Knowledge organization systems as enablers to the conduct of science

Patrick Lambe
Straits Knowledge, Singapore

Abstract

The sophistication of knowledge organization systems (KOS) has evolved rapidly over the past thirty years, largely driven by information technology innovations. Two key assumptions have been a) that KOS-work is the preserve of information professionals acting as skilled intermediaries, and b) that it is largely focused on enabling the finding and discovery of information. This paper challenges both assumptions with reference to the conduct of science in the 21st century, by describing the ways in which access to KOS skills and tools is already broadening beyond information professionals to scientists, and by describing how knowledge organization systems enable sense-making of trends within science and new knowledge creation, beyond simple access and discovery roles. It closes with remarks on the implications for information professionals engaged in KOS-related work.

Introduction: a new role for knowledge organization systems

In 2008 Brian Vickery wrote a survey article 'On knowledge organization' which to my mind is the clearest and most succinct survey of the nature, structure and functions of knowledge organization systems (KOS) I have yet seen (Vickery, 2008). Vickery takes a developmental view, identifying four stages in the evolution of KOS:

- Pre-coordinate era: KOS are static designed structures such as card catalogues or indexes.
- Post-coordinate era: KOS consist of concepts that can be assigned to entities, and combined dynamically in search queries; semantic relationships can be assigned through facets, and relationships between terms defined in thesauri.
- Internet era: hierarchy- and facet-enabled browsing and filtering of content; navigation and discovery through hyperlinks; search via automated indexing of the contents of documents.
- Semantic Web era (still emerging): ontology-enabled search and discovery, where semantic relations between terms are provided to machines to enable meaningful deduction of the content of a document.

He closes by discussing the challenges of updating structured KOS in the ever-growing expansion of information availability, and the apparent need to conceal the ever-growing sophistication of computer handling of documents and their semantic content and to provide a simple user-interface. He concludes that in spite of the rapid and significant evolution of KOS, we are still 'at the start of a long period of experimentation and development in the evolution of knowledge organization systems' (Vickery 2008).

Notwithstanding his perceptive observations about the likely future evolution of KOS, Vickery begins his article with an assertion about the work of knowledge organisation that I believe to be no longer true:

> 'To organise knowledge is to gather together what we know into a comprehensive organised structure, to show its parts and their relationships. This is the work of scholars and encyclopaedists. It is not the role of the information profession. Our tasks are to make knowledge (whether organised or unorganised) available to those who seek it, to store it in an accessible way, and to provide tools and procedures that make it easier for people to find what they seek in those stores.' (Vickery, 2008,para. 1).

I believe two significant changes have occurred to make Vickery's position overly limiting for those who work in the field of knowledge organization. These changes have been happening for a while, but they became most visible when we made the transition between Vickery's post-coordinate era and the Internet era.

Assumption 1: There should be a separation between scholars and information professionals

The role distinctions between a) creators and users of knowledge, and b) the information professions that have traditionally supported them through KOS, have blurred and in many cases broken down.

Knowledge creators require much greater levels of information literacy simply to navigate their increasingly complex knowledge spaces and do their work, and they need many of the KOS capabilities traditionally reserved to the information professions.

Conversely, domain-agnostic information professionals can increasingly rarely afford to stand over against the knowledge domains they support, wielding a standardised set of KOS tools and approaches. We are increasingly forced to steep ourselves in the disciplines and domains that we support, and to develop specialized instruments designed to navigate their contours. To take the sciences as an example, information professionals need to become scientists and scientists need to be transfused with KOS skills and capabilities.

Assumption 2: The main purpose of a KOS is to support search and discovery

The role of a KOS is no longer confined to the support of search and discovery alone. The work of classification, of building taxonomies and of designing ever more sophisticated KOS, is integral to the work of constructing knowledge itself (Bowker and Star, 1999). Classification is firstly a sense-making activity that places interpretation and meaning upon the world; it solidifies into conceptual vocabularies with which we can manipulate the world (Lambe 2007).

