A note on the theory and practice of knowledge organization and knowledge representation

Alan GILCHRIST

Cura consortium and Metataxis
cura@fastnet.co.uk

Resumen

Se discute que, dado que el conocimiento es abstracto y que cada persona tiene una percepción única de su entorno y de las propiedades y comportamiento de sus componentes, se debe colegir que las personas comprometidas en la representación y organización del conocimiento deben basar su trabajo en los registros físicos, a los cuales podemos denominar transportadores de información o mensajes. Los productos basados en el análisis de estos mensajes pueden por consiguiente ser considerados modelos del conocimiento. Estos modelos son creados para reducir la complejidad y ganar una comprensión más clara de los aspectos del mundo que nos rodea, pero deben ser continuamente comprobados y revisados en un entorno de trabajo. La evaluación de los productos del conocimiento se realiza frecuentemete por los científicos de la información en el proceso de asegurar la recuperación de la información, mientras que los productos de representación del conocimiento también se apoyan en la organización del conocimiento, que puede ser entendida, hasta cierto punto, como auto-evaluada. Se sigue que se podría ganar mucho de una colaboración más estrecha entre los que trabajan en los campos de la representación del conocimiento, de la organización del conocimiento y lo que están implicados en la provisión de información a los usuarios.

Palabras clave: Organización del conocimiento. Representación del conocimiento. Recuperación de la información. Interdisciplinariedad. Evaluación.

Abstract

It is argued that because knowledge is abstract and every person has a unique perception of his environment and the properties and behaviour of its components, it follows that those people engaged in Knowledge Organization and Knowledge Representation must base their work on physical records, which we may call carriers of information, or messages. The products based on analysis of these messages can then be considered as models of knowledge. Models are created in order to reduce complexity and to gain a clearer understanding of aspects of the world around us, but they must be continuously tested and revised in a working environment. The testing of the products of Knowledge Organization is often carried out by information scientists in their provision of information retrieval, whereas while the products of Knowledge Representation also rely on Knowledge Organization, they may be considered, to some extent, to be self-testing. It follows that much can be gained by a closer collaboration between those engaged in Knowledge Organization, Knowledge Representation and those engaged in delivering information to end users.

Keywords: Knowledge representation. Knowledge organization. Information retrieval. Interdisciplinarity. Evaluation.

1. Foreword

Let me start by stating that what follows is a personal view, coloured by the fact that I am a professional consultant, rather than an academic; and that, while I am a member of the Executive committee of ISKO(UK) my views do not necessarily represent the views of my Committee colleagues.

2. Introduction

The etymology of work in the Library and Information Sciences has often been poor and potentially misleading, and this etymology is tending to get more confused as others make their contributions to the increasingly wide and complex arena of "information processing". For example, before full text retrieval became possible, the neologism 'information retrieval' should more accurately have been called 'reference retrieval'; and the term 'information scientist', coined in the UK meaning 'a scientist employed to provide information to other scientists' gave rise to the term 'Information Science' which, in those early days, was a somewhat presumptuous claim.

Today, with librarians, information scientists, website designers, knowledge engineers and

others all engaged in aspects of information processing, it is often difficult to know what is meant by such words as 'facet', 'taxonomy' or 'ontology' and above all to one of the most overloaded words 'knowledge', which we find in the compound terms 'Knowledge Management' (KM), 'Knowledge Organization' (KO). 'Knowledge Representation' (KR), 'Knowledge Engineer' (which seems to have replaced the term 'Artificial Intelligence'), and even 'Knowledge worker'. So let us take a closer look at this key word 'Knowledge'.

3. Knowledge

The study of Knowledge per se is Epistemology: "the theory or science of the method or grounds of knowledge" (Oxford English Dictionary, 2011).

As such, it seems to be mainly concerned with such questions as "What does it mean 'to know'"; and what can we mean by the terms 'scientific knowledge' or 'religious knowledge'; or are there limits to what we can know. This leads to the thought that doubt must come after knowledge, which is elaborated in the concept called "fallibilism", defined as "the philosophical principle that human beings could be wrong about their beliefs, expectations, or their understanding of the world", and that "any claim justified today may need to be revised or withdrawn in light of new evidence, new arguments, and new experiences" (Wikipedia contributors, 2011).

