KNOWLEDGE-BASED SYSTEMS AND COPYRIGHT

Patrice A. Lyons

A Distributed Information Model as laid out by Dr. Robert Kahn of CNRI serves a very general purpose. It could either encompass a structure that is tailored along the lines of other models under consideration, or admit various overlapping structures within the model. The advantage of this model is that it allows the network environment to serve as a marketplace in which both formal and informal goods and services may be exchanged. Within the marketplace framework, individuals and organizations may enter into agreements among each other or collectively that dictate the terms and conditions under which goods and services may be accessed and utilized. In this essay, I am going to discuss the nature and possible legal status of a particular component of this model: knowledge-based systems.

The key point that distinguishes the Distributed Information Model from other models that may be similar is the use of active intelligent computer programs that travel within the network and access network resources on behalf of end-users. These programs contain sufficient information for them to carry out the wishes of the user. They are expected to become increasingly sophisticated over time, as the technology improves and the results of individual queries are accessible to them. This frees the user from knowing many of the details of the network in order to enhance ease of access and use.

Lyons is an attorney practicing copyright, communications, and trademark law in Washington, DC.

Knowledge-based Systems

Potential information resources accessible within this model are limited by the desire of rightsholders to make them available and the ability of information technology to access them for retrieval. At the present time, the environment might include databases, knowledge-based systems, brokering systems of various kinds, gateways, transformational systems (send data and get back a graph or send in a page reference and get back a figure on the page), agents that help users and their systems in the formulation of queries, and mediators who help plan database retrieval strategies. For purposes of this discussion, knowledge-based systems are considered to be either conventional databases with smart retrieval software, where the output is more than simply information stored in the database, or knowledge bases in the more usual sense of artificial intelligence (AI), where rules, facts, heuristics, and other forms of experiential knowledge are embedded in a computer program. A knowledge-based system in an AI sense may or may not have a separable and distinguishable database system associated with it.

Many of the issues that would arise in the context of such a model are well known and have been intensively studied. For example, considerable experience has been acquired in providing access over computer networks to conventional databases. In this paper, I will focus on an aspect that has received less attention, how intelligent programs might interact over computer networks with knowledge-based systems to reply to user queries. To take the interactive element into account,

I propose that consideration be given to a new "right to access" under section 106 of the copyright law, title 17 U.S.C.

Proposed Right to Access

The introduction of an access right for producers of knowledge-based systems was suggested to me by a comment made a few years ago by a noted computer scientist, Allen Newell. When addressing the form of algorithms, he observed:

> As anyone in computer science knows, the boundary between data and program—that is, what is data and what is procedure—is very fluid. In fact, as our discussion of the forms of algorithms indicated, there is no principled distinction in terms of form or representation of which is which. What counts is the total body of knowledge represented somehow in the assembled symbolic expressions. This totality determines the ultimate behavior of the machine.[1]

It occurred to me that, instead of treating data and program as separate entities in this context, it would be better to view the recasting of any preexisting elements, whether discrete or embedded in a database or knowledge base, to form what may be termed a "knowledge-based system," as the creation of a complex computer program.[2] A computer model for analyzing the various processes of typhoons that is under development at Florida State University (FSU) may serve to clarify the nature of these programs.[3]

Typhoon Model

The Typhoon Model, also known as a "Large-Scale Numerical Partial Differential Equation Model," may be used to describe the life cycle, wind patterns, speed, intensity, and strength of a typhoon. Generally, this is a mathematical way of describing how typhoons may evolve over time. Such models are sets of equations that describe how elements interact: once you know the starting conditions, you are able to predict how the situation will evolve. The model developed at FSU may contain some or all of the following material:

1. statistical data with respect to winds, currents, oceans, atmospheric pressures, and temperatures;

2. formulas such as mathematical relationships between the various elements; and

3. heuristics (i.e., rules of thumb that describe what is going on).

Whether the model is viewed as an "expert system," a "knowledge-based system," a computer program, a computer database, or just a model, the bottom line is that people may be using it as a basis for making judgments or decisions that could have important consequences in their lives. Whatever the actual label attached to the data, expertise, and mathematical equations in the model, and, apart from the possible copyright, patent, or other legal claims in particular units of data or programs that may be embedded in the work, a legal framework is needed to cover the consequences of accessing the system as a whole to provide guidance, advice, or other information—in this case with respect to typhoons.

