Knowledge management technology

by A. D. Marwick

Selected technologies that contribute to knowledge management solutions are reviewed using Nonaka's model of organizational knowledge creation as a framework. The extent to which knowledge transformation within and between tacit and explicit forms can be supported by the technologies is discussed, and some likely future trends are identified. It is found that the strongest contribution to current solutions is made by technologies that deal largely with explicit knowledge, such as search and classification. Contributions to the formation and communication of tacit knowledge, and support for making it explicit, are currently weaker, although some encouraging developments are highlighted, such as the use of text-based chat, expertise location, and unrestricted bulletin boards. Through surveying some of the technologies used for knowledge management, this paper serves as an introduction to the subject for those papers in this issue that discuss technology.

The goal of this paper is to provide an overview of technologies that can be applied to knowledge management and to assess their actual or potential contribution to the basic processes of knowledge creation and sharing within organizations. The aim is to identify trends and new developments that seem to be significant and to relate them to technology research in the field, rather than to provide a comprehensive review of available products.

Knowledge management (see, for example, Davenport and Prusak[1]) is the name given to the set of systematic and disciplined actions that an organization can take to obtain the greatest value from the knowledge available to it. "Knowledge" in this context includes both the experience and understanding of the people in the organization and the information artifacts, such as documents and reports, available within the organization and in the world outside. Effective knowledge management typically requires an appropriate combination of organizational, social, and managerial initiatives along with, in many cases, deployment of appropriate technology. It is the technology and its applicability that is the focus of this paper.

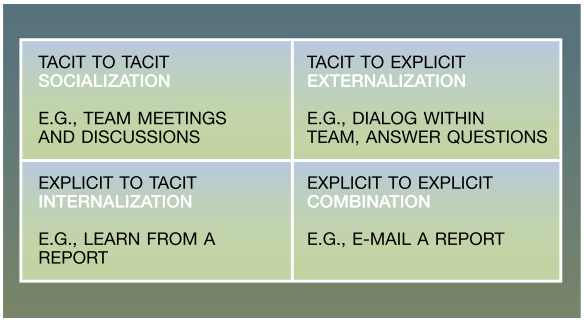
To structure the discussion of technologies, it is helpful to classify the technologies by reference to the notions of tacit and explicit knowledge introduced by Polanyi in the 1950s[2,3] and used by Nonaka[4,5] to formulate a theory of organizational learning that focuses on the conversion of knowledge between tacit and explicit forms. Tacit knowledge is what the knower knows, which is derived from experience and embodies beliefs and values. Tacit knowledge is actionable knowledge, and therefore the most valuable. Furthermore, tacit knowledge is the most important basis for the generation of new knowledge, that is, according to Nonaka: "the key to knowledge creation lies in the mobilization and conversion of tacit knowledge."[5] Explicit knowledge is represented by some artifact, such as a document or a video, which has typically been created with the goal of communicating with another person. Both forms of knowledge are important for organizational effectiveness.[6]

 0018-8670/01/$5.00 © 2001 IBM

These ideas lead us to focus on the processes by which knowledge is transformed between its tacit and explicit forms, as shown in Figure 1.[5] Organizational learning takes place as individuals participate in these processes, since by doing so their knowledge is shared, articulated, and made available to others. Creation of new knowledge takes place through the processes of combination and internalization. As shown in Figure 1, the processes by which knowledge is transformed within and between forms usable by people are

- *Socialization (tacit to tacit):* Socialization includes the shared formation and communication of tacit knowledge between people, e.g., in meetings. Knowledge sharing is often done without ever producing explicit knowledge and, to be most effective, should take place between people who have a common culture and can work together effectively (see Davenport and Prusak,[1] p. 96). Thus tacit knowledge sharing is connected to ideas of communities and collaboration. A typical activity in which tacit knowledge sharing can take place is a team meeting during which experiences are described and discussed.
- *Externalization (tacit to explicit):* By its nature, tacit knowledge is difficult to convert into explicit knowledge. Through conceptualization, elicitation, and ultimately articulation, typically in collaboration with others, some proportion of a person's tacit knowledge may be captured in explicit form. Typical activities in which the conversion takes place are in dialog among team members, in responding to questions, or through the elicitation of stories.
- *Combination: (explicit to explicit):* Explicit knowledge can be shared in meetings, via documents, e-mails, etc., or through education and training. The use of technology to manage and search collections of explicit knowledge is well established. However, there is a further opportunity to foster knowledge creation, namely to enrich the collected information in some way, such as by reconfiguring it, so that it is more usable. An example is to use text classification to assign documents automatically to a subject schema. A typical activity here might be to put a document into a shared database.
- *Internalization (explicit to tacit):* In order to act on information, individuals have to understand and internalize it, which involves creating their own tacit knowledge. By reading documents, they can to some extent re-experience what others previously learned. By reading documents from many sources, they have the opportunity to create new knowledge by combining their existing tacit knowledge with the knowledge of others. However, this process is becoming more challenging because individuals have to deal with ever-larger amounts of information. A typical activity would be to read and study documents from a number of different databases.

Figure 1 Conversion of knowledge between tacit and explicit forms (after Nonaka[5])



These processes do not occur in isolation, but work together in different combinations in typical business situations. For example, knowledge creation results from interaction of persons and tacit and explicit knowledge. Through interaction with others, tacit knowledge is externalized and shared.[7] Although individuals, such as employees, for example, experience each of these processes from a knowledge management and therefore an organizational perspective, the greatest value occurs from their combination since, as already noted, new knowledge is thereby created, disseminated, and internalized by other employees who can therefore act on it and thus form new experiences and tacit knowledge that can in turn be shared with others and so on.[7] Since all the processes of Figure 1 are important, it seems likely that knowledge management solutions should support all of them, although we must recognize that the balance between them in a particular organization will depend on the knowledge management strategy used.[8]

Table 1 shows some examples of technologies that may be applied to facilitate the knowledge conversion processes of Figure 1. These technologies and others are discussed in this paper. The individual technologies are not in themselves knowledge management solutions. Instead, when brought to mar-

Figure 2 Information extraction using template filling[11]

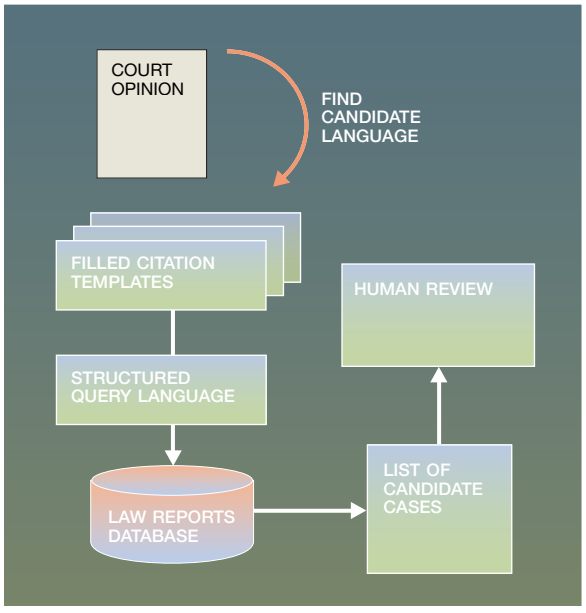


Table 1 Examples of technologies that can support or enhance the transformation of knowledge

Tacit to Tacit	**Tacit to Explicit**
E-meetings	Answering questions
Synchronous collaboration (chat)	Annotation
Explicit to Tacit	**Explicit to Explicit**
Visualization	Text search
Browsable video/audio of presentations	Document categorization

ket they are typically embedded in a smaller number of solutions packages, each of which is designed to be adaptable to solve a range of business problems. Examples are portals, collaboration software, and distance learning software. Each of these can and does include several different technologies.