The physicist Lawrence Krauss has been quoted as saying: "Uncertainty is a central component of what makes science successful". This constant state of flux concerning what we think we know forces us to consider what we can mean by the word knowledge in a practical and every day sense; and the only way we, in the LIS professions, can deal with it is to recognize that knowledge is held in the individual brain —and that shared knowledge is a consensus (inevitably a compromise) of shared perceptions.

It follows that, strictly speaking, KM is an impossibility, and that we must also be careful what we mean by KO. Knowledge is shared either orally (and this is where KM attempts to facilitate such transfer); or through written records (which is the domain of LIS).

Nonaka and Takeuchi (1995) have convincingly made this distinction in their book "The knowledge creating company" in which they describe "tacit knowledge" as that which is stored in the brain and "explicit knowledge" as that which is recorded, (which we may fairly call 'information'). They also, in what has become known as the SECI Model, show the four transformations between tacit and explicit knowledge:

- Tacit to tacit (socialization): a classical component of KM
- Tacit to explicit (externalization): transfer of knowledge from minds to formal records
- Explicit to explicit (combination): creation of new records through collation
- Explicit to tacit (internalization): assimilation of knowledge from records.

Though it has just been claimed that we can consider explicit knowledge to be the same as information in the normal sense of the word, we should continue to be careful with our choice of words and remember that one 'item' of information assimilated by one person does not have exactly the same meaning as it might have for another.

We might, then, argue that the term 'Information Management' is also misleading; and, indeed, Miller (2002) has argued that "information has no intrinsic meaning", distinguishing between meaning and 'messages', or in the terms above explicit knowledge as records or (in their widest sense) documents.

Not only may interpretations vary according to the different mental 'knowledge stores' of the two individuals, but context and use also come into play, and as Myers and Myers (1998) say

> ...words don't have meanings. There is no direct relationship between the thing you are talking about and the words you use. Only as these words are related through the thoughts of a person do they have meaning. Meaning is not in the object or in the symbol but in the interaction of these through the human [communication process].

(It may be observed here, in passing, that the word knowledge has acquired a rather grander sound than the humbler and more prosaic word information - which is why some vendors of retrieval software erroneously claim to be selling knowledge retrieval).

4. Information

If, then, we should consider that our only approach to knowledge is through the recorded message there are some important implications.

First, that messages are issued about very many and diverse sorts of information. So, in principle, exponents of KO and IS should be concerned with facilitating access to records of 'many and diverse sorts of information', and should be prepared to process all messages where there is a demand for such messages to be retrieved.

Second, it is not for us to dictate what are good or bad messages (though we may well have informed opinions about them). Floridi (2010) has proposed a tree structure of information concepts which includes "untrue information", which can be either misinformation (unintentional) or disinformation (intentional). Librarians and information scientists process messages to facilitate their retrieval, and only indirectly to provide access to knowledge, and consequently are bound to deal with 'untrue information" as well as "true information". When does true information become untrue —and vice versa? Are we not concerned, for example, with the history of science and what arguments were once put forward for geocentrism? Are we not obliged to record the latest findings from bioarchaeological evidence correcting earlier notions of Homo Sapiens, Neanderthal Man and their interaction? We know that Beijing was once called Peking and that Slovenia was part of Jugoslavia, but we must update our maps and gazeteers, while maintaining the previous names for the records.

The third point arising from the obligation to deal with messages is that, if we accept KO to be largely about the semantics of Structured Vocabularies (SV) (1) and IS to be about the delivery of information, then it is suggested that there is an increasingly large overlap between the two activities, which implication will be discussed later in this paper.

5. Models

Returning to the concept of fallibilism, defined earlier, we may acknowledge that the only way to rise above the tricky ground of knowledge, information and meaning is to think in terms of 'approximations'.

Neil Gershenfeld, another American physicist, has been quoted as saying that "the most common misunderstanding about science is that scientists seek and find truth. They don't —they make and test models".

Are not all 'mentefacts' (2) in some sense models? And if this is true, then our SVs can also be considered as models 'approximating' to the contents for which they are designed.

Some support for this view may be found in the concept of 'pragmatic epistemology' enunciated by Heylighen (1993), who says:

> This philosophy still dominates most present work in cognitive science and artificial intelligence. According to pragmatic epistemology, knowledge consists of models that attempt to represent the environment in such a way as to maximally simplify problem-solving.

Is this not what we do when we construct a model that we call a structured vocabulary? And is this structured vocabulary not a compilation of words without meanings until they are used in indexing (tagging) to support an aspect of problem-solving?

