledge in predictable, commonly understood ways
- It overlays salient and useful structures onto a diffuse knowledge domain, which enables sensemaking to occur on significant patterns and relationships within the knowledge domain, including identification of gaps in knowledge, and enabling testable hypotheses to be made.

A KOS is able to do these three things because it combines the ability to work with **lexical** characteristics, identify salient **relationships** between entities, and support **visual representation** of an entire knowledge domain. To associate a KOS simply with one of these characteristics at a time and to miss the others, is to miss its value for knowledge organization in support of new knowledge creation.

Let's take a couple of famous illustrations from the history of science.

Carl Linnaeus

Throughout the fifteenth century, with the spreading of wealth through trade and the growth of scholarship, the passion for collecting "curiosities" was taken up on a large scale by scholars and scientists across Europe, and their collections were increasingly used as instruments of learning about the natural world. Arrangements of curiosities became part of a larger endeavour to construct a systematic knowledge of the natural world. Collections started to become more systematic and supportive of enquiry, sensemaking and discovery.

These were the seeds of modern empirical science. By the beginning of the seventeenth century, however, writers like Francis Bacon were thoroughly dismissive of the higgledy-piggledy arrangements of the rich and famous:

"There is such a multitude and host as it were of particular objects, and lying so widely dispersed, as to distract and confuse the understanding; and we can therefore hope for no advantage … unless we put its forces in due order and array by means of proper, and well arranged, and as it were living tables of discovery of these matters which are the subject of investigation…"

Bacon's impatience was echoed just over a century later by the methodical biologist Carl Linnaeus who was dismissive of the "complete disorder" he found in the home of the last great universal collector of his time, Sir Hans Sloane – founder of the collection that became the British Museum. After Sloane, in fact, collectors divided themselves into discrete disciplines. The world of knowledge had become too complex to comprehend and represent in one single arrangement.

In the midst of this complexity, Linnaeus' great gift to science was threefold. Beginning with his *Systema Natura* in 1735, he introduced a far simpler principle of distinguishing between species based on anatomical observation than had ever been proposed before. Beginning in 1737 with his *Critica Botanica* he laid down the rules for his binomial naming system for species which riled his critics immensely (because he substituted so many older naming conventions with his own), but when widely adopted created the first standardized way of describing species. This immeasurably enhanced scientific coordination and collaboration.

Finally, his hierarchical, nested classification tree structure turned out to be a perfect vehicle to express the genealogical relationships that gained such prominence during the emerging evolutionary theories of the late eighteenth and early nineteenth centuries.

Linnaeus' new taxonomic method simplified the task of categorization, imposed rigorous rules (and therefore consistency), and happened on a form of representation that history turned into a lucky bet. From the point of view of advancing scientific method, his focus on analysis, rules and standardized approaches, gave an incalculable advantage.

We can see in Linnaeus' taxonomy design two of the three elements of a

KOS – lexical stabilization to enable coordination between scientists, and a meaningful structure (a hierarchical rule-based tree structure) to establish predictable and (as it turned out from subsequent science) salient relationships between the entities being described.

Dmitri Mendeleev

Dmitri Mendeleev's periodic table of elements was an attempt to figure out patterns of behaviour across chemical elements. His endeavour was essentially a sensemaking endeavour illustrating the third function of a KOS – he was playing with the organization of the elements to see if he could explain deviations, simplify, understand and explain the relationships between them.

Mendeleev used a different taxonomy structure, not the classical hierarchy associated with Linnaeus. He used the matrix structure, where the entities are arranged according to their properties along two dimensions –he arranged the elements in columns by similarity of properties and horizontally by regular patterns of behaviour or periodicity. Like Linnaeus, he happened upon a salient and useful way of organizing before the underlying science behind his arrangement had been uncovered – electron structures had not yet been identified.

Arranging the elements in this way did two interesting things for science. First, it helped to make sense of the "periodicity" of elements – where elements exhibit similar properties at regular intervals of atomic mass increase. Secondly, representing the elements in a matrix display enabled scientists to identify gaps in the table where elements that were previously unknown should exist.

Hence the KOS helped explain behaviours and gave predictive power by identifying new elements that scientists could hunt for – and were subsequently discovered or manufactured in the laboratory – simply because their "place" in the taxonomy was visibly unfilled. Discovering and displaying the periodicity of behaviour through organizing by mass and electron structure allowed scientists to predict the existence of new elements – essentially to create new knowledge.

This by the way turns out to be a strong feature of matrix representations for taxonomies. They are extremely useful for sensemaking as well as for new knowledge creation or discovery.

Linnaeus and Mendeleev created knowledge organisation systems and standardised scientific languages to enable greater coordination, inter-connection and sensemaking across their respective scientific communities.

The elements of a KOS

A KOS can have three different orders of complexity. As science becomes more complex and inter-related, the complexity of the needed KOS increases:

(a) At the most basic level are **controlled vocabularies**, with principles

for recognition, inclusion and exclusion, which provide a common reference language for describing science and enabling coordination.