The approach to the technology of knowledge management in this paper emphasizes human knowledge. Sometimes in computer science "knowledge management" is interpreted to mean the acquisition and use of knowledge by computers, but that is not the meaning used here. In any case, automatic extraction of deep knowledge (i.e., in a form that captures the majority of the meaning) from documents is an elusive goal. Today the level of automatic extraction is deemed to be rather shallow because only a subset of the meaning, sometimes a very limited one, can be captured, ranging from recognition of entities such as proper names or noun phrases to automatic extraction of ontological relations of various kinds (e.g., References 9 and 10), and there is no system that can reason (in the sense of deducing something new from what it already knows) over the extracted knowledge in a way that even approaches the capabilities of a human. As an example of the current state of the art in applications for extracting knowledge automatically, Figure 2 shows a system[11] for analyzing reports of appellate court decisions to find the precedents they may affect. Court opinions are analyzed to find language that refers to other cases that the opinion may modify or invalidate. The candidate cases are retrieved from a database of law reports and are presented to an analyst for final judgment. The results are used to enrich the database with appropriate cross-references. Here the approach is that a template defines the fragment of knowledge to be sought, and the system tries to fill it by extracting information from the text. However, the candidate pieces of extracted knowledge must still be presented to a human for review and final decision, so that the value of the system is in increasing the productivity of the human analysts. For the foreseeable future, knowledge management in business will be about human knowledge in its various forms.

The use of technology in knowledge management is not new, and considerable experience has been built up by the early pioneers. Even before the availability of solutions such as Lotus Notes**[12] on which many contemporary knowledge management solutions are based, companies were deploying intranets, such as EPRINET,[13] based on early generations of networking and computer technology that improved access to knowledge "on line." Collaboration and knowledge sharing solutions also arose from the development of on-line conferencing and forums[14] using mainframe computer technology. Today, of course, intranets and the Internet are ubiquitous, and we are rapidly approaching the situation where all the written information needed by a person to do his or her job is available on line. However, that is not to say that it can be used effectively with the tools currently available.

It is important to note that knowledge management problems can typically not be solved by the deployment of a technology solution alone. The greatest difficulty in knowledge management identified by the

respondents in a survey[15] was "changing people's behavior," and the current biggest impediment to knowledge transfer was "culture." Overcoming technological limitations was much less important. The role of technology is often to overcome barriers of time or space that otherwise would be the limiting factors. For example, a research organization divided among several laboratories in different countries needs a system that scientists with common interests can use to exchange information with each other without traveling, whereas a document management system can ensure that valuable explicit knowledge is preserved so that it can be consulted in the future. Two caveats must be stated at this point. First is the point made by Ackerman[16] that in many respects the state of the art is such that many of the social aspects of work important in knowledge management cannot currently be addressed by technology. Ackerman refers to this situation as a "social technical gap." Second, the coupling between behavior and technology is two-way: the introduction of technology may influence the way individuals work. People can and do adapt their way of working to take advantage of new tools as they become available, and this adaptation can produce new and more effective communication within teams (e.g., the effect of introducing solutions based on Lotus Notes on process teams in a paper mill described by Robinson et al.[17] or the adaptations made by people in a customer support organization studied by Orlikowski[18] after Notes was introduced).

Other surveys of technology for knowledge management can be found in the book, *Working Knowledge* by Davenport and Prusak,[1] and in a paper by Jackson.[19] Prospects for using artificial intelligence (AI) techniques in knowledge management have been discussed recently by Smith and Farquhar.[20]

In the following sections of this paper the technologies that support the processes of Figure 1 are described in more detail and illustrated with examples drawn largely from current research projects.

Tacit to tacit

The most typical way in which tacit knowledge is built and shared is in face-to-face meetings and shared experiences, often informal, in which information technology (IT) plays a minimal role. However, an increasing proportion of meetings and other interpersonal interactions use on-line tools known as groupware. These tools are used either to supplement conventional meetings, or in some cases to replace them. To what extent can these tools facilitate formulation and transfer of tacit knowledge?

Groupware. Groupware is a fairly broad category of application software that helps individuals to work together in groups or teams. Groupware can to some extent support all four of the facets of knowledge transformation. To examine the role of groupware in socialization we focus on two important aspects: shared experiences and trust.

Shared experiences are an important basis for the formation and sharing of tacit knowledge.

Shared experiences are an important basis for the formation and sharing of tacit knowledge. Groupware provides a synthetic environment, often called a virtual space, within which participants can share certain kinds of experience; for example, they can conduct meetings, listen to presentations, have discussions, and share documents relevant to some task. Indeed, if a geographically dispersed team never meets face to face, the importance of shared experiences in virtual spaces is proportionally enhanced. An example of current groupware is Lotus Notes,[12] which facilitates the sharing of documents and discussions and allows various applications for sharing information and conducting asynchronous discussions to be built. Groupware might be thought to mainly facilitate the combination process, i.e., sharing of explicit knowledge. However, the selection and discussion of the explicit knowledge to some degree constitutes a shared experience.

A richer kind of shared experience can be provided by applications that support real-time on-line meetings—a more recent category of groupware. On-line meetings can include video and text-based conferencing, as well as synchronous communication and chat. Text-based chat is believed to be capable of supporting a group of people in knowledge sharing in a conversational mode.[21] Commercial products of this type include Lotus Sametime** and Microsoft NetMeeting**. These products integrate both instant messaging and on-line meeting capabilities. Instant

Table 2 Sources of evidence for an expertise location system

A profile or form filled in by a user
An existing company database, for example one held by the Human Resources department
Name-document associations
Questions answered

messaging is found to have properties between those of the personal meeting and the telephone: it is less intrusive than interrupting a person with a question but more effective than the telephone in broadcasting a query to a group and leaving it to be answered later.

In work on the Babble system,[22] chat was evaluated by at least some users as being ". . . much more like conversation," which is promising for the kind of dialog in which tacit knowledge might be formed and made explicit. However, not all on-line meeting systems have the properties of face-to-face meetings. For example, the videoconferencing system studied by Fish et al.[23] was judged by its users to be more like a video telephone than like a face-to-face meeting. Currently, rather than replacing face-to-face meetings, many on-line meetings are found to complement existing collaboration systems and the well-established phone conference and are therefore probably more suited to the exchange of explicit rather than tacit knowledge. On-line meetings extend phone conferences by allowing application screens to be viewed by the participants or by providing a shared whiteboard. An extension is for part of the meeting to take place in virtual reality with the participants represented by avatars.[24] One research direction is to integrate on-line meetings with classic groupware-like applications that support document sharing and asynchronous discussion. An example is the IBM-Boeing TeamSpace project,[25] which helps to manage both the artifacts of a project and the processes followed by the team. On-line meetings are recorded as artifacts and can be replayed within TeamSpace, thus allowing even individuals who were not present in the original meeting to share some aspects of the experience.

Some of the limitations of groupware for tacit knowledge formation and sharing have been highlighted by recent work on the closely related issue of the degree of trust established among the participants.[26] It was found that videoconferencing (at high resolution—not Internet video) was almost as good as face-to-face meetings, whereas audio conferencing was less effective and text chat least so. These results suggest that a new generation of videoconferencing might be helpful in the socialization process, at least in so far as it facilitates the building of trust. But even current groupware products have features that are found to be helpful in this regard. In particular, access control, which is a feature of most commercial products, enables access to the discussions to be restricted to the team members if appropriate, which has been shown[22] to encourage frankness and build trust.

Another approach to tacit knowledge sharing is for a system to find persons with common interests, who are candidates to join a community. In Foner's Yenta System,[27] the similarity of the documents used by people allowed the system to infer that their interests were similar. Location of other people with similar interests is a function that can be added to personalization systems, the goal of which is to route incoming information to individuals interested in it. There are obvious privacy problems to overcome.

Expertise location. Suppose one's goal is not to find someone with common interests but to get advice from an expert who is willing to share his or her knowledge. Expertise location systems have the goal of suggesting the names of persons who have knowledge in a particular area. In their simplest form, such systems are search engines for individuals, but they are only as good as the evidence that they use to infer expertise. Some possible sources of such evidence are shown in Table 2.

The problem with using an explicit profile is that persons may not be motivated to keep it up to date, since to them it is just another form to fill in. Thus it is preferable to gather information automatically, if possible, from existing sources. For example, a person's resume or a list of the project teams that he or she has worked on may exist in a company database. Another automatic approach is to infer expertise from the contents of documents with which a person's name is associated. For example, authorship (creation or editing) of a document presumably indicates some familiarity with the subjects it discusses, whereas activities such as reading indicate some interest in the subject matter. Two approaches to using document evidence for expertise location suggest themselves: either the documents can be classified according to some schema, thus classifying their authors; or when a user submits a query to the expertise location system, it searches the documents,

transforms the query to a list of authors (suitably weighted), and returns the list as the result of the expertise search.