The word "model" may not be a precise enough description here, since the term is used to represent a potentially wide spectrum of entities from plaster of paris molds to paradigms that are just words. It is also possible to take exception to the notion that such models accurately predict certain behavior. Predicting is usually understood to mean telling something that will happen in the future. Taken in that sense, the word "predicting" may be imprecisely used in connection with such models, when you are really describing a relationship between variables that cannot be measured accurately enough or that is inherently chaotic.

When discussing a particular model and its use, it may be appropriate to refer to it as a knowledge-based system. Such a system is really a large set of computer codes connected to a database. The system massages the database through complex equations. It is an operational system, consisting of hundreds of algorithms and pieces of data, that processes input and produces output.

The Typhoon model could be viewed as a computer database or even an "expert system," but the distinction may not be useful. While there may be some parts of the system that may be deemed "expert," the system may also contain simple listings of data that may be categorized as a "database."

Application of Existing Public Performance Right

While I argue that there is a need for a new "right to access" knowledge-based systems, because the output is not measurable simply with reference to the information stored in such systems, the process of establishing this right may be viewed by some as troublesome. In this context, further consideration should be given to the nature and scope of protection that computer programs already enjoy under the U.S. copyright law. Unlike sound recordings, where Congress specifically

excluded a right of performance from the scope of protection, computer programs were not so limited. Since the copyright law is silent on the subject of the right of public performance in computer programs, it may be asserted that the owners of copyright in computer programs do enjoy this right.

While a right to access would clearly take into account the interactive nature of access to knowledge-based systems, the processing of information by such a system and the communication of output to members of the public may also be viewed as the public performance of these computer programs. The performance of such programs was discussed in a recent article on the Cyc project.[4] In describing the current state of the art in capturing knowledge in artificial intelligence programs, it was noted that:

> [M]uch of the "I" in these "AI" programs is in the eye—and "I" of the beholder. Carefully selecting just the fragments of relevant knowledge leads to adequate but brittle performance: when confronted by some unanticipated situation, the program is likely to reach the wrong conclusion. It is all too easy to find examples of such brittle behavior: a skin disease diagnosis system is told about a rusty old car, and concludes it has measles; a car loan authorization system approves a loan from someone whose "years at the same job" exceeds the applicant's age....[5]

The authors appear to be referring to the processing and output of data by a knowledge-based system as "performance." The transmission of the output of such performance over computer networks may be deemed a further performance. There would appear to be room to apply the copyright concept of performance to these activities.

Knowledge Embedded in Computer Programs

In considering the programs that would qualify for this right to access, or the application, in some form, of the right of public performance under the existing copyright statute, let us consider the various types of computer programs that have been developed over the years. Perhaps the most widely useful type of computer program has been the word processing program. Would a right of access or public performance apply to such ubiquitous programs? Should there be limitations on any new exclusive right to access under section 117 of the copyright law?

Most every computer program contains some understanding, whether implicit or explicit, that is crucial to performing a particular task. However, most informed people would probably agree that, if there is knowledge embedded in a word processing program, the knowledge is pretty trivial: like one word follows another; or text is made up of concatenated sentences (i.e., one after another).

When you purchase a copy of a word processing program, you are buying a product that contains the embedded (and usually implicit) knowledge. You are acquiring a program that allows you to manipulate text. This function—text manipulation—is what is purchased and what producers have agreed to sell.

Where a word processing program is capable of taking what someone wrote, and recasting it to read more clearly or more understandably, you enter an area where the software is not just performing a straightforward and clearly delineated processing task. The program requires more complex knowledge of the form that may not be known or accessible to everyone. The program needs to know something about writing style, rules of grammar, some characteristics of different types of documents, as well as the intended audience for a particular work (e.g., if it is to be submitted to a particular court, it must know the format and semantics required by the court). While some of this knowledge may be in the form of heuristics and facts, the work being protected is still a computer program.