In describing a variety of science-related activities, scientist Wolff-Michael Roth explains it thus:

> 'Karen, my students, the workers at the fish hatchery, and I all engaged in classification without thinking about it as such or making it an object of reflection. In fact, we no longer distinguished the classification schemes and the names associated with them from the things that we classify. The schemes and category names have

> become transparent, allowing us to directly deal with things in the world. Categories and the things they classify have become conflated – there is no longer a distinction between maps and the territories they describe.' (Roth, 2005, p.583).

To conduct their work more effectively, to question their assumptions, to innovate across disciplinary boundaries, and to create new ways of working, scholars need to become more aware of the knowledge organization work they do, and they need to be able to design and adapt their knowledge organization practices to support their knowledge goals.

James Evans and Jacob Foster have recently described this ability to reap insight from reflection on how knowledge is used and organized, as 'metaknowledge':

> 'Using informatics archives spanning the scientific process, from data and preprints to publications and citations, researchers can now track knowledge claims across topics, tools, outcomes, and institutions. Such investigations yield metaknowledge about the explicit content of science, but also expose implicit content—beliefs, preferences, and research strategies that shape the direction, pace, and substance of scientific discovery.' (Evans and Foster, 2011).

No wonder then that a recent National Science Foundation workshop on *Changing the Conduct of Science in the Information Age* had as one of its highest priority recommendations to 'foster formal and informal training to develop scientists' skills in knowledge and data access.' (National Science Foundation, 2011).

This returns us to the blurring of boundaries between information professionals and scholars.

In this paper I wish to illustrate the breakdown of these assumptions about the role of information professionals and the scope of knowledge organization work, by describing the emerging role of KOS in the conduct of 21st century science.

I will first look at the role of knowledge organization work in the history of science, and then extract some principles about knowledge organization applied to the more complex and fast moving conduct of science in the 21st century. I will then describe how these principles are being applied practically in the US Federal Government's management of science policy and science funding. I will close with some remarks about the implications for information professionals.

Part 1: KOS in the conduct 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;
- maintain continuity between past (science memory), current (science activity) and emerging ways (new knowledge creation) of describing science.

Taxonomies and formal KOS play a sophisticated role in delivering these capabilities, but this role is often poorly or partially understood. In fact, KOS turn out to 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 sense-making 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 (Lambe, 2007). 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, extracted from my book *Organising Knowledge* (Lambe, 2007*)*.

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, sense-making 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…' (Blom, 2003, p 46).

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. (Lambe, 2007).

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 sense-making 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 sense-making as well as for new knowledge creation or discovery. (Lambe, 2007).

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

Part 2: Five principles of knowledge organization to support the conduct of science

The elements of a KOS for supporting the conduct of science

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 sense-making, 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 organization 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 of science texts, is an important example of such a mechanism.

Perhaps our mandate for the role of KOS beyond search and discovery can be taken from a recent report to the National Science and Technology Council: 'The ability to achieve innovation in a competitive global information society hinges on the capability to swiftly and reliably find, understand, share, and apply complex information from widely distributed sources for discovery, progress, and productivity.' (Interagency Working Group on Digital Data, 2009).

'Finding' is only the first verb in the series: an effective KOS also needs to be able to support understanding (by bringing contextual associations and clarifying relationships), sharing (by identifying other parties to whom this information might be relevant, or who may have knowledge to add to it), and application (by delivering it in a form that can be combined and used easily).

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 sense-making work it needs to support.

Human factors in using KOS

Modern science is now too fluid and complex to be supported by simpler KOS 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 each have inherent limitations when used on their own. Any controlled vocabularies in use, and any taxonomy systems in use, really need the richer environment of ontologies behind them, to perform the sense-making, 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 (Lambe 2007).

This has implications for the human users of a KOS who tend to favor simpler lexical work (for example 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 sense-making 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 sense-making 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 (eg through topic maps derived from patent descriptions) which are unmoored from broader perspectives on science, and which fail to connect trends and developments with scientific memory, or to connect 'faster' knowledge developments with the 'slower' and more stable core of science description and measurement.

Understanding science as a social system broadens the scope of a KOS beyond the formal publications of science

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 sense-making capabilities of a KOS, eg seeing what is missing, cannot easily be served by this. Semantic technologies depend on – and reflect – the known, they shed little light on what is to be discovered or created.