The current state of the art is to use the first three sources of evidence listed in Table 2: explicit profiles, evidence mined from existing databases, and evidence inferred from association of persons and documents. For example, the Lotus Discovery Server** product contains a facility whereby an individual's expertise is determined using these techniques,[28] while it and the Tacit Knowledge Systems KnowledgeMail** product[29] analyze the e-mail a person writes to form a profile of his or her expertise. Given the properties of on-line discussions, discussed below, it is reasonable to suppose that a fourth source of evidence could be the content of the questions answered by a person in such a system, with the added advantage that such a person is already willing to be helpful. This example is a simple case of the social interaction dimension in expertise location which, as found in empirical studies (e.g., Reference 30), is an important factor but is not yet reflected in available applications, perhaps because of the difficulty of capturing aspects such as the expert's communication skills, in order to rate how useful he or she is likely to be.

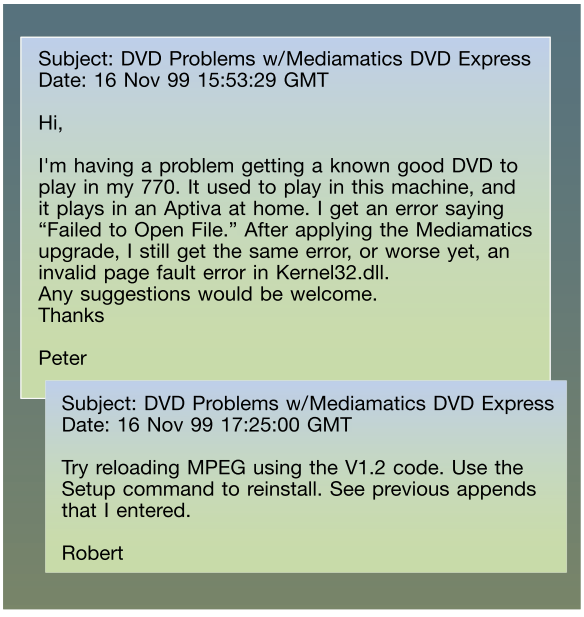
Tacit to explicit

According to Nonaka, the conversion of tacit to explicit knowledge (externalization) involves forming a shared mental model, then articulating through dialog. Collaboration systems and other groupware (for example, specialized brainstorming applications[31]) can support this kind of interaction to some extent.

On-line discussion databases are another potential tool to capture tacit knowledge and to apply it to immediate problems. We have already noted that team members may share knowledge in groupware applications. To be most effective for externalization, the discussion should be such as to allow the formulation and sharing of metaphors and analogies, which probably requires a fairly informal and even freewheeling style. This style is more likely to be found in chat and other real-time interactions within teams.

Newsgroups and similar forums are open to all, unlike typical team discussions, and share some of the same characteristics in that questions can be posed and answered, but differ in that the participants are typically strangers. Nevertheless, it is found that

Figure 3 An example of an exchange in an internal company forum



many people who participate in newsgroups are willing to offer advice and assistance, presumably driven by a mixture of motivations including altruism, a wish to be seen as an expert, and the thanks and positive feedback contributed by the people they have helped.

Within organizations, few of the problems experienced on Internet newsgroups are found, such as flaming, personal abuse, and irrelevant postings. IBM's experience in this regard is described by Foulger.[14] Figure 3 shows a typical exchange in an internal company forum, rendered here using a standard newsgroup browsing application. It illustrates how open discussion groups are used to contribute knowledge in response to a request for help. Note both the speed of response and the fact that the answerer has made other contributions previously. The archive of the forum becomes a repository of useful knowledge. Clearly the question answerer in this case has made a number of contributions and could be considered to be an expert. Although the exchange is superficially one of purely explicit knowledge, the expert must first make a judgment as to the nature of the problem and then as to the most likely solution, both of which bring his or her tacit knowledge into play. Once the knowledge is made explicit, persons with similar problems can find the solution by consulting the archive. A quantitative study[32] of this

phenomenon in the IBM system showed that the great majority of interchanges were of this question-and-answer pattern, and that even though a large fraction of questions were answered by just a few persons, an equal proportion were answered by persons who only answered one or two questions. Thus the conferencing facility enabled knowledge to be elicited from the broad community as well as from a few experts.

Explicit to explicit

There can be little doubt that the phase of knowledge transformation best supported by IT is combination, because it deals with explicit knowledge. We can distinguish the challenges of knowledge management from those of information management by bearing in mind that in knowledge management the conversion of explicit knowledge from and to tacit knowledge is always involved. This leads us to emphasize new factors as challenges that technology may be able to address.

Capturing knowledge. Once tacit knowledge has been conceptualized and articulated, thus converting it to explicit knowledge, capturing it in a persistent form as a report, an e-mail, a presentation, or a Web page makes it available to the rest of the organization. Technology already contributes to knowledge capture through the ubiquitous use of word processing, which generates electronic documents that are easy to share via the Web, e-mail, or a document management system. Capturing explicit knowledge in this way makes it available to a wider audience, and "improving knowledge capture" is a goal of many knowledge management projects. One issue in improving knowledge capture is that individuals may not be motivated to use the available tools to capture their knowledge. Technology may help by improving their motivation or by reducing the barriers to generating shareable electronic documents.

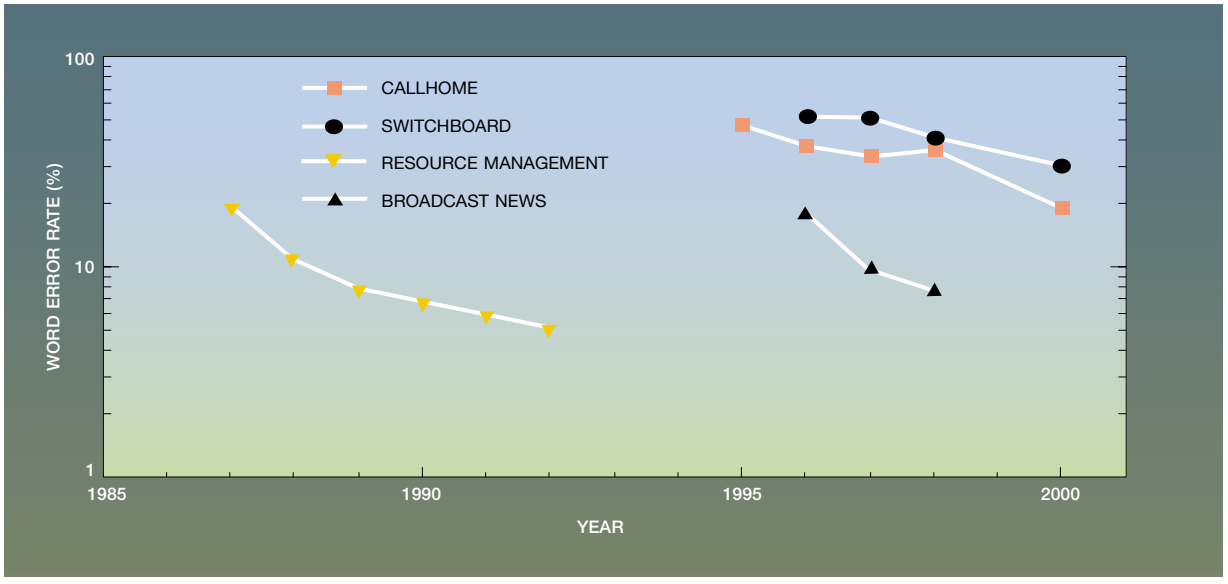
One way to motivate people to capture knowledge is to reward them for doing so. If rewards are to be linked to quality rather than quantity, some way to measure the quality of the output is needed. Quality in the abstract is extremely difficult to assess, since it depends on the potential use to which the document is to be put. For example, a document that explains basic concepts clearly would be useful for a novice but useless to someone who is already an expert. If we focus on usefulness as a measure of quality, and if we substitute "use" for "usefulness," then we have something that IT systems can measure. In fact, portal infrastructures that mediate access to documents can easily accumulate metrics of document use, and hence can estimate usefulness and quality. The next generation of products will include such features.[28]

Another measure of quality is the number of times a document has been cited, as in the scholarly literature, or the number of times it has been hyperlinked to, as on the Internet. A citation or hyperlink is evidence that the author of the citing or linking document thought that the target document is valuable. The most valuable or authoritative documents can be detected in Internet applications by analyzing the links between Web pages, thus measuring the cumulative effects of numerous value judgments (e.g., see References 33 and 34). The numeric quality estimate that can be derived is useful in information retrieval, where it can be used to boost the position of high-quality documents in the search results list. This method has been applied to citation analysis in scientific papers by the ResearchIndex search engine[35,36] and to Web search by the Google search engine.[37]

Citation analysis of this kind detects quality assessments made in the course of authoring documents. Quality judgments by experts are another way to capture their knowledge. There are, of course, many deployed solutions in which documents undergo a quality review through a refereeing process, often facilitated by a workflow application. In this case, the quality judgment acts as a gate, and documents judged to be of low quality are not distributed. However, technology also makes it feasible to record judgments as annotations of existing documents.[38] Here, the association of an annotation with a document is recorded in some infrastructure, such as a special annotation server that the user's browser accesses to find annotations of the Web page being viewed. Numeric data stored in databases can also be annotated[39] to record various interpretations, judgments, or cautions. Annotations may also support collaboration around documents,[40] although, as in other applications where the underlying documents may be altered, the annotation system needs to be robust in the face of changes.

