Information management in research collaboration

H. CHEN, K. J. LYNCH, A. K. HIMLER AND S. E. GOODMAN

MIS Department, University of Arizona, Tucson, Arizona 85721, USA

(Received 15 April 1990 and accepted in revised form 12 March 1991)

Much of the work in business and academia is performed by groups of people. While significant advancement has been achieved in enhancing individual productivity by making use of information technology, little has been done to improve group productivity. Prior research suggests that we should know more about individual differences among group members as they respond to technology if we are to develop useful systems that can support group activities.

We report results of a cognitive study in which researchers were observed performing three complex information entry and indexing tasks using an Integrated Collaborative Research System. The observations have revealed a taxonomy of knowledge and cognitive processes involved in the indexing and management of information in a research collaboration environment. A detailed comparison of knowledge elements and cognitive processes exhibited by senior researchers and junior researchers has been made in this article. Based on our empirical findings, we have developed a framework to explain the information management process during research collaboration. Directions for improving design of Integrated Collaborative Research Systems are also suggested.

1. Introduction

Much of the work in business and academia is performed through collaboration by members of groups of people (Bair, 1985). Managers and intellectual workers, in particular, spend a large portion of their time working in groups. Information systems (IS) managers, for example, spend between 60 and 70% of their time in group activities (Hymowitz, 1988; Ives & Olson, 1981; Mintzberg, 1983; Mosvick & Nelson, 1986). While significant advancement has been achieved in enhancing individual productivity by making use of information technology, little has been done to improve group productivity. In the past decade, IS researchers recognized this problem and started to develop technology that can support the group collaboration process. This type of system has been referred to as "groupware". It consists of software, hardware and communications designed to support teams of people working on a particular problem. Examples of this type of system are electronic mail, videoconferencing, integrated collaborative research systems and group decision support systems. Among these types of groupware, integrated collaborative research systems (ICRS) are most relevant to intellectual workers in both academia and business. ICRS typically support the activities related to a research cycle: assembling information from a wide variety of sources, organizing and manipulating that information, analysing the information and writing papers. The projects involved may cover extended time periods, and employ multiple researchers who collaborate on a series of related topics. Systematic and relentless information collection, indexing and management is crucial to the success of research collaboration.

While information resources are becoming more highly computerized, access to these resources is often difficult because of the indeterminism involved in the process by which information is collected and indexed and the latitude users have in choosing terms to express a query. In a complex, computerized task environment, the problem of storing and retrieving documents can be difficult for three reasons:

1. it can require a significant amount of knowledge of the subject area in which information is sought;
2. it requires knowledge about the way the information system functions; and
3. it requires knowledge about the classification scheme of the information system.

These problems are even more evident in an environment that involves multiple users and unstructured information. Grudin (1988) identified several problems in the design and evaluation of computer-supported cooperative work (CSCW) applications and proposed the following research direction:

“We need to have a better understanding of how groups and organizations function and evolve than is reflected in most of the (CSCW) systems that have been developed. At the same time, we also need to know more about individual differences in responding to technology if we are to develop systems that can support entire groups.”

Our research attempted to address these questions by studying in detail how members of a research group used an ICRS. The system was designed to support collaborative research on computer-related technology transfer and international technological trends and policy. The research group we studied conducts foreign-area studies and technology assessment of international developments in the information technologies, focusing on the former-Communist countries (Goodman *et al.*, 1990). The system allows users (researchers for the most part) to enter, index and organize information (we refer to these processes as “information management” in the context of this article) from various sources and provides assistance to users in information retrieval and analysis.

The goal of this research was to understand how users of a sophisticated ICRS used the system to organize and index their information. In particular, we attempted to investigate the types of knowledge elements and cognitive processes that were essential in performing these tasks. We report the results of a cognitive psychology-based study in which six subjects of varying degrees of expertise performed three tasks of different levels of difficulty. Among the subjects, three were considered experts and the rest were novices. The subjects were asked to think aloud during the experiment. Their interactions with the system were also logged. An interview was conducted after completion of each task. Each task lasted between 20 min and 2 h. Protocol Analysis and Problem Behavior Graph techniques were used in this research to analyse the qualitative data.

The specific results we report in this paper are:

1. a taxonomy that categorizes knowledge elements for information management and a comparison of characteristics of experts’ and novices’ knowledge (in Section 5.1);

2. a detailed process model of the information entry and indexing procedure and the specific characteristics of the experts and the novices observed during the process (in Section 5.2);
3. a framework depicting the relationships between the knowledge elements and the information entry and indexing process (in Section 5.3); and
4. the implications of our findings for the understanding of research collaboration and ICRS design (in Section 6).