In a recent National Science Foundation workshop on *Changing the Conduct of Science in the Information Age,* Hans Pfeiffenberger, Peter Elias and Cameron Neylon all pointed to scientific work which is 'off the books' of the formal documentation of science in journals, conferences and patents – whether it be (Pfeiffenberger, 2010; Elias, 2010; Neylon, 2010):

- science contributions by non-researchers (eg public participation in large scale science projects such as the Sloan Digital Sky Survey, which is driven by data from almost a million 'citizen astronomers');
- participation in large-scale science infrastructure in roles that look very much like administrative or managerial roles, but on which nevertheless the conduct of science depends; or
- 'behind the scenes' participation in science work (eg in highly skilled technical support roles).

This is not new: Diana Crane pointed out almost forty years ago that a significant portion of scientific work and its 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, what they decide to collaborate on, and how new knowledge gets validated (Crane, 1972). In fact, there is evidence that citation patterns in published science articles bear little resemblance to the actual reading patterns of scientists in the conduct of their research (Bollen et al, 2009). Published science does not necessarily reflect how science actually gets done.

At face value, the application of semantic technologies to the published literature of science holds little visible promise for describing and understanding this kind of invisible or 'off the books' scientific activity. Yet to shape the conduct of science from a policy point of view in directing funding, or from a scientist's point of view in choosing a productive area for research, this kind of visibility into the 'real' topography of science activity is increasingly important.

Moreover, the formal public literature of science, while still its most visible representation, is not the most relevant content to represent the full span of work of a scientist. Publication and citation activity is most relevant to early career scientists. 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" eg using one base algorithm or insight in new combinations and not making any real progress at all.'

Within 'serious' science, mid to mature career scientists develop other skills and activity which are not so easily tracked, but which are critical for the conduct of science:

- 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.

As for patents, there are whole areas of science where 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 science funders' point of view (whether government or private) requires them to be treated as trade or military secrets and protected know-how.

The use of social media platforms such as blogging, micro-blogging ('tweeting'), wiki-collaboration, open source publishing and peer review, bookmark sharing, data-sharing, and tagging activity (any blend of controlled, uncontrolled, 'farmed' or guided tagging adds a layer of meaning and organization potential), are much more suitable platforms to render a form of visibility for this kind of 'invisible' science activity. But the less formally controlled an information channel is, the harder it is for semantic technologies to fully represent them.

The ability to trace, map and interpret social networks turns out to be an important way of representing the topography of a knowledge domain, in particular to identify potential 'hotspots' of collaboration that signal emerging new knowledge. This kind of representation has not hitherto been considered a part of traditional KOS-related work (although it has been applied to the study of knowledge flows) (Cross and Parker, 2004). The study of information visualization in general becomes an important component in the KOS toolkit.

Information professionals cannot stand aloof from the practice of science and focus simply on the published literature of science in order to support it with knowledge organization work. They need to be cultivating a deep familiarity with its activities and challenges, in order to be able to map and represent it in novel and productive ways.

Principle 4: a KOS that effectively supports the conduct of science must be able to identify, observe and represent 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 that somehow needs to be connected to the formal Army knowledgebase 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. Two of the 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 (Dixon et al, 2005).

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 improvised explosive devices [IEDs]) than the formal knowledge and learning systems of the Army. The quality of these field-based lessons learned 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 and more slowly for incorporation into official Army doctrine, tactics, techniques and procedures. Lives were saved.

This informal, peer-to-peer professional sharing initially started on a password protected internet site, but its value, as well as the security risks it posed, were 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 soldiers.

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 often uncontrolled, but can be sampled and analysed for significant new patterns which need to be incorporated into the formal knowledge organization system. They can also be guided by suggesting previously used tags in an auto-suggest function as people begin to type their keywords.

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/guided 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 and visualization tools as ‘human-oriented middleware’ between emerging new language and concept usage, and existing ontologies.