Although the most common way to capture knowledge by far is to write a document, technology has made the use of other forms of media feasible. Digital audio and video recordings are now easily made, and an expert may find that speaking to a camera or microphone is easier or more convenient than

Figure 4 Improvement in various automatic speech transcription tasks over time



writing, particularly if the video is of a presentation that has to be made in the ordinary course of business, or if the audio recording can be made in an otherwise unproductive free moment. It is also now relatively easy to distribute audio and video over networks. However, nontext digital media have the disadvantage of being more difficult to search and to browse than text documents and, hence, are less usable as materials in a repository of knowledge. Browsing of video has been improved by summarization techniques that automatically produce a gallery of extracted still images, each of which represents a significant passage in the video.[41] If the video is of someone giving a presentation, images of the speaker alone will not convey as much as a summary that includes images of any visual aids, such as slides or charts, that accompany the narrative. Several systems that key a recording of a presentation to the slides have been described.[42–44]

Although video searching systems have been built that use image searching[45] of extracted frames,[46,47] they are hampered by the difficulty of composing a semantically meaningful image query. A more fruitful approach to searching is to extract text from the multimedia object, if possible. Although in some cases the video may contain text (on images of text slides), in most cases the challenge is to convert speech to text.

Speech recognition. Improvements in the accuracy of automatic speech recognition (ASR) hold out the promise of usable speaker-independent recognition with unconstrained vocabulary in the foreseeable future. Figure 4 shows progress with time in a number of standardized speech recognition tasks. Word error rates were reported in the Speech Recognition Workshop conferences of the National Institute of Standards and Technology. The accuracy varies with the difficulty of the task. The resource management task involves reading speech with a 1000-word vocabulary. Broadcast news uses recordings with an approximately 20K word vocabulary, whereas the CallHome and switchboard are telephone (lower speech quality) recognition tasks with unconstrained vocabulary. In all cases the accuracy shows steady improvement with time.

Accuracy for speech recorded under controlled conditions is already acceptable, but the error rate for poor quality recordings (for example, from the telephone) is still high enough to cause problems for applications unless the vocabulary is constrained. However, the trends depicted in Figure 4 show that future improvements can reasonably be expected and will lead to new ways to capture knowledge.

Although perfect or near-perfect transcription produces a text transcript that can be browsed like any

other piece of text, ways to make an imperfect transcript usable as a browsing aid are being investigated.[48,49] In this work even an imperfect transcript supports browsing because certain words and phrases, which are judged to be significant and for which the estimated accuracy of ASR is high, are highlighted.

The most common problem in a search is that a query retrieves many documents irrelevant to the user's needs.

Such techniques can be used to make the replay of audio more usable even where the transcript as a whole is unreadable because of the density of errors. The highlights can be used to find the passage of interest.

Search. The most important technology for the manipulation of explicit knowledge helps people with the most basic task of all: finding it. Since the trend in most organizations is for essentially all documents to become available in electronic form on line, the challenge of on-line access has been transformed into the challenge of finding the materials relevant for some task. Furthermore, the total amount of potentially relevant information, including what is on the Internet and company intranets and what is available from commercial on-line publishers, continues to grow rapidly. Thus text search, which only 10 years ago was a tool primarily used by librarians to search bibliographic databases, has become an everyday application used by almost everyone. Not surprisingly, the new uses of text search have motivated new work on the technology.

Another driving factor in the use of on-line explicit knowledge is the diversity of sources from which it is available. It is not uncommon for users to have to look in several databases or Web sites for potentially relevant information. Since there is little standardization, users have to cope with different user interfaces, different search language conventions, and different result list presentations. Portals—described in another paper in this issue[50]—are a popular approach to reducing the complexity of the user's task. The key aspect that allows a portal to do this is that it maintains its own meta-data about the information to which it gives access. In the current state of the art, the meta-data may be quite simple, consisting of a list of sources and a search index formed from the content of the sources. Even this simple function provides great value because it relieves the user of the need to visit all the sources to find out whether they contain relevant information. The user is therefore made more productive, and the quality of his or her work is improved. Most portal systems use a single search index, which requires that the documents in the domain of interest have to be retrieved by "spidering" or "crawling" at indexing time. The alternative, using distributed search as in, for example, the Harvest project,[51] has not proved to be popular for knowledge management applications, perhaps because advances in hardware have made it cheaper to build a central index. Recent developments in peer-to-peer applications, such as Gnutella[52] and the collaboration application Groove,[53] have promoted a new interest in distributed search, which may lead to new advances.

The index that is built by a text search engine consists of a list of the words that occur in the indexed documents, along with a data structure (the inverted file) that allows the documents in which the words occurred to be determined efficiently at search time.[54] Users can therefore use query words that they expect to occur in the documents. The problem is that not all the documents will use the same words to refer to the same concept and, therefore, not all the documents that discuss the concept will be retrieved. In a world of information overload this situation is not usually a problem, but for applications where it is important to have high recall, an alternative approach can be used in which documents are assigned meta-data that describe the concepts they discuss in a controlled vocabulary. This is a classical approach used in bibliographic databases. However, where searches are being done by untrained end users rather than librarians, the evidence is that searching with natural language gives better results than does searching with a controlled vocabulary.[55]

The most common problem in a search is that a query retrieves many documents that are irrelevant to the user's needs, known as the problem of search precision (a measure of accuracy). Precision is of paramount importance in a world of "info-glut." However, results from TREC (Text REtrieval Conference)[56] indicate that the accuracy of natural language search engine technology has reached a plateau in recent years. What are the prospects of improvements to the search function that will benefit knowledge management

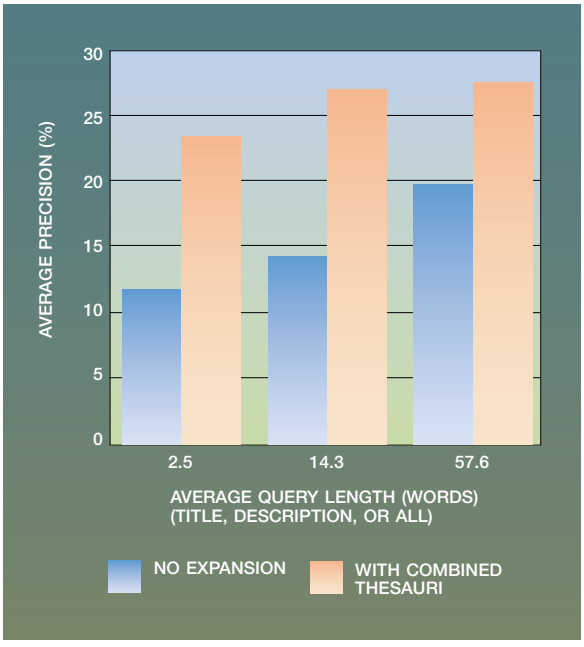
systems? Two areas of potential improvement can be identified: increased knowledge of the user and of the context of his or her information need, and improved knowledge of the domain being searched.

The notion that increased knowledge of the user can be beneficial comes from the realization that in almost all search systems today the only information about the user's information need that is available to the system is the query. The most common query submitted to Web-based search services is two words, and the average query length is only about 2.3 words.[57] Obviously, this amount is not much information. A challenging research area is to gather better information about the context of a search and to build search engines that can use this information to good advantage.