Though this point is not central to the argument in this paper, it could be argued that we (whether working in the areas of KO or IS) must no longer be concerned only with the models of SV which, incidentally, have called on, for example General Systems Theory and the Theory of Integrative Levels in the approach to building bibliographic classifications.

We must now concern ourselves more thoroughly with such models as semantic maps as used in ontologies, as well as models that we might use in that part of IS called information architecture, such as enterprise architecture models and domain models used to describe the environments in which information retrieval operates.

We should also be conversant with modelling techniques such as mind maps and the Unified Modelling Language, which has been used to model a complete set of the components of a thesaurus and their relationships to each other.

In essence, KO is about the analysis of explicit knowledge, just as information architecture is largely about the analysis of the environments in which the products of KO are deployed. Furthermore, such analysis may not be confined to the support of conventional information retrieval which has expanded in recent years as electronic information becomes more manipulable.

This expansion was well illustrated in a number of interesting papers given at the second Biennial Conference of ISKO(UK) last July. In particular Campbell (in press) showed how application of Farradane's relational indexing to parallel texts in the languages of health professionals and lay writers discovered significant, but not immediately recognizable, ambiguities between the two texts.

In another paper, Lambe (in press) argued that in the current complex state of scientific research, involving multidisciplinary teams, KO had a role to play in identifying and recording new boundaries and overlaps between apparently disparate subject areas as an aid to sense-making in the pursuit of scientific discovery. Finally, Petras proposed that Knowledge Organization Systems (KOS) could be considered to be deployed in seven distinct applications in information retrieval systems (in press).

6. Knowledge Representation

A Knowledge Representation is also a model, incorporating a structured vocabulary, but one with more features, thereby making it more powerful for certain defined operations. This link between KR and modelling is well described by Davis et al. (1993) who argue that

> Any intelligent entity that wishes to reason about its world encounters an important, inescapable fact: reasoning is a process that goes on internally, while most things it wishes to reason about exist only externally.

These authors go on to propose that a KR is best described in terms of five fundamental roles:

> (1) a surrogate, a substitute for the thing itself, used to enable an entity to determine consequences by thinking rather than acting.
>
> (2) a set of ontological commitments, i.e. an answer to the question: In what terms should I think about the world?
>
> (3) a fragmentary theory of intelligent reasoning, expressed in terms of three components (i) the representation's fundamental conception of intelligent reasoning; (ii) the set of inferences the representation sanctions; and (iii) the set of inferences it recommends.
>
> (4) a medium for pragmatically efficient computation.
>
> (5) a medium of human expression, i.e. a language in which we say things about the world.

KR, then, has a distinct relationship with KO (or can be seen as an applied form of KO). The products of both are surrogates using a medium of human expression, but KR employs a statement of intent and some formal logic manipulated by an inference engine.

A major and important difference is that KRs work most effectively in relatively closed systems with specific purposes, such as suggesting treatment on the basis of medical diagnoses.

The challenge for the Semantic Web is to vastly extend such logical reasoning and, we should note here, to ensure that the underpinning and necessary SVs are not only available, but as well-formed and maintainable as possible.

A second difference that leads on from the first is that the vocabularies supporting KRs are "special" as in the early use of the term to distinguish such classification schemes from the larger universal schemes such as the Dewey Decimal Classification.

A further distinction is that the special and universal schemes used in 'bibliographic' retrieval systems are in some sense predictive, especially where flexibility of co-ordination of concepts is supported by a structure created by facet analysis.

Where the two converge is in the development of what we may call an ontology. In KR this is the vocabulary component that is manipulated by a logic layer such as OWL and mounted on the Internet using a language such as RDF.

This ontology will usually have genus/species relationships —designated as Broader Terms (BTG)/Narrower Terms (NTG) in thesauri, and as 'isa' in KR), and whole/part relationships (designated as BTP/NTP in thesauri and 'is part of' in KR.

A major difference is that the KR also has Related Terms which, unlike most thesauri, are specifically defined (such as the unidirectional thesaural RT between Irbesartan and Hypertension which can be designated in a KR as the bidirectional relationship 'is used to treat'/'treated by').

However, with currently available software it is now possible to create thesauri for information retrieval applications incorporating not only defined hierarchies and relationships but to add notes and sortable codes to each term. These 'enriched' thesauri, sometimes called ontologies, can be extended to perform as terminology repositories to facilitate interoperability within, and even between, organizations. (Note, however, that closely defined RTs may be mostly specific to an organization or application, because whereas hierarchies must be either generic/specific, whole/part or instantial (and even here they may be further defined) thesaurus standards do not list universal RT types, other than as guidelines drawn from facet analysis).