What emerges is a picture of an interlocking ecosystem of knowledge organization tools and practices, both formal and informal, both tightly structured and loosely structured, both designed for specific audiences/uses and abstracted for universal application, each performing a distinct role. None of these tools and practices can adequately serve the needs

of science on their own, but together they provide mutual support for the diversity of needs faced by scientists, policy makers, funders and science administrators.

To get a sense of how this might work in practice, we turn now to an initiative by the US White House's Office of Science and Technology Policy, originally designed to measure the economic impact of the economic stimulus funds applied to science funding, STAR METRICS (Science and Technology in America's Reinvestment — Measuring the Effects of Research on Innovation, Competitiveness and Science).

Part 3: Star metrics and the science of science policy in the United States

Measuring the relationship between investments and outcomes of science is a tricky business. The National Science Foundation (which with the National Institutes of Health is driving the STAR METRICS project) has traditionally tried to track investments and apportionment to fields of science through surveys, but these are difficult to administer, time consuming to process and analyse, and it is hard to match investments to outcomes in a consistent way.

By contrast, the STAR METRICS project 'aims to match data from institutional administrative records with those on outcomes such as patents, publications and citations, to compile accomplishments achieved by federally funded investigators.' (Lane, 2010). The mining of existing administrative data has been shown in pilot projects to substantially cut the amount of time and effort involved in collecting this information.

The need to match inputs with outcomes requires the building of a sophisticated data infrastructure, and linking diverse data sources such as administrative records, researcher profiles, funding proposals, patents, publications and citation databases, etc in a common environment.

To achieve this, STAR METRICS has almost unwittingly found itself effectively building a KOS for many different aspects of the conduct of science, and this has seen a surge of interest from many of the players involved in science in the USA. The reason for this is plain: the data infrastructure that will serve as a mechanism for measuring the impact of Federal investment in science and technology will also pay off in many other ways.

Linking profiles of researchers with the products of their research and providing connections with the investment data from administrative records has wider benefits than just showing impact:

- funders and legislators will be able to judge the impact of their investment decisions on a national scale with much greater clarity;
- policymakers will be able to judge both impact of funding with greater sense of accuracy, and identify important trends in scientific research; topic models will be able to identify science activity 'hotspots' enabling the earlier identification of emerging new research areas;
- science publishers will have more channels for exposing their content when it is linked to the relevant researchers' profiles and to collaboration platforms;

- open access science publishers will have much greater 'mainstream' visibility and scientists participating in it will be able to present integrated portfolios of work associated with their profiles;

- standardized and validated profiles for researchers can be referenced and reused in funding applications and reports to funding agencies, substantially reducing their administrative burden; researchers will be able to identify potential collaborators and 'hot' research areas with greater ease;

- universities will have much greater visibility into how they are performing in science research, they will be able to track economic impact of their research within their own states, the administrative burden of reporting on their use of Federal funds will be substantially reduced, and benchmarking against other institutions will be much easier.

Julia Lane, who co-leads the STAR METRICS project with Stefano Bertuzzi from NIH expects that when researcher profiles are linked to publications and online collaboration platforms 'the researchers' use of the Internet to communicate and publish will enable STAR METRICS to track the creation and transfer of knowledge properly for the first time.' (Macilwain, 2010).

There are, of course, multiple challenges in developing such a data infrastructure, and notwithstanding the high levels of interest in STAR METRICS, it is still early days. Key data sources are scattered across research institutions, federal agencies, and third party databases, they are not in standardized formats, or systematically shared (Lane and Bertuzzi, 2011).

KOS tools are recognized as an essential part of this data infrastructure, whether they be topic modeling and text-mining tools, or taxonomies and ontologies which provide definitions for 'a set of relations between different areas of scientific knowledge and the maintenance of continuity between past, current, and emerging ways of describing science.' (Lane and Bertuzzi, 2011, pp.679, 680).

Whatever the challenges, the STAR METRICS example provides an exceptionally clear view into the agenda for large scale knowledge organization systems in the 21st century, systems that can have broad societal and economic impact. We in the information professions have to raise the level of our game to deal with these challenges.

Conclusion: implication for the information professions

There is an interesting implication from all of this for information professionals now engaged in knowledge organization work.