The goal of gathering and using more information about the domain being searched is one that is well-established, but progress so far has been limited. It is common to use a thesaurus—a kind of simple domain model—as an adjunct to a search, although this is more common in systems designed for specialists. Expansion of a query with synonyms is known to improve the recall in a text search, but expansion is only effective in well-defined domains where the ambiguity of words, and the validity of term relationships, is not an issue. To improve precision in broad-domain searching by reducing the ambiguity of ordinary words using thesauri or other structures such as ontologies has been a goal of much research, with many negative results (e.g., Reference 58). Recently, however, some encouraging findings have been obtained.[54] Using WordNet[59] (a large manually built thesaurus that is widely available), combined with automatically built data structures encoding co-occurrence and head-modifier relations, Mandala et al.[60] showed significant improvements in average precision, a measure of accuracy, as shown in Figure 5. The results were obtained using TREC data, from queries derived from the search topics using the title field, the title and description fields, or all the fields in the topic. Woods et al.[61] also reported improvements by using a different approach to encoding knowledge of the domain, in this case a semantic network that integrated syntactic, semantic, and morphological relationships.

Taxonomies and document classification. Knowledge of a domain can also be encoded as a "knowledge map," or "taxonomy," i.e., a hierarchically organized set of categories. The relationships within the hierarchy can be of different kinds, depending on the application, and a typical taxonomy includes several different kinds of relations. The value of a taxonomy is twofold. First, it allows a user to navigate to documents of interest without doing a search (in practice, a combination of the two strategies is often used if it is available). Second, a knowledge map allows documents to be put in a context, which helps users to assess their applicability to the task in hand. The most familiar example of a taxonomy is Yahoo!,[62] but there are many examples of specialized taxonomies used at other sites and in company intranet applications.

Figure 5 Improved average precision in text search using combined thesauri for query expansion[60]



Manually assigning documents to the categories in a taxonomy requires significant effort and cost, but in recent years automatic document classification has advanced to the point where the accuracy of the best-performing algorithms exceeds 85 percent (F_1 measure) on good quality data.[63] This degree of accuracy is adequate for many applications and is in fact comparable to what can be achieved by manual classifiers in a well-organized operation,[64] although the accuracy of automatic classification over different types of data varies quite widely.[65] An attractive feature of the current generation of automatic classifiers is their inclusion of machine-learning algorithms

Table 3 Essay grading with an automatic text classifier[66]

	Exact Grade (%)	Adjacent Grade (%)
G1: auto vs manual*	55	97
G1: manual A vs B	56	95
G2: auto vs manual*	52	96
G2: manual A vs B	56	95

*The performance of the classifier is compared with two human markers, A and B, and it performs almost as well. In each comparison, the proportion of test essays where the same or an adjacent grade was assigned is given. Here "manual" refers to the average of the two human graders, whereas G1 and G2 are two open-domain essay-writing tasks.

that train themselves from example data, whereas the previous generation required construction of a complex description of the category in the form, for example, of an elaborate query. Selecting documents as training examples is a simpler task.

Automatic classification, although simple in concept, is capable of surprisingly refined distinctions, given enough training data. For example, it has been known for some time (see the brief review in Kukich[66]) that automatic essay marking systems can assign grades to student essays with an accuracy and consistency only slightly worse than human graders, and recently it has been shown that a document classifier can perform well in this application.[67] Table 3 shows the results of comparing two human graders and an automatic classifier. The automatic classifier performed very nearly as well as the human graders, both in accuracy and consistency, even though the test essays were on unconstrained subjects.

Despite the power of automatic classification, there are many challenges in implementing solutions using taxonomies. The first challenge is the design of the taxonomy, which has to be comprehensible to users (so that they can use it for navigation with no or minimal training) and has to cover the domain of interest in enough detail to be useful. There are a number of strategies for building a taxonomy,[68] including the use of document clustering to propose candidate subcategories. However, human input is probably required to ensure that the taxonomy reflects business needs (e.g., it emphasizes some aspect that may be significant but is not a strong theme in the documents). Thus, clustering can be seen as an adjunct to human effort. One usability challenge is to ensure that the user of a taxonomy editor can understand the clusters that are proposed, using automatically generated labels. The labels typically contain words or phrases that are chosen to represent the documents in the cluster; recently a technique for using extracted sentences has been proposed.[69,70]

Taxonomies have proved to be a popular way in which to build a domain model to help users to search and navigate, so much so that the trend seems to be for each group of users of any size to have their own taxonomy. This popularity is understandable because as on-line tools become central to individuals' work, they naturally want to see the information displayed within a schema that reflects their own priorities and worldview, and that uses the terminology that they use. This trend is likely to lead to a proliferation of taxonomies in knowledge management applications. It follows that there will be an increasing focus on the need to map from one taxonomy to another so as to bridge between the schemas used by different groups within an organization.

Portals and meta-data. As already mentioned, portals provide a convenient location for the storage of meta-data about documents in their domain, and two examples of such meta-data, search indexes and a knowledge map or taxonomy, have been discussed. In the future, increasing use of natural language processing (NLP) in portals is likely to generate new kinds of meta-data. The general trend is for more structured information—meta-data—to be automatically generated as part of the indexing service of the portal. It is efficient to generate these meta-data when the document has been retrieved for text indexing. The value of the meta-data is in encapsulating information about the document that can be used to build selected views of the information space, such as a list of the documents in a given subject category, or mentioning a geographic location, through a database lookup in response to a user click. This makes exploration of the information easier and more rewarding, in effect providing the user with a new experience based on the exploration on which new tacit knowledge can be built as part of the internalization process to be discussed later.

Summarization. Document summaries are examples of meta-data of this kind. The value of a summary is that it allows users to avoid reading a document if it is not relevant to their current tasks. Figure 6 shows results from Tombros and Sanderson[71] who showed that users performing a simple information-seeking task had to read many fewer full documents when they used a system that provided summaries than when the system provided document titles alone. Automatic generation of summaries is an ac-

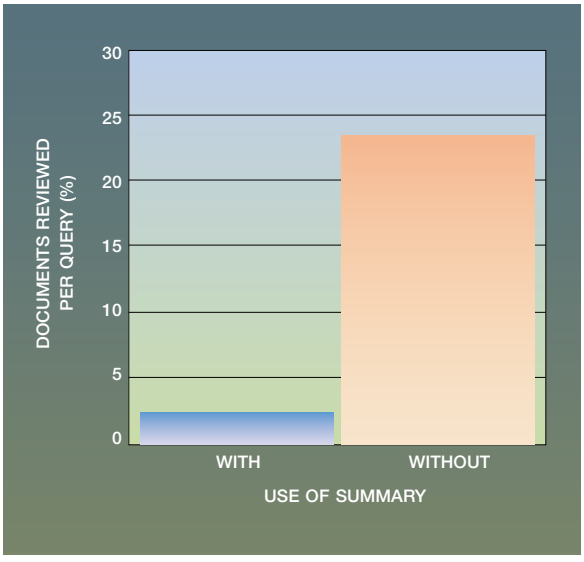
tive area of research. Commercially available summarizers use the sentence-selection method, originated by Luhn in 1958,[72] in which an indicative summary is constructed from what are judged to be the most salient sentences in a document. However, the summary may be incoherent, e.g., if the selected sentences contain anaphors. Construction of more coherent summaries, implying the use of natural language generation, currently requires that the subject domain of the documents be severely restricted, as for example, to basketball games.[73] Summarization of long documents containing several topics is improved by topic segmentation[74] and can be further condensed for presentation on handheld devices,[75] whereas summarization of multiple documents, either about the same event[76] or in an unconstrained set of domains,[70] is another challenge being addressed by current research. For other recent work see References 77 through 79.

Explicit to tacit

Technology to help users form new tacit knowledge, for example, by better appreciating and understanding explicit knowledge, is a challenge of particular importance in knowledge management, since acquisition of tacit knowledge is a necessary precursor to taking constructive action. A knowledge management system should, in addition to information retrieval, facilitate the understanding and use of information. For example, the system might, through document analysis and classification, generate metadata to support rapid browsing and exploration of the available information. It seems likely that the future trend will be for information infrastructures to perform more of this kind of processing in order to facilitate different modes of use of information (e.g., search, exploration, finding associations) and thus to make the information more valuable by making it easier to form new tacit knowledge from it. Other processing of explicit knowledge, already described, can support understanding. For example, putting a document in the context of a subject category or of a step in a business process, by using document categorization, can help a user to understand the applicability or potential value of its information. Discovery of relationships between and among documents and concepts helps users to learn by exploring an information space.

