Situating Knowledge Management on the Interpretation Continuum

Leo Obrst* and Patricia Carbone†

*VerticalNet, Inc.
700 Dresher Rd., Suite 100
Horsham, PA 19044
lobrst@vertical.net

†The MITRE Corporation
1820 Dolley Madison Blvd.
McLean, VA 22102
carbone@mitre.org

Abstract

Knowledge Management (KM) is an emerging discipline which seeks to assemble, codify, and manage a repository of human knowledge specific to an organization. KM necessarily involves organizational science and structure (sociology, industrial psychology and engineering), management strategies and practices (business management, business process re/engineering, economics, decision science), and information technologies (including knowledge representation and knowledge engineering). This paper attempts to situate the notion of knowledge management on the *interpretation continuum*, by gauging knowledge as "interpretation."

What is a Knowledge Management System?

Knowledge Management (KM) is an emerging discipline which seeks to assemble, codify, and manage a repository of human knowledge specific to an organization. KM necessarily involves organizational science and structure (sociology, industrial psychology and engineering), management strategies and practices (business management, business process re/engineering, economics, decision science), and information technologies (including knowledge representation and knowledge engineering).

Knowledge Management requires intelligent information processing, including information dissemination, information retrieval, software agents, and text summarization. For each of these technologies to perform their respective desired functions, however, they need metadata, domain knowledge, or an ontology that enables the functions to perform their intelligent service. The term "ontology" itself is subject to the same buzzword-spawning process that afflicts "knowledge management". To give substance to the notion of ontology, which has a well-defined and extensive technical tradition and a more recent computational research line which incorporates the results and methods of that tradition, an *ontology* is the vocabulary for expressing the entities and relationships of a conceptual model for a general or particular domain, along with semantic constraints on that model which limit what that model means. Both the vocabulary and the semantic constraints are necessary in order to correlate that information model with the real world domain it represents. According to this notion, an ontology encompasses metadata and domain theories. Metadata is exactly that data which describes the semantics of the underlying simple (object-level) data. A domain theory is just a specialized ontology, i.e., a vocabulary identifying the semantic entities, relationships, and attributes of the domain model, and constraints on that domain. Hence, an ontology acts as the skeleton of a knowledge base, and is comparable to the notion of schema in database terminology.

What does knowledge management mean? In general, "knowledge management systems" is the current term for a set of technologies which evolved with slight modification from what was termed digital libraries, enterprise information systems, and enterprise engineering. These systems are typically web-based and link people in an organization with other people and their documents (i.e., "knowledge"), usually through a search engine.

Consider the typical manner in which people use web browsers currently. They click to link to a document for which they've either searched, using simple keywords, or which is already indexed on the page they are viewing. The document is then displayed before them. They must then read the document, and using their own internalized conceptual model of the world and that document's domain, interpret the meaning of the document. The knowledge in the document is not necessarily available to them: it may require extensive background information, long experience, or many years of formal education. A retrieved document on the topic of interacting bosons in particle physics has knowledge which an average person cannot extract. The knowledge cannot be captured and transferred because the average person cannot interpret the words semantically, cannot decipher their intended meaning. Why? Because the person does not have a sufficiently rich conceptual model of that domain. With today's applications, we lose sight of the fact that they require the individual human user to be their semantic

interpreter, i.e., the user must figure out the knowledge contained in a document without any computer software interpretation of the meaning of that document.

A true knowledge management system encapsulates knowledge and at least partially symbolically interprets that knowledge for a user, meaning that it categorizes the knowledge in a way and to a fine degree so that the knowledge is directly usable by the user. Having a knowledge model that is conceptually close to what the user desires ensures usability. The visionary promise of robust knowledge management is a system which assists the user in obtaining and interpreting knowledge at an arbitrary level of understanding commensurate with a user's own knowledge and desired level and mode of presentation. It is the degree of automatic interpretation which is controversial, however.

Knowledge as Interpretation

With the rise of the availability of information and traffic related to information on the Internet comes a concomitant need to exploit and manage such information flow and and its storage using more intelligent means. Because the information is actually an information continuum (to be termed "interpretation continuum" in this paper) which ranges from completely unstructured data to very structured knowledge, a variety of intelligent methods must be employed to filter, to fuse, and to represent data which ultimately becomes user- and institution-level knowledge.

Data, Information, Knowledge

The terms data, information, and knowledge are usually used in ill-defined ways, which is acceptable in colloquial conversations where only a rough, intuitive notion of their distinctive meanings are needed. There is a danger, however, that such usage may result in an obliteration of crucial technical distinctions, with the possible result that these terms either become synonyms or equivalently that one term begins to stand for all, but in a watered-down fashion devoid of content.

This paper will not try to give precise formal definitions to these three terms, but instead informally delineate their meanings. By doing so, we hope to show that what we have called the "information continuum" in the introduction above is actually better termed the "interpretation continuum".

In this view, data and knowledge are simply endpoints on (for our purposes here) a linear continuum, as the following figure suggests.

Figure 1. The Interpretation Continuum

That which is nearer the left side of the continuum is termed data; that which is nearer the right side is termed knowledge, though in general the distinction between any two points on the line is more appropriately expressed in terms of their lesser or greater structure. But structure itself, though important, is not the crucial determining or characteristic factor for the continuum: interpretation is. Structure is a side effect of the degree of interpretation required. Knowledge is the relatively complex symbolic modeling (representation) of some aspect of a universe of discourse (i.e., that which we can talk about as human beings); data is a relatively simple symbolic modeling.

Knowledge = Data + Interpretation

Information is here used in its more technical sense derived from Shannon: information is a measure of the extent to which a piece of knowledge tells you something which you did not previously know. Hence, the information contained in some piece of data/knowledge depends on what a person already knows and in general will vary from individual to individual. Information represents change in knowledge, i.e.,

New Knowledge = Old Knowledge + Information

But what now is interpretation and why is it important?

> *Interpretation is the mapping between some structured set of data and a model of some set of objects in a universe of discourse with respect to the intended meaning of those objects and the relationships between those objects.*

Interpretation, therefore, is the mapping between notations (which we will call "glyphs"), e.g., strings of characters from some alphabet (for text) or some set of defined binary encodings (for graphics, video, etc.), and what those notations are intended to mean in a human-defined universe of discourse. Notations (or glyphs) are meaningless symbols unless they are given an interpretation, i.e., mapped to objects in a model. Interpretation is semantics: it is interpreting the syntactic glyphs with respect to their intended semantics.

Now, typically the model lies in the mind of the human. We understand, which means: we *symbolically represent in some fashion the world, the objects of the world, and the relationships among those objects*. We have the semantics of (some part of) the world in our minds: it is structured and interpreted. When we view a textual document, we see glyphs on a page and interpret those with respect to what they mean in our mental model; that is, we supply the semantics. If we wish to assist in the dissemination of the knowledge embedded in a document, we make that document available to other human beings, expecting that they will provide the semantic interpreter (their mental

models) which will make knowledge out of the notations (glyphs) on the document pages. There is no knowledge there without interpretation; interpretation makes knowledge out of squiggles on a page.

However, if we wish to have the computer assist in the dissemination of the knowledge embedded in a document, we need to partially automate the interpretation process, which means we need to construct and represent in a computer-usable way a portion of our mental model. What does that mean? We turn to this question in the next section.

Knowledge Issues: Acquisition, Representation, Manipulation

Much of the data now being made accessible to individuals via the Internet are structured in relational databases. Other kinds of data exist as textual and graphical documents, video, speech, and other specialized data stores. All of these data sources, however, are relatively unstructured to moderately structured. Relational databases, which have both an intensional database (the generic information or schema, i.e., part of the metadata) and an extensional database (the instances or tuples of the schema), still greatly rely either on application code for the semantic interpretation of their data or, as is the case in general for documents and relatively less structured data, on human beings such as database administrators, users, etc. Their primary purpose is for storage and ease of access to data, not complex use; software applications (with the data semantics embedded in nonreusable code via programmers) and human beings must focus on data use, manipulation, and transformation, all of which require a high degree of interpretation of the data.

To partially automate the interpretation process (i.e., to shift some interpretation ability to the computer from the human), in recent years stochastic methods have been employed. These range from the data mining of statistically relevant patterns in relational databases, to the use of statistical techniques in natural language processing, image processing, etc. What most of these shallower (that is, interpretation-nonintensive) methods have in common is the attempt to discover statistically significant correlations of data, words, relatively unstructured units, and by those correlations infer semantic significance. It is assumed that co-location or co-occurrence within some defined context signifies a semantic relation. What that semantic relation is, of course, must be conceived (interpreted) by a human being who inspects the statistically significant patterns. These methods however, remove at least a portion of the noise from relatively unstructured data, performing a filtering function, and display before the interpreting human a subset of what can become knowledge once the human interprets it. Such methods attempt to bypass the knowledge bottleneck encountered by a previous generation of technologies epitomized by expert systems: capturing complex knowledge requires much work in elicitation from domain experts.

Deep knowledge management, it was thought, requires time-consuming, responsible, and consistent knowledge acquisition; consistent (or at least partition-bounding the inconsistent), semantically correct, and expressive knowledge representation; tractable, semantically-licensed, sound but nonmonotonic reasoning; and ongoing knowledge maintenance. Such an effort was in general deemed too resource-intensive by industry. Coupled with extravagant claims made by a small set of researchers, and thus generative of unrealizably high expectations, this perception quashed or at least curtailed efforts toward deep representation. The need for such deep representation, however, continues and in fact has increased: to maximize the knowledge we can utilize, we must shift more of the burden of interpretation onto our computers.

Figure 2. displays the interpretation continuum mentioned earlier, but now annotated with additional information on methods and technologies (along the bottom) and types of knowledge (along the top) as one progresses from data rightward to knowledge.

If all agree that both shallow and deep methods and representations should be employed, that increased knowledge utility, hence increased computer-assisted interpretation is desirable, then the question becomes "How might shallow and deep methods and representations be combined?" A related question is: "Are there promising technologies which might help combine shallow and deep?"

Conclusions and Challenges

In the not too distant past, knowledge management research concentrated on the representation and storage of and reasoning against the metadata or structured knowledge of a domain. As such, knowledge management was synonymous with “knowledge base management”. With the recent emphasis on building “knowledge management” for a corporation, however, there has been a displacing surge of interest in the more cultural and political issues to encourage and enforce knowledge management in an enterprise. This recent emphasis has its origins in business process re/engineering and business management methodologies which seek to build a “knowledge corporation”. Without some kind of knowledge representation and storage of the domain and the inherent metadata associated with a vast array of differently structured data, and sound methods to automatically reason about the knowledge, however, the technologies to facilitate a true knowledge management system will not be available nor useful. We must go rightward on the *interpretation continuum*.

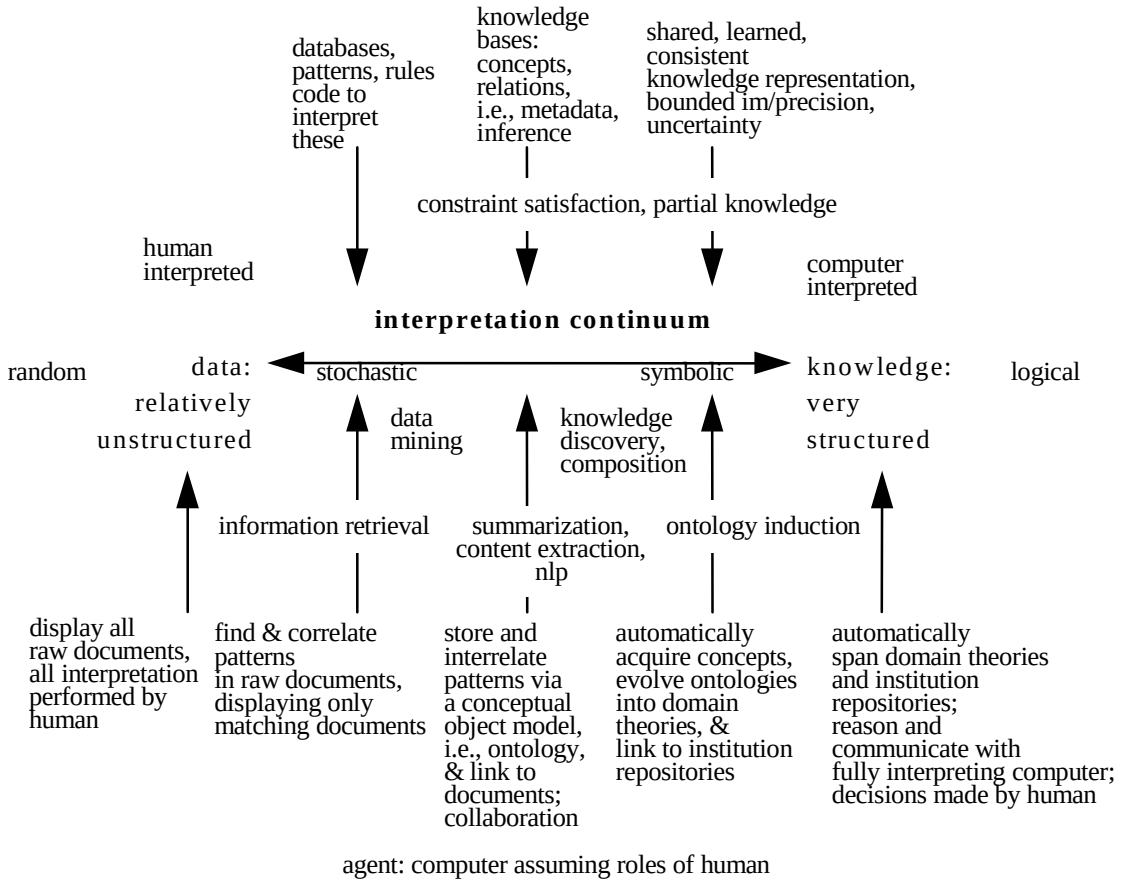

Figure 2. The Interpretation Continuum Revisited