(b) Next in order of complexity are **taxonomies** which put structure around the controlled vocabularies (along with principles for how those structures are maintained), and which enable sensemaking, identification of gaps, and inter-relationships among areas of science.

(c) As scientific knowledge becomes even more complex, taxonomies can no longer represent all of the salient kinds of relationships within a single comprehensible structure. We need ways of visualizing different patterns of relationships across multiple domains. **Ontologies** are systems of taxonomies, where relationships are also defined across different taxonomies, taxonomy elements and vocabularies. They enable large scale pattern-sensing and sophisticated interpretation filters on a complex scientific activity landscape.

(d) Finally, a knowledge organisation system requires mechanisms for detecting and recognising new language, new usages and new relationships between areas of science. This is essential to keeping the KOS vocabularies, taxonomies and ontologies current and reflective of current and emerging reality. The maturing field of **topic maps** based on semantic analysis is an important example of such a mechanism.

Principle 1: the complexity of a KOS needs to match the complexity of the domain it attempts to describe, and the complexity of the coordination, connection and sensemaking work it needs to support.

Human factors in using KOS

Modern science is now too fluid and complex to be supported by simpler KOS's such as controlled vocabularies and taxonomies. This is why keyword or topic-based approaches, or single taxonomy approaches to the description and measurement of science have inherent limitations by themselves. Any controlled vocabularies in use, and any taxonomy systems in use, really need the richer environment of ontologies behind them, to perform the sensemaking, memory and coordination functions that a KOS should properly provide for the complex and shifting landscape of science.

One of the drawbacks with ontologies however is that machines find it much easier to navigate and process the information from ontologies than humans do. Humans have significant cognitive constraints in terms of attention, memory span and tracking relationships, which means that they are much more suited to navigating and processing individual taxonomies than multi-dimensional ontologies.

This has implications for the human users of a KOS who tend to favor simpler lexical work (eg keywords or topic words) or simplistic taxonomy structures over investment in the information enrichment required to support ontologies. Actors such as publishers, authors, audiences,

scientists, science administrators, funders, analysts, policy makers, all require human-scale representations of scientific knowledge – and this means at the vocabulary level, or at the taxonomy level.

If ontologies are to support the human actors in the science landscape, ontologies require context-sensitive human interfaces to create intelligible representations that are meaningful to their respective audiences, but still provide those functions of standardization of language, meaningful connections of content (including from past to future), and sensemaking capability. Vocabularies need to be connected to taxonomies, and taxonomies need to be connected to ontologies.

Principle 2: when the complexity of the KOS exceeds human cognitive capabilities, designed interfaces using taxonomies are necessary to serve the working needs of users in their own normal working contexts.

Humans also resist lexical control, especially if the controlled language is not natural to their own context.

The typical managerial response to the human aversion to working with – and contributing to – a complex KOS in a disciplined and consistent way, is to use semantic technologies to analyse natural or semi-controlled language texts and to make inferences about topics and relationships between topics to feed the ontology-supported approach. These technologies have great potential for sidestepping human aversion to control and consistency, and they are also very powerful for identifying emerging trends in science – too much control suppresses new or variant language about science, and so suppresses signals of new science. Semantic technologies can also infer relationships between concepts, based on association patterns.

However, to perform the larger functions of coordination of language, meaningful connections and sensemaking in support of science, human intervention is required to judge and identify the most salient relationships, and to establish connections between domains as well as between past and future science language.

Principle 3: it is not sufficient to use semantic technology to describe science activity. This does not get at all the functions of a KOS. Linnaeus and Mendeleev had the impact they had, because they engaged in a work of design, not simply description.

In practice in today's world, the task is no longer within the grasp of gifted and determined individuals such as Linnaeus and Mendeleev. We require institutional interventions, in the form of development and maintenance of standardised vocabularies, taxonomies and ontologies, and the environments where they can be deployed.

Any KOS intended to meet the needs of understanding and progressing science will require some elements of designed structure and the

disciplined application of human design. Otherwise we end up with naturalistic representations of current trends which are unmoored from broader perspectives on science, and which fail to connect trends and developments with scientific memory, or "faster" knowledge developments with the "slower" and more stable core of science description and measurement.

Science as a social system

Semantic technologies have another drawback, which is that they work best on reasonably well-structured textual content (eg scientific papers, proposals to a set format, funding and administrative records, project reports, patents) within a well-defined "language community" – eg scientists working within a given discipline, who already share, to a large extent, a common language. More advanced sensemaking capabilities of a KOS, eg seeing what is missing, cannot easily be served by this.

Hans Pfeiffenberger, Peter Elias and Cameron Neylon have all pointed in their white papers to scientific work which is "off the books" of the formal documentation of science – whether it be science contributions by non-researchers, participation in large-scale science infrastructure, or behind the scenes participation in science work.

Diana Crane pointed out almost forty years ago (*Invisible College: diffusion of knowledge in scientific communities*) that a significant portion of scientific work and validation is in fact "invisible" – and the visible manifestations of science conceal an intricate social network of relationships, trust and perceived authority, underlying how science gets funded, how scientists decide to collaborate, and how new knowledge gets validated. At face value, the application of semantic technologies holds little visible promise for describing and understanding this kind of invisible or "off the books" scientific activity.