We present an overview of research collaboration and information management in Sections 2 and 3, respectively. The research design is described in Section 4 while Sections 5 and 6 discuss our research findings. We conclude this paper in Section 7.

2. Research collaboration

Human collaboration has been a subject of study for researchers in the areas of sociology, philosophy and anthropology. In the past decade, information technologies have emerged that support the group collaboration process. Although the technologies are quite diverse, the problem being examined is fundamental to business managers and intellectual workers.

Information technology-based systems for the support of group work can be classified into two broad categories: group decision support systems (GDSS) and computer-based systems for cooperative work (CSCW) (Dennis *et al.*, 1988). DeSanctis and Gallupe (1985) defined GDSS as "an interactive computer-based system which facilitates solution of unstructured problems by a set of decision makers working together as a group". Johnson *et al.* (1986) define CSCW as, "the use of computer and electronic communication tools as media for communicating". A central concern in CSCW is coordinated access to shared information (Greif, 1986; Gorry *et al.*, 1988). As Dennis *et al.* (1988) indicated, the distinction between CSCW and GDSS may be blurred due to rapid advancements in groupware development.

An important sub-area within CSCW is termed "research collaboration". This sub-area is concerned with how intellectual workers (researchers) collaborate in order to generate joint research. Much work in academia and business is accomplished by groups of people (Bair, 1985). Examinations of patterns of authorship reveal that collaborative research is the norm in most research-oriented environments. But, so far, improvements in information technology do not seem to facilitate this process (Over, 1982; Kraut, Galegher & Egidio, 1986).

Prior research has shown that research collaboration involves the collection of information by individual members and the reciprocal exchange of information. Researchers share information by some means and thus effect changes in the thinking and actions of the people involved in the collaboration process (Goodman & Abel, 1986; Gorry *et al.*, 1988; Lynch, 1989). Research collaboration involves a cycle with a number of activities: assembling information from a wide variety of sources, organizing and managing that information, retrieving relevant information for analysis, sharing ideas between members in the group and document preparation. Long-term, related research efforts by groups of researchers require high-level information modeling and management capabilities and the ability to build on past

analytical achievements by allowing models to evolve as understanding is enhanced. Integrated Collaborative Research Systems (ICRS) support projects covering an extended time period, employing multiple researchers who collaborate on a series of related topics, and base their work on information that is primarily composed of non-numeric data (Lynch *et al.*, 1990).

There are some systems that support part of the activities in the research cycle. CES is a Collaborative Editing System for a group of co-authors working asynchronously on a shared document (Greif, 1986). "Quilt" is also a computer-based tool for collaborative document preparation. It provides annotation, messaging, computer conferencing and notification facilities to support communication and information sharing among the collaborators on a document (Leland, Fish & Kraut, 1988).

Few systems support the full research collaboration cycle. "Visual Notebook" was designed to help the members of a biomedical group coordinate efforts and share information and to improve group functioning through the automated importation and storage of relevant information from external sources (Gorry *et al.*, 1988). The system helps biomedical researchers scan and filter information from the environment. It helps researchers perform information management, decision making and document preparation using the information collected. "The Analyst", developed by Xerox (1989), is another product in this category; it includes a desk-top publishing module along with extensive, group-oriented hypermedia capabilities. A full-scale ICRS in use by the Mosaic Group at the University of Arizona is the Arizona Analyst Information System (AAIS). This was designed to support collaborative research on computer-related technology transfer and international technological trend and policy analysis (Lynch, 1989). Support is provided for collaboration between researchers across the entire research cycle by facilitating entry, indexing, retrieval, analysis and writing of large volumes of textual information. The AAIS provides a means of information modeling and management and inter-researcher communication in a dynamic, multi-task, collaborative environment (McHenry, Lynch & Goodman, 1988). AAIS was the system used for our empirical study. More details will be supplied in the following sections.