After decades of worrying about the marginalization of our profession through the computerization of information provision and the automation of many of our traditional roles, the transition into the Internet and Semantic Web eras have given us a new prominence as our knowledge organization skills and tools have been 'discovered' and used to support the deployment of discovery and search technologies.

Despite this new (and somewhat unfamiliar) popularity for our profession, we must admit that most of the innovation in KOS has been driven externally by technologists, not by the information professions themselves. We have often responded to the challenges in very interesting and creative ways, but we have not been a driving force.

If we remain in a passive role, if we remain bound to a parochial 'separatist' view of our role, and if we limit ourselves to a view of our work as primarily supporting information search and discovery, we risk becoming marginal once more when our tools, methods and special skill sets have been absorbed and the technology moves on. Relevance is about understanding the reach and relevance of our work beyond our traditional roles and continually reinventing our roles.

Our close look at the KOS requirements for the conduct of science shows that we can no longer stand separated from the domains that we enable, we will need to become as much involved in skills and tools transfer as we were previously involved in the application of those skills and tools, and we have to re-imagine the scope and breadth of our work beyond the published canon, and beyond information search and discovery alone.

Acknowledgements

An early version of this paper was written for the National Science Foundation (NSF) workshop *Changing the conduct of science in the information age* held in November 2010, and I would like to thank the NSF for inviting me to participate in that workshop. In particular I would like to thank Julia Lane and Jeri Mulrow of the NSF, Jeff Alexander of SRI International, and Professor Shaogang Gong of the University of London, with whom several of the insights in this paper were developed.

References

Blom, P. (2003) *To have and to hold: an intimate history of collectors and collecting*. New York: Overlook Press.

Bollen, J., et al. (2009) Clickstream data yields high-resolution maps of science, *PLoS ONE*, 4(3).

Bowker, G. C., & Star, S.L. (1999) *Sorting things out: classification and its consequences*. Cambridge, MA: MIT Press.

Crane, D. (1972) *The invisible college: diffusion of knowledge in scientific communities*. Chicago: University of Chicago Press.

Cross, R. L., & Parker, A. (2004) *The hidden power of social networks: how work really gets done in organizations.* Cambridge, MA: Harvard Business School Press.

Dixon, N. M., et al (2005) *Company Command: unleashing the power of the army profession*. West Point, NY: Center for the Advancement of Leader Development and Organizational Learning.

Elias, P. (2010) Digital technology and the conduct of scientific research. [white paper for workshop] In *Changing the Conduct of Science in the Information Age*. Arlington, VA: National Science Foundation.

Evans, J. A., & Foster, J. G. Metaknowledge, *Science*, 331 (6018), 721-725.

Interagency Working Group on Digital Data (2009). *Harnessing the power of digital data for science and society: report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council*. Washington, DC: National Science and Technology Council.

Lambe, P. (2007*) Organising knowledge: taxonomies, knowledge and organizational effectiveness*. Oxford: Chandos.

Lane, J. (2010) Let's make science metrics more scientific. *Nature,* 464 (March), 488-9

Lane, J., & Bertuzzi, S. (2011) Measuring the results of science investments. *Science* , 331 (6018), 678-80

Macilwain, C. (2010) What science is really worth. *Nature*, 465 (June), 682-4

National Science Foundation (2011). *Changing the conduct of science in the information age: summary report of workshop held on November 12, 2010*. Arlington, VA: National Science Foundation.

Neylon, C. (2010) Attribution and identity for researchers and research objects [white paper for workshop] In *Changing the Conduct of Science in the Information Age.* Arlington, VA: National Science Foundation.

Pfeiffenberger, H. (2010) 'Focusing on social constructs' white paper for workshop on *Changing the Conduct of Science in the Information Age* (Arlington, VA: National Science Foundation)

Roth, W-M. (2005) Making classifications (at) work: ordering practices in science. *Social Studies of Science*, 35(4), 581-621

Vickery, Brian (2008) *On knowledge organization* . Retrieved from http://classic-web.archive.org/web/20080404103206/www.lucis.me.uk/knowlorg.htm.

Corresponding author:

Patrick Lambe can be contacted at plambe@straitsknowledge.com