Publication and citation activity is most relevant to early career scientists. Mid to mature career scientists develop other skills which are not so easily tracked: their ability to win funding through their ability to conceptualise requirements for funding sponsors both private and public; their track record in generating tangible outputs such as new conceptual tools or solutions; their ability to attract good students and collaborators; their participation in agenda-setting panels and meetings, many of them not transparent to the visible domain of publications or institutional records.

Publication activity in mid career scientists can in fact conceal lack of progress in science – as one senior scientist put it to me "It's perfectly possible to spend your career and earn a living generating a publications trail simply by rearranging the furniture using one base algorithm or insight and not making any real progress at all."

In whole areas of science patents are considered inappropriate ways of protecting new knowledge for exploitation, either because they represent new tools or solutions without specific defined purpose, or because their exploitation from a funders' point of view (both government and private) requires them to be treated as trade secrets and protected know-how.

Principle 4: a KOS that effectively supports the conduct of science must be able to observe informal social activity and relationships beyond the boundaries of traditional formal outputs and records of science activity.

Making invisible work visible

There are promising approaches from other domains which recognize and exploit the social dimension of knowledge creation. The US military also has to meet challenges in connecting "faster" and "slower" streams of knowledge, particularly in capturing lessons learned from combat mission experiences, and connecting these lessons with the much slower moving bodies of Army doctrine.

In combat zones such as Afghanistan and Iraq, the tactics of insurgents adapt constantly, and the language used to describe new dangers and risks is also constantly changing. Formal knowledge description and codification systems such as the Army Lessons Learned knowledgebase and doctrine manuals cannot recognise and incorporate this fast-moving knowledge quickly enough for personnel requirements in the field of operations. Hence to the formal knowledge systems of the Army, there is also a domain of "invisible" work which somehow needs to be connected to Army knowledge in a managed way.

Company Command is the name of an initiative started informally in the early 2000s by a group of US Army company commanders to enable and scale informal sharing between company commanders in combat zones via bulletin boards and a Web 2.0 style collaboration site. The two founders of the site, Nate Allen and Tony Burgess, said that they wanted to recreate in an online platform the end of day front porch conversations they themselves used to have about their professional practice.

The Company Command site turned out to serve an immediate need in Afghanistan and Iraq, because it was much better at picking up and disseminating fast-moving knowledge about insurgency tactics (such as new methods of laying IEDs) than the formal knowledge and learning systems of the Army. Quality was recognized as provisional, and validation systems were very simple; however, this was a peer-to-peer network, where people knew each other socially or by reputation, so validation was "good enough" for immediate use, while the formal systems weighed and discriminated lessons more systematically.

This informal, peer-to-peer professional sharing initially started on a password protected internet site, but its value (and the security risks it posed) was quickly recognized and it was incorporated into the military network. Now the US Army is taking lessons from this experience and increasingly experimenting with Web 2.0 collaboration tools to provide more channels for the informal and previously invisible knowledge sharing and knowledge creation activity among its officers and men.

Connecting fast knowledge to slow knowledge

The challenge still remains of how to connect this informal, socially driven content, now rendered visible, to the more formal knowledge systems of the Army. To think of this in KOS terms, we use the metaphor of a street, a department store, and a warehouse.

The street is the place where people maintain social and situational awareness of what is going on around them. This is the place where you can see the latest fashions and fads, catch the latest news headlines, and calibrate yourself with your social peers. In knowledge terms, this is the place of current awareness, who is doing what, social interactions, and faster moving knowledge, much of it ephemeral, but some of it providing signals of emerging trends. The vocabularies used here are uncontrolled, but can be sampled and analysed for significant new patterns.

The department store has windows onto the street for passersby to view its wares. But inside, it is organized deliberately to enable shoppers to find collections of related content. It is organized into departments suited to specific kinds of audience. In KOS terms, this is the area of formal knowledge arrangements using taxonomies designed for specific groups and their needs.

The warehouse contains all the stocks of knowledge on display in the department stores, organized and tagged for multiple reuse in many different stores, and in multiple possible arrangements. In KOS terms, this is the area of ontologies, capable of generating different arrangements and visualizations of content.

Connecting the street, department store and warehouse means having the ability to analyse and learn from emerging patterns on the street (social, collaborative spaces reflecting informal conversations about work practices with uncontrolled user-driven vocabularies), and then to incorporate new terms and relationships between terms into the ontology-driven warehouse, and thence into new arrangements of content for the department store windows and internal store arrangements.

In creating environments for informal knowledge sharing that leverage existing peer relationships and natural patterns of social interaction and reputation building, the US Army has brought conversations into a place where language can be mined for insights, and fed into the KOS ontology and taxonomies. We can make a case that the same mechanism needs to be employed within the domain of science.

Principle 5: a KOS that effectively supports the conduct of science must be able to observe and connect formal and informal activity streams, using designed taxonomy structures as 'human-oriented middleware' between emerging new language and existing ontologies.