3. Foundation of research collaboration: information management

In this section, we first summarize the problems associated with information management. Bates (1986) proposed a framework based on general system theory for understanding the major problems associated with on-line information management: information indexing and retrieval. The framework highlights the role of uncertainty and indeterminism in both the process of indexing and the process of searching. According to this framework, the problems associated with information processing can be understood more clearly by considering the indeterminism involved in the process of information classification and in the terms and strategies searchers use in formulating a query (Chen & Dhar, 1987). Specifically, the factors that make indexing and retrieval problematic are:

1. *Uncertainty*: the indexing process is partly indeterminate and probabilistic. Evidence suggests that different indexers, well-trained in an indexing scheme,

might assign indexes for a given document differently. It has also been observed that an indexer might use different terms for the same document at different times (Jacoby & Slamecka, 1962; Stevens, 1965). An even higher degree of uncertainty has been observed in users' searches. One study revealed that, on average, the probability of any two people using the same term to describe an object ranged from 7 to 18% (Furnas, 1982);

2. *Redundancy*: due to the uncertainty in indexing and searching, generating an exact match between the user's term and that of the indexer becomes difficult. Bates (1986) argues that for a successful match, the searcher must somehow generate as much "variety" (in the cybernetic sense, as defined by Ashby (1973)) in the search as is produced by the indexers in their indexing. The variety produced by an indexer can also be viewed as *redundancy* in the sense that it consists of partially overlapping meanings applied to a document. In practice, however, information systems may discourage redundancy and enforce *consistency*. The following principles have been used in various indexing systems (Bates, 1986; Chan, 1986):
 - (i) *Controlled keywords*: the principle of controlled keywords holds that any document should be indexed using the keywords provided by the system. This principle, however, may place a burden upon the indexers. For untrained searchers, the problem of matching the search terms with the indexes still exists;
 - (ii) *Whole document indexing*: an indexer may be trained to index the whole document, not parts or concepts within it. Each document is to be entered under a keyword that is specific to the content of the document, neither broader nor narrower in scope than the scope of the document's contents. In an on-line environment, this principle reduces "access points" to the document. Information retrieval may become more difficult;
 - (iii) *Cross-reference structure*: Linkages are created between keywords and thus augment the semantic structure of the keywords. This structure is evident in most thesauri such as the Library of Congress Subject Headings (LCSH). Cross-reference structure can assist information retrieval and enhance query articulation.

In summary, the indeterminism involved in the information management process decreases the likelihood that a user will retrieve relevant information.

4. Study design and method

The Arizona Analyst Information System (AAIS) has played a major role in the work of the Mosaic researcher group at the University of Arizona (Goodman *et al.*, 1990). The group conducts foreign-area studies and assessment of information technologies, focusing on the Soviet Union, Eastern Europe and the People's Republic of China. Today, the group comprises more than 20 members. The current group boasts fluency in seven languages, proficiency in five others and its members have a working capability in four more. The AAIS databases have over 40,000 entries from 20,000 sources and have served over 50 researchers. AAIS databases have supported three dissertations, 54 refereed papers, 19 chapters in books, 63 trip

reports and 35 unpublished manuscripts and technical reports. AAIS databases are also used to support research ranging from *ad hoc*, time-critical questions to long-term policy analyses for the US Congress. A major part of the group collaboration effort is the collection, indexing and management of time-critical, text-based information.

Six of these researchers (or “analysts”, as they are referred to in the group) participated in the study. For the purpose of this study, we classified analysts as “novice” or “expert”, based on their experiences. Experts are analysts who have worked with the system for four years or more and possess a focused area of expertise within the subject area. Novices are analysts who have worked with the system for less than three years and exhibit less subject knowledge. Among the six subjects, three were considered experts and the remaining three subjects were novices. These subjects were asked to perform three tasks (of their own choosing) in document entry and indexing. The documents involved ranged from an article in a foreign language journal or a chapter in a book to somebody’s business card, a brochure or a group member’s trip reports. These tasks were classified by the analysts as easy, medium or hard. The criteria used by the analysts were: the number of topics involved, the amount of time expected to be required and the technical difficulty and structure of the documents. Since the difficulty levels of the tasks strongly depended on the expertise of the subjects, no objective measure could be used. Instead, the analysts’ perceived levels of difficulties (subjective difficulty) were adopted. But overall, experts handled materials that were more difficult than those handled by novices (i.e. the experts’ materials were more complex and technical, the source documents were more unstructured and the amount of time required was longer).

Subjects were asked to think aloud during their interaction with the system. These protocols were tape-recorded and the interaction between the analysts and the system was logged. Most of the interactions lasted between 20 min and 2 h. After each interaction, the analysts were interviewed about the indexing process and any problems encountered.