A quite different set of technologies applies to the formation of tacit knowledge through learning, especially in the domain of on-line education or distance learning. Within organizations, on-line learning has the advantage of being able to be accomplished without travel and at times that are compatible with other work. A wide variety of tools and applications support distance learning.[80] The needs of the corporate training market, emphasizing self-directed learning rather than instructor-led learning, have led to a focus on interactive courseware based on the Web or on downloaded applications. In the future, modules of self-directed training will be found in portals, along with other materials.

Figure 6 The proportion of documents read by subjects using information retrieval systems to perform a task[71]



Information overload is a trend that motivates the adoption of new technology to assist in the comprehension of explicit knowledge. The large amounts of (often redundant) information available in modern organizations, and the need to integrate information from many sources in order to make better decisions, cause difficulties for knowledge workers and others.[81] Both of these trends result directly from the large amounts of on-line information available to knowledge workers in modern organizations. Information overload occurs when the quality of decisions is reduced because the decision maker spends time reviewing more information than is needed, instead of reflecting and making the decision. Various approaches to mitigating information overload are feasible. The redundancy and repetition in the information can be reduced by eliminating duplicate

or overlapping messages (related to the Topic Detection and Tracking track at TREC[82]). An agent can filter or prioritize the messages, or compound views can make it easier to review the incoming information. Finally, visualization techniques can be applied in an attempt to help the user understand the available information more easily.

Different visualizations of a large collection of documents have been used with the goal of making subject-based browsing and navigation easier. These methods include text-based category trees, exemplified by the current Yahoo! user interface. Several graphical visualizations have also been described. Themescape[83] uses (among other things) a shaded topographic map as a metaphor to represent the different subject themes (by location), their relatedness (by distance), and the proportional representation of the theme in the collection (by height), whereas VisualNet[84] uses a different map metaphor for showing subject categories. Another approach is represented by the "Cat-a-Cone" system[85] that allows visualization of documents in a large taxonomy or ontology. In this system the model is three-dimensional and is rendered using forced perspective. Search is used to select a subset of the available documents for visualization.

Other visualization experiments have attempted to provide a user with some insight into which query terms occur in the documents in a results list, as was done in Hearst's TileBars[86] and the application described by Veerasamy and Belkin.[87] However, the evaluation described in the latter paper showed that the advantage of the visualization in the test task was small at best. A later study,[88] which compared text, two-dimensional, and pseudo three-dimensional interfaces for information retrieval, found that the richer interfaces provided no advantage in the search tasks that were studied. This result may explain why graphical visualization has not been widely adopted in search applications, whereas text-based interfaces are ubiquitous.

Perhaps a more promising application of visualization is to help a user grasp relationships, such as those between concepts in a set of documents as in the Lexical Navigation system described by Cooper and Byrd[89] or the relationships expressed as hyperlinks between documents.[90] This use is more promising because of the difficulty of rendering relationships textually. Furthermore, figuring out the relationships within a set of documents is a task that requires a lot of processing, and computer assistance is of great value.

Conclusion

This paper has surveyed a number of technologies that can be applied to build knowledge management solutions and has attempted to assess their actual or potential contributions to the processes underlying organizational knowledge creation using the Nonaka model. The essence of this model is to divide the knowledge creation processes into four categories: socialization (tacit knowledge formation and communication), externalization (formation of explicit knowledge from tacit knowledge), combination (use of explicit knowledge), and internalization (formation of new tacit knowledge from explicit knowledge). The value of this model in the present context is that it focuses attention on tacit knowledge (which is featured in three of the four processes) and thus on people and their use of technology.

Because all four of the processes in the Nonaka model are important in knowledge management, which aims to foster organizational knowledge creation, we might seek to support all of them with technology. Although early generations of knowledge management solutions (solutions typically integrate several technologies) focused on explicit knowledge in the form of documents and databases, there is a trend to expand the scope of the solutions somewhat to integrate technologies that can, to some extent, foster the use of tacit knowledge. Among these technologies now being applied in some knowledge management solutions are those for electronic meetings, for text-based chat, for collaboration (both synchronous and asynchronous), for amassing judgments about quality, and for so-called expertise location. These technologies are in addition to those for handling documents, such as search and classification, which are already well-established yet are still developing.

Despite these trends, there are still significant shortfalls in the ability of technology to support the use of tacit knowledge—for which face-to-face meetings are still the touchstone of effectiveness. As Ackerman has pointed out, this lack of ability is not just because the designers of the applications do not appreciate how important the human dimension is (although that is true in some cases). We simply do not understand well enough how to accommodate this dimension in computer-supported cooperative work. Many of the factors that mediate effective face-to-

face human-human interactions are not well understood, nor do we have good models for how they might be substituted for or synthesized in human-computer interactions. We can expect gradual progress in this direction, perhaps aided by improvements in the general fidelity with which people's faces, expressions, and gestures are rendered in (for example) high-bandwidth videoconferencing, but there can be no assurance of an immediate breakthrough because of the complexity of the problem and the current shortfall in the basic understanding of its elements.

However, the survey in this paper has highlighted many factors that provide grounds for some optimism when we consider how technology can help in knowledge management. Technology can assist teams, who in today's world may meet only occasionally or even never, to share experiences on line in order to be able to build and share tacit knowledge, and more generally to work effectively together, even if the efficiency is less than in face-to-face meetings. From the perspective of tacit knowledge formation and sharing, the relative informality of text-based chat is probably superior to more structured discussions, which may, however, be effective for sharing explicit knowledge. The importance of limiting access to team members has been highlighted by recent work. The chat archive, and other recordings of on-line meetings, have the added advantage of being able to help in the socialization of people who miss parts of the original interaction. It is also encouraging that recent work by Olson and Olson and their collaborators has shown that studio-quality video is helpful in some tasks related to knowledge management, such as collaboration (in some cases) and trust building.[26]

Another encouraging use of technology is to help persons who need to share knowledge to find each other. Expertise location systems are in their infancy in industrial practice but hold out the promise of being able to identify individuals with the right knowledge. Even without actually identifying a person, unrestricted forums and bulletin boards have been shown to be effective in eliciting assistance both from experts and from the broader community. It seems likely that appropriate integration of this approach with chat on the one hand and expertise location on the other will result in more effective access to and communication of the knowledge in an organization.

Another way to tap the knowledge of experts is through capturing their judgments, expressed as annotation, hyperlinks, citations, and other interactions with documents. Portal infrastructures, which mediate and can collect metrics on the interaction of people and documents, are ideal for amassing this kind of information. Currently, portal products are just becoming capable of accumulating meta-data of this kind. Another trend is for their meta-data to become richer and to support a broader range of tasks. In particular, the meta-data can support the formation of new tacit knowledge from the explicit knowledge indexed by the portal, for example, by situating documents within a new conceptual framework represented by a knowledge map. It is becoming cheaper to use several different frameworks for this purpose, and thus to match them better to the needs of different groups of users, because the accuracy of automatic text classification is improving and, for some classes of content such as news stories, is already as good as the accuracy of human indexers.

Technology will clearly become more helpful in dealing with information overload. Techniques such as summarization can reduce the load of persons attempting to find the right documents to use in some task. There is some promise, as yet unfulfilled, that intelligent agents may in the future help persons to prioritize the messages they receive. And the meta-data stored by portals can be used to draw visualizations of large amounts of information, although, contrary to intuition, graphical visualizations seem not to be better than their text-based equivalents, at least for information retrieval tasks.

Finally, it should be emphasized again that this paper has dealt with human knowledge, not with the formation or use of expert systems or similar knowledge-based systems that aim to replace human reasoning with machine intelligence. The current capability of machine intelligence is such that, for the great majority of business applications, human knowledge will continue to be a valuable resource for the foreseeable future, and technology to help to leverage it will be increasingly valuable and capable.

Cited references

1. T. H. Davenport and L. Prusak, *Working Knowledge: How Organizations Manage What They Know*, Harvard Business School Press, Boston, MA (1998).
2. M. Polanyi, *The Tacit Dimension*, Routledge & Kegan Paul, London (1996).