7. Testing models

Mention was made earlier of 'pragmatic epistemology', which must be an offshoot of a branch of philosophy called 'Pragmatism' which, proposed by the American philosopher Charles Sanders Peirce and elaborated by his compatriots William James and John Dewey, is defined as "A method of understanding facts and events in terms of cause and effect, and of inferring practical lessons or conclusions from this process" (Oxford English Dictionary, 2011).

The testing of models created by KO for KR, KOS and information retrieval must, to a large extent be pragmatic: a process of inference rather than reliance on circumstantial detail. There are many more types of models than those discussed above, but all share the property of being representations of complexity so that we

can better understand and deal with the world around us.

But, because they are approximations and because the things that we model are constantly changing, we must also continuously test our models and revise them so as to maintain some understanding of our world.

Scientists are constantly testing models and challenging hypotheses (just think about the amazing discoveries and hypotheses emanating from the experiments in the Large Hadron Collider on an almost daily basis) and subsequently describing these models and hypotheses in the learned journals. Their papers may then be analysed and indexed by information scientists, using semantic models devised by 'Knowledge Organizers', and here we arrive at an important point.

The models that we call SVs must, like all models, be tested; and the only practical way in which they can be tested is in their application, that is to say through their efficacy in relating to the written records for which they are designed. It is not too fanciful to suggest that such testing is an aspect of 'applied epistemology', defined as (MacGraw…, 2003):

> The use of machines or other models to simulate processes such as perception, recognition, learning, and selective recall, or the application of principles assumed to hold for human categorization, perception, storage, search, and so on, to the design of machines, machine programs, scanning, storage, and retrieval systems.

So we may now say that KO has, like many other disciplines, a pure and an applied component; the former being concerned with the theory and techniques of KO, the latter with the application of the complete mentefacts such as SVs.

This indicates, not surprisingly, that there is a large overlap between the interests and activities of the two communities of KO and IS, and one which, perhaps, is likely to become more pronounced.

There has been mention in this paper, for convenience, of the words indexing and tagging without further elaboration; but in an environment of distributed processing and full-text retrieval these words are now used more loosely and tagging is not confined to specialists.

At the same time, information scientists are increasingly directly involved with the creation of 'special' SVs and interoperability between them. In other words they have become more concerned with KO at the application level, while other information scientists continue to help maintain, develop and map the big schemes such as Dewey, MeSH and EUROVOC.

8. Conclusions

In recognizing that knowledge is an abstract entity confined to the individual human brain, the proponents of KO and IS must both work pragmatically, albeit that they work with concepts. But even here they must remember that concepts are also abstract and that words have no intrinsic meaning.

Information is also abstract and information scientists must work with 'messages' which are the attempts of knowledge creators to disseminate their opinions.

According to the SECI model knowledge creators record their ideas as explicit knowledge (Externalisation) which are then analysed by IS (Internalisation) so that they can 'describe' them by processes such as abstracting and indexing (a sort of Combination), making these new products available for intelligent information retrieval. And so the cycle continues with users reading the retrieved documents (Internalization) and creating new messages, sharing these with others in discussion or giving papers at conferences (Socialization) and recording them (Externalization).

The eminent information scientist Robert Fairthorne once said "It is not the job of information scientists to give information, but information about information". One might add that it is not the job of those concerned with KO to give knowledge, but information about knowledge; and this is achieved through the construction of models.

Throughout these various information chains models are being continuously created and pragmatically tested, from which it follows that the efforts of those working in KO, KR and IS overlap to such an important degree that they should work closely together. Unfortunately, present institutional and educational structures are not providing effective support for such collaboration.

9. Postscript

ISKO(UK) was established in February 2007, and so is a relatively new Chapter of ISKO. Its objectives are "to promote research and communication in the domain of knowledge organization, within the broad field of information science and related disciplines", and "to seek to establish links with all groups and fora in the UK involved in information and knowledge organization".

In its four years it has held 14 afternoon meetings, many attended by up to 100 people, on such diverse topics as KO techniques (semantic analysis technology, facet analysis); new developments (Semantic Web, Linked Data); and subject specific events (Legal information, Cultural Heritage, Records Management). The ISKO(UK) website carries accounts of all these meetings and, in some cases, audio recordings of the talks.

The biennial two-day Conferences have each attracted around 140 delegates, and have been addressed by speakers from 18 countries, which suggests that our collaborative objectives do, and should, extend beyond the UK.

In recent months some disquiet has been expressed in the UK about the "fragmentation" of the information profession and whether anything should be done about it. Fragmentation means "the process or state of breaking into small or separate parts".

Historically, in the UK, there has been a 'process' of fragmentation of the library and information institutions, but the current 'state' of fragmentation has more to do with the evolution of an information-centric world. The UK Library Association was established as long ago as 1877, and in 1924 Aslib (Association of Special Libraries and Information Bureaux) was formed because some thought that the Library Association was paying insufficient attention to the needs of the newly emerging industrial and business libraries. In the 1950s there was another 'break-away' movement when Jason Farradane and others established the Institute of Information Scientists and the first School of Information Science in the world, believing that information science was a distinct academic subject.

In 2002, the Institute, suffering from financial difficulties, merged with the Library Association to form the Chartered Institute of Librarians and Information Professionals —an odd title which seems to suggest on the one hand that librarians are not information professionals, but on the other fails to define what is meant by the term information professionals. The merger has been less than successful, and so it was not perhaps surprising when one of the contributors to the 'fragmentation' debate observed that ISKO(UK) appeared to have filled the gap left by the demise of the Institute of Information Scientists. Be that as it may, the fact remains that there are many and diverse 'information professionals', however that term may be defined (as well as the myriad numbers of 'information amateurs' who sit at their PCs authoring, publishing, disseminating, storing and retrieving information). At this moment in the evolution of the information-centric world it would be impossible, and wrong, to 'institutionalize' the disparate range of information professionals, but KO and IS are too important to ignore the diffusion of 'information professionals'.

In her Presidential address to ISKO, Maria Lopez-Huertas called for the opening up of ISKO to all those in any way concerned with KO; and also stated that KO should address actual needs and so provide a better service to the Information Society. There is still much to achieve.

Notas

(1) Including, according to the BS and ISO standards, thesauri, classifications, taxonomies and ontologies.

(2) A word coined by the classificationist Barbara Kyle as an amalgamation of mental and artefact.

Referencias

Campbell, D. Grant. (In press). Revisiting Farradane's relational indexing in a consumer health context. // Gilchrist, A.; Vernau, Judi (Eds). Facets of Knowledge Organization: proceedings of the ISKO(UK) Biennial Conference, 4th/5th July 2011, London. Bingley: Emerald Group Publishing.

Davis, Randall; Shrobe, Howard; Szolovits, Peter (1993). What is a KR? // AI Magazine. 14:1, 17-33. http://medg.lcs.mit.edu/ftp/psz/k-rep.html (2011-06-11).

Floridi, Luciano (2010). Information: a very short introduction. Oxford: Oxford University Press.

Heylighen, F. (1993). Epistemology, introduction. // Heylighen, F.; Joslyn, C.; Turchin, V. (editors). Principia Cybernetica Web (Principia Cybernetica, Brussels). http://pespmc1.vub.ac.be/EPISTEMI.html (2011-06-11).

Lambe, Patrick (In press). Knowledge organization systems as enablers to the conduct of science. // Gilchrist, A.; Vernau, Judi (Eds). Facets of Knowledge Organization: proceedings of the ISKO(UK) Biennial Conference, 4th/5th July 2011, London. Bingley: Emerald Group Publishing.

McGraw Hill Dictionary of Scientific and Technical Terms. New York: McGraw Hill. 2003.

Miller, F. J. (2002). I=0 (Information has no intrinsic meaning). // Information research. 8:1. http://informationr.net/ir/8-1/paper140.html(2011-06-11).

Myers, G.E.; Myers, M.T. (1998). The dynamics of human communications. New York: McGraw Hill.

Nonaka, Ikujiro; Takeuchi, Hirotaka (1995). The knowledge creating company. Oxford: Oxford University Press.

Oxford English Dictionary. Oxford: Oxford University Press, 2011. www.oed.com

Petras, Vivien (In press). Knowledge organization systems and their consequences for information retrieval. // Gilchrist, A.; Vernau, Judi (Eds). Facets of Knowledge Organization: proceedings of the ISKO(UK) Biennial Conference, 4th/5th July 2011, London. Bingley: Emerald Group Publishing.

Wikipedia contributors (2011). Fallibalism. // Wikipedia. Wikimedia Foundation. http://en.wikipedia.org/wiki/Fallibilism(2011-06-11).