The technique we used to analyse the collected data was *protocol analysis*, which involves detailed analysis of the think-aloud protocols and logs. Our study was qualitative in nature. Our focus was on the detailed investigation of the cognitive skills exhibited among the group members and how these skills affected their behaviors. Protocol analysis technique is typically very time-consuming, but we believe the findings can help us understand the group collaboration process at a much more detailed level. Our study can also help us identify the design criteria for more useful research collaboration systems.

Protocol analysis requires categorizing the protocols and logs according to the processes that could have generated them (Ericsson & Simon, 1984). Various information processing concepts and tools, such as *problem space* and *Problem Behavior Graph*, are important for identifying these processes.

In order to perform a task, a person needs to create an internal representation of the task. This internal representation is referred to as the *problem space*. A *problem space* can be envisioned as a set of nodes that indicate the knowledge states involved in the problem solving process. A successful problem solving session can be perceived as a traversal from an *initial state*, using various operators to arrive at

some intermediate states, to a *final goal state*, a knowledge state that contains the solution to the problem. The approach a problem solver adopts in traversing the problem space is referred to as the *process model* of problem solving. A problem solving process is in essence the formulation of the problem space elements in response to demands of the task environment (Newell, 1980; Newell & Simon, 1972).

Different problem solvers, varying in their expertise in the task, may possess different knowledge states, adopt different operators, and display different process models. A large body of research in the past decade has reported the differences between the ways novices and experts solve real world problems. The research findings help identify reasons for the better performance of experts. Examples of the domains that have been studied include: program debugging (Gugerty & Olson, 1986), mathematical programming (Orlikowski & Dhar, 1986), physics problems (Larkin, 1983; Chi, Feltovich & Glaser, 1981) and conceptual data modeling (Batra & Davis, 1989).

A number of studies have used the *Problem Behavior Graph* (PBG) as a means to study the problem solving process. This representation describes problem-solving activities in a time sequence, from an initial state (a vague description of needs) to a goal state (a solution that satisfies the needs), using nodes and labeled links. This detailed representation of the problem solving process is derived by first splitting the user's interaction logs and verbal protocols into their *semantic elements*, which consist of the *knowledge elements* and the *operator elements* (Newell & Simon, 1972; Waterman & Newell, 1971). The *knowledge elements* specify the kinds of knowledge the subject has about the task. The *operator elements* are a finite set of actions that take a state of knowledge as input and produce a new state of knowledge as output. Users typically employ a finite number of operators to change the knowledge states.

Bouwman (1983) compared the decision-making processes of experts and novices in financial analysis tasks. The verbal protocols obtained in this study were used to construct a PBG to indicate the problem solver's process model of financial analysis. Bouwman's study revealed that the financial analysis task can be broken down into three stages: examination of existing information; integration of observations and findings; and reasoning. These findings were used to construct computer programs that simulate financial experts' behaviors. Huguenard, Prietula and Lerch (1989) presented a study in which the fragility of expertise in *reactive scheduling* was investigated. PBG was used to identify the various strategies taken by the experts and novices, and these researchers observed that simple modifications to the task environment were sufficient to degrade the performance of the experts, sometimes to the level of the novices. Chen and Dhar (1990a) studied the on-line information retrieval behavior of novice and expert searchers in a library setting and identified various search strategies (approaches or processes) that were adopted by the subjects. The expert searchers employed three search strategies that were more effective and efficient and were considered *strong* methods of problem solving. The novice searchers, on the other hand, adopted two *weak* search strategies. The *strong* search strategies were simulated and incorporated into the design of a knowledge-based information retrieval system (Chen & Dhar, 1990b).

One conclusion that can be drawn from research investigating problem-solving activity is that different problem solvers may possess different knowledge elements,

which in turn affect the formulation of the problem, the selection of problem-solving methods and eventually the performance of the task. Prior research has shown that experts often “chunk” a large amount of information, adopt “strong” problem-solving methods and perceive problems at a higher level of abstraction. In our current study, we also compare the differences between experts and novices in their knowledge and problem-solving methods when performing information management-related tasks.

5. Information management in research collaboration: empirical findings

In this section we present the findings from our empirical study. We first describe the analysts’ knowledge elements that were essential for the information management tasks. The knowledge elements are presented as a taxonomy. In the subsequent sub-section, we then describe the process involved in information management. We make explicit comparison between the experts’ and novices’ knowledge elements and processes. In