3. M. Polanyi, "The Tacit Dimension," *Knowledge in Organizations*, L. Prusak, Editor, Butterworth-Heinemann, Woburn, MA (1997).
4. I. Nonaka, "The Knowledge Creating Company," *Harvard Business Review* **69**, 96–104 (November–December 1991).
5. I. Nonaka and H. Takeuchi, *The Knowledge Creating Company*, Oxford University Press, Oxford, UK (1995).
6. I. Nonaka and H. Takeuchi, "A Dynamic Theory of Organizational Knowledge Creation," *Organizational Science* **5**, No. 1, 14–37 (1994).
7. I. Nonaka and N. Konno, "The Concept of 'Ba': Building a Foundation for Knowledge Creation," *California Management Review* **40**, No. 3, 40–54 (1998).
8. M. T. Hansen, N. Nohria, and T. Tierney, "What's Your Strategy for Managing Knowledge?" *Harvard Business Review* **77**, 106–116 (March–April 1999).
9. B. Boguraev and C. Kennedy, "Applications of Term Identification Technology: Domain Description and Content Characterization," *Natural Language Engineering* **1**, 1–28 (1998).
10. A. Maedche and S. Staab, "Mining Ontologies from Text," *Knowledge Acquisition, Modeling and Management* (*EKAW*), Springer, Juan-les-Pins (2000).
11. P. Jackson, K. Al-Kofahi, C. Kreilick, and B. Grom, "Information Extraction from Case Law and Retrieval of Prior Cases by Partial Parsing and Query Generation," *Proceedings of the 1998 ACM CIKM: 7th International Conference on Information and Knowledge Management*, Bethesda, MD (November 3–7, 1998), pp. 60–67.
12. L. Kalwell, Jr., S. Beckhardt, T. Halvorsen, R. Ozzie, and I. Greif, "Replicated Document Management in a Group Communication System," *Proceedings of the Conference on Computer Supported Cooperative Work*, Portland, OR (1988).
13. M. M. Mann, R. L. Rudman, T. A. Jenckes, and B. C. McNurlin, "EPRINET: Leveraging Knowledge in the Electric Utility Industry," *Knowledge in Organizations*, L. Prusak, Editor, Butterworth-Heinemann, Woburn, MA (1997), pp. 73–97.
14. D. A. Foulger, *Medium as Process: The Structure, Use and Practice of Computer Conferencing on IBM's IBMPC Computer Conferencing Facility*, Ph.D. thesis, Department of Communications, Temple University, Philadelphia, PA (1991).
15. R. Ruggles, "The State of the Notion: Knowledge Management in Practice," *California Management Review* **40**, No. 3, 80–89 (1998).
16. M. S. Ackerman, "The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility," *Human-Computer Interaction* **15**, 179–203 (2000).
17. M. Robinson, M. Kovalainen, and E. Auramäki, "Diary as Dialogue in Papermill Process Control," *Communications of the ACM* **43**, No. 1, 65–70 (January 2000).
18. W. Orlikowski, "Improvising Organizational Transformation over Time: A Situated Change Perspective," *Information Systems Research* **7**, No. 1, 63–92 (1996).
19. C. Jackson, *Process to Product: Creating Tools for Knowledge Management*, http://www.brint.com/members/online/120205/jackson/secn1.htm (2001).
20. R. G. Smith and A. Farquhar, "The Road Ahead for Knowledge Management," *AI Magazine* **21**, No. 4, 17–40 (2000).
21. T. Erickson and W. A. Kellogg, "Social Translucence: An Approach to Designing Systems That Support Social Processes," *ACM Transactions on Computer-Human Interactions* **7**, No. 1, 59–83 (2000).
22. E. Bradner, W. A. Kellogg, and T. Erickson, "The Adoption and Use of 'Babble': A Field Study of Chat in the Workplace," *Sixth European Conference on Computer Supported Cooperative Work*, Copenhagen (1999).
23. R. S. Fish, R. E. Kraut, R. W. Root, and R. E. Rice, "Video as a Technology for Informal Communication," *Communications of the ACM* **36**, 48–61 (1993).
24. T. T. Midwinter and P. J. Sheppard, "Ecollaboration—'The Drive for Simplicity'," *BT Technology Journal* **18**, No. 2, 107–115 (2000).
25. W. Geyer, S. Daijavad, T. Frauenhofer, H. Richter, K. Truong, L. Fuchs, and S. Poltrock, "Virtual Meeting Support in TeamSpace," *Demonstration, CSCW'00, ACM Conference on Computer-Supported Cooperative Work*, Philadelphia, PA (2000). Also see http://www.research.ibm.com/teamspace/index.html.
26. N. D. Bos, D. Gergle, J. S. Olson, and G. M. Olson, "Being There Versus Seeing There: Trust Via Video," *Proceedings of CHI 2001*, short papers (2001).
27. L. N. Foner, "A Multi-Agent, Referral-Based Matchmaking System," *Agents '97: First International Conference on Autonomous Agents*, Marina del Rey, CA (1997).
28. R. Copeland, *Mine Your Intellectual Assets*, InformationWeek.com, http://www.informationweek.com/824/lotus.htm (2001).
29. Tacit Knowledge Systems, http://www.tacit.com/products/knowledgemail.html.
30. D. W. McDonald and M. S. Ackerman, "Just Talk to Me: A Field Study of Expertise Location," *ACM Conference on Computer Supported Cooperative Work* (*CSCW '98*), Seattle, WA (1998).
31. J. F. Nunamaker, A. R. Dennis, J. S. Valacich, D. R. Vogel, and J. F. George, "Electronic Meeting Systems to Support Group Work," *Communications of the ACM* **34**, No. 7, 40–61 (1991).
32. W. G. Pope and S. L. Keck, *Diffusing Private Knowledge in Organizations Via Unrestricted Electronic Bulletin Boards*, Pace University Working Paper (2001).
33. J. M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*.
34. S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg, "Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text," *Computer Networks and ISDN Systems* **30**, 65–74 (1998).
35. ResearchIndex is at http://www.csindex.com.
36. S. Lawrence, C. L. Giles, and K. Bollacker, "Digital Libraries and Autonomous Citation Indexing," *Computer* **32**, 67–71 (1999).
37. S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Computer Networks and ISDN Systems* **30**, 107–117 (1998).
38. I. A. Ovsiannikov, M. A. Arbib, and T. H. McNeill, "Annotation Technology," *International Journal of Human-Computer Studies* **50**, 329–362 (1999).
39. Overview of the Sophia Project, at http://www.almaden.ibm.com/st/projects/managementsolutions/sophia/.
40. J. Cadiz, A. Gupta, and J. Grudin, "Using Web Annotations for Asynchronous Collaboration Around Documents," *Proceedings of CSCW*, Philadelphia, PA (2000).
41. R. Lienhart, P. Silvia, and E. Wolfgang, "Video Abstracting," *Communications of the ACM* **40**, 54–62 (1997).
42. L. He, E. Sanocki, A. Gupta, and J. Grudin, "Auto-Summarization of Audio-Video Presentations," *Proceedings of Multimedia '99* (1999).
43. J. Foote, J. Boreczky, and L. Wilcox, "Finding Presentations in Recorded Meetings Using Audio and Video Features,"