When you reach the stage of software development illustrated by the weather models, the situation becomes chaotic (defined as a "confused unorganized state existing before the creation of distinct forms"). Unlike static programs, where the data are embedded over a long period of time, the data in a weather model may be externally updated minute by minute.

Conclusion

If you have a system that embodies a database containing factual data, a knowledge base containing heuristics, and a science program or mathematical model,[6] it is possible to view the system as an integrated database system or an augmented mathematical model. It may be preferable, however, to view such works as knowledge-based systems,[7] and reach a consensus on the scope of copyright protection for these complex computer programs that encourages further development of these works of authorship, while assuring access by the public to the knowledge embedded therein.

NOTES

1. A. Newell, "The Models Are Broken, the Models Are Broken," 47 *U. of Pittsburgh L. Rev.* 1023, 1033 (1986).

2. As defined in section 101 of the copyright law, title 17 U.S.C., "a 'computer program' is a set of statements or instructions to be used directly or indirectly in a computer in order to bring about a certain result." For purposes of this definition, the word "statements" has been viewed as "data." See Stern Electronics, Inc. v. Kaufman, 669 F.2d 852, 855 (2d Cir. 1986).

3. See U.S. Congress, Office of Technology Assessment, Seeking Solutions: High Performance Computing for Science—Background Paper, OTA-BP-TCT-77, at 7 (1991) (Supertyphoon Hope).

4. For a detailed description of the effort known as Cyc, see D.B. Lenat and R.V. Guha, *Building Large Knowledge-Based Systems*. (Reading, MA: Addison-Wesley, 1990).

5. D.B. Lenat, R.V. Guha, K. Pittman, D. Pratt, and M. Shepherd, "CYC: Toward Programs with Common Sense," 33 *Communications of the ACM* 30, 32 (1990).

6. Mathematical model: may be termed "algorithmic." It contains procedures (i.e., statements that the programs can execute exactly).

7. For background information on development of such systems, see W.S. Mark and R.L. Simpson, Jr., "Knowledge-Based Systems: An Overview," *IEEE Expert* 12 (1991).

(*continued from page 72*)

titles, the advantages may be few in early years.

For qualified users, the goal would certainly be to make more material available locally. If that does not happen, the system has failed. The broader public is not really affected directly, although in a national license system they could become a market.

Disadvantages and Barriers

For the publisher, the greatest potential disadvantage is clear: if the licensee either intentionally abuses the license or is unable to restrict use to authenticated users, then not only is the potential market diluted but pricing algorithms begin to fall apart. For site licenses to be affordable, the publisher needs some assurance that the license will be honored.

For the libraries, there are related hurdles. Universities in general have limited control over their very inventive user population. Universities can agree to use their best efforts to enforce restrictions on access and use, but in the end, if someone chooses to deliberately "liberate" the database and share it across the network, the university will not take any legal responsibility. Some effort must be put on a system design that effectively monitors extraordinary use (such as repeated downloading of large files) without infringing upon user privacy.

Libraries also have a concern about ownership of the licensed material. What happens if the license is terminated? Must everything licensed to that point be returned, as is so frequently required now? It should be possible to design licenses that will permit retention of material, so long as the license terms governing use remain in effect.

The barriers to creating site licenses—regional or national—are linked to these real or potential problems. The system will have to offer real benefits (cost savings ***and*** service improvements) and this possibility has not yet been proven. All participants in the transaction—publisher, licensee, and customer/user—must feel modestly protected financially. There need to be technical delivery standards and, as is clear, pricing and legal models. The pricing models are particularly difficult to create. None of these barriers is impossible, but it will take cooperation and hard work.

A Next Step

As a result of discussions at the March 1991 Coalition for Networked Information meeting, a group of 15 universities has volunteered to try to deal with these licensing issues. The goal is to start small: individual university or university system site licenses for a cluster of journals. The universities are working now with Elsevier as the initial test publisher. This will be an intensive program of at least three years, aimed at understanding and overcoming precisely the barriers described above. As the program moves along, the critical question must be: are we improving service without increasing cost? If the answer is "no," then a different approach must be tried that will be win-win for the university, the publisher, and, more importantly, the user.