IEEE International Conference on Acoustics, Speech, and Signal Processing (1999).
44. S. Srinivasan, D. Ponceleon, A. Amir, and D. Petkovic, "What Is in That Video Anyway?: In Search of Better Browsing," *IEEE International Conference on Multimedia Computing and Systems*, Florence, Italy (1999).
45. W. Niblack, R. Barber, W. Equitz, M. Flickner, E. H. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin, "The QIBC Project: Querying Images by Content, Using Color, Texture, and Shape," *Storage and Retrieval for Image and Video Databases*, *SPIE Proceedings Series*, Vol. 1908, San Jose, CA (February 1993), pp. 173–187.
46. M. Flickner, H. S. Sawhney, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by Image and Video Content," *Computer* **28**, No. 9, 23–32 (1995).
47. M. G. Christel, T. Kanade, M. Mauldin, R. Reddy, M. Sirbu, S. M. Stevens, and H. D. Wactlar, "Informedia Digital Video Library," *Communications of the ACM* **38**, No. 4, 57–58 (1995).
48. J. W. Cooper, M. Viswanathan, and Z. Kazi, "Samsa: A Speech Analysis, Mining and Summary Application for Outbound Telephone Calls," *Proceedings of the 34th Annual Hawaii International Conference on System Sciences HICSS-34* (2001).
49. E. W. Brown, S. Srinivasan, A. Coden, D. Ponceleon, J. W. Cooper, and A. Amir, "Toward Speech as a Knowledge Resource," *IBM Systems Journal* **40**, No. 4, 985–1001 (2001, this issue).
50. R. Mack, Y. Ravin, and R. J. Byrd, "Knowledge Portals and the Emerging Digital Knowledge Workplace," *IBM Systems Journal* **40**, No. 4, 925–955 (2001, this issue).
51. C. M. Bowman, P. B. Danzig, D. R. Hardy, U. Manber, and M. F. Schwartz, "The Harvest Information Discovery and Access System," *Computer Networks and ISDN Systems* **28**, 119–125 (1995).
52. Gnutella is a protocol for information-sharing technology; hub at http://gnutella.wego.com.
53. Groove Networks, http://www.groove.net/.
54. R. Baeza-Yates, B. Ribeiro-Neto, and R. Baeza-Yates, *Modern Information Retrieval*, Addison-Wesley Publishing Co., Reading, MA (1999).
55. D. D. Lewis and K. Sparck Jones, "Natural Language Processing for Information Retrieval," *Communications of the ACM* **39**, No. 1, 92 (1996).
56. *Proceedings of the Eighth Text Retrieval Conference* (*TREC-8*), E. M. Voorhees and D. K. Harman, Editors, at http://trec.nist.gov/pubs.html/ (2000).
57. A. Spink, D. Wolfram, B. J. Jansen, and T. Saracevic, "Searching the Web: The Public and Their Queries," *Journal of the American Society of Information Science* **53**, No. 2, 226–234 (2001).
58. E. M. Voorhees, "On Expanding Query Vectors with Lexically Related Words," *Second Text Retrieval Conference* (*TREC-2*) (1993).
59. *WordNet: An Electronic Lexical Database* (*Language, Speech and Communication*), C. Fellbaum, Editor, MIT Press, Cambridge, MA (1998).
60. R. Mandala, T. Tokunaga, and H. Tanaka, "Combining Multiple Evidence from Different Types of Thesaurus for Query Expansion," *Proceedings of SIGIR '99* (1999).
61. W. A. Woods, L. A. Bookman, A. Houston, R. J. Kuhns, and P. Martin, "Linguistic Knowledge Can Improve Information Retrieval," *Language Technology Joint Conference*, ANLP Sessions, Seattle, WA (April 29–May 4, 2000).
62. Yahoo! is at http://www.yahoo.com.
63. Y. Yang and X. Liu, "A Re-Examination of Text Categorization Methods," *Proceedings of SIGIR '99* (1999).
64. A. Joscelyne, "Automatic Information Refining: A Reuters Success Story," *Language Industry Monitor*, http://www.lim.nl/monitor/reuters.html (May/June 1991).
65. T. Zhang and F. J. Oles, "Text Categorization Based on Regularized Linear Classification Methods," *Information Retrieval* **4,** No. 1, 5–31 (2001).
66. K. Kukich, "Beyond Automated Essay Scoring," *IEEE Intelligent Systems* **15**, No. 5, 22–27 (September–October 2000).
67. L. S. Larkey, "Automatic Essay Grading Using Text Categorization Techniques," *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia (1998).
68. W. Pohs, G. Pinder, C. Dougherty, and M. White, "The Lotus Knowledge Discovery System: Tools and Experiences," *IBM Systems Journal* **40**, No. 4, 956–966 (2001, this issue).
69. R. Kubota Ando, "Latent Semantic-Space: Iterative Scaling Improves Precision of Inter-Document Similarity Measurement," *Proceedings of SIGIR '00* (2000).
70. R. K. Ando, B. K. Boguraev, R. J. Byrd, and M. S. Neff, "Multi-Document Summarization by Visualizing Topical Content," *Proceedings of ANLP/NAACL 2000 Workshop on Automatic Summarization* (2000).
71. A. Tombros and M. Sanderson, "Advantages of Query Biased Summaries in Information Retrieval," *Proceedings of SIGIR '98* (1998).
72. H. P. Luhn, "The Automatic Creation of Literature Abstracts," *IBM Journal of Research and Development* **2**, No. 2, 159–165 (1958).
73. J. Robin and K. R. McKeown, "Corpus Analysis for Revision-Based Generation of Complex Sentences," *Proceedings of the National Conference on Artificial Intelligence*, Washington, DC (1993).
74. B. K. Boguraev and M. S. Neff, "Discourse Segmentation in Aid of Document Summarization," *Proceedings of the 33rd Hawaii International Conference on System Sciences*, Maui, HI (2000).
75. B. Boguraev, R. Bellamy, and C. Swart, "Summarization Miniaturization: Delivery of News to Hand-Helds," *Proceedings of Workshop on Automatic Summarization, Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA (2001).
76. D. R. Radev and K. R. McKeown, "Generating Natural Language Summaries from Multiple On-Line Sources," *Computational Linguistics* **24**, 469–500 (1998).
77. I. Mani, D. House, G. Klein, L. Hirshman, L. Obrst, T. Firmin, M. Chrzanowski, and B. Sundheim, *The TIPSTER Summac Text Summarization Evaluation*, Technical Report, Mitre Corporation, McLean, VA (1998).
78. *Advances in Automatic Text Summarization*, I. Mani and M. Maybury, Editors, MIT Press, Cambridge, MA (1999).
79. *Proceedings of ANLP/NAACL 2000 and 2001 Workshops on Automatic Summarization*.
80. *Advanced Learning Technology: Design and Development Issues*, J. C. Kinshuk and T. Okamoto, Editors, IEEE Computer Society, Los Alamitos, CA (2000).
81. D. Shenk, *Data Smog: Surviving the Information Glut*, Harper, San Francisco (1998).
82. TREC Topic Detection and Tracking, see http://www.nist.gov/speech/tests/tdt/index.htm.
83. J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow, "Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Doc-

uments," *Proceedings of IEEE Information Visualization '95*, Atlanta, GA (1995).
84. VisualNet is described at http://www.map.net.
85. M. Hearst and C. Karadi, "Cat-a-Cone: An Interactive Interface for Specifying Searches and Viewing Retrieval Results Using a Large Category Hierarchy," *Proceedings of SIGIR '97*, Philadelphia, PA (1997).
86. M. A. Hearst, "TileBars: Visualization of Term Distribution Information in Full Text Information Access," *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems*, Denver, CO (May 1995), pp. 59–66.
87. A. Veerasamy and N. J. Belkin, "Evaluation of a Tool for Visualization of Information Retrieval Results," *ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich (1996).
88. M. M. Sebrechts, J. Cugini, S. J. Laskowski, J. Vasilakis, and M. S. Miller, "Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces," *SIGIR '99 22nd International Conference on Research and Development in Information Retrieval*, Berkeley, CA (1999).
89. J. W. Cooper and R. J. Byrd, "Lexical Navigation: Visually Prompted Query Expansion and Refinement," *Proceedings of Digital Libraries '97*, Philadelphia, PA (1997).
90. I. Ben-Shaul, M. Herscovici, M. Jacovi, Y. S. Maarek, D. Pelleg, M. Shtalhaim, V. Soroka, and S. Ur, "Adding Support for Dynamic and Focused Search with Fetuccino," *Proceedings of WWW8*, Toronto (1999).

Accepted for publication June 15, 2001.

Alan D. Marwick *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 704, Yorktown Heights, New York 10598 (electronic mail: marwick@us.ibm.com).* Dr. Marwick received B.Sc. and D.Phil. degrees in physics from the University of Sussex in Britain, then worked on the application of nuclear methods of analysis to research problems in materials science and the effect of radiation on solids, first at the AEA Harwell Laboratory in Britain and then at the Watson Research Center. More recently he has led groups working on on-line access to the scientific literature, digital libraries, information retrieval, natural language processing, and knowledge management. In addition to his technical interests, he works on the practical problems of applied research and technology transfer in an industrial environment. Dr. Marwick currently manages the Knowledge Management Technology Department at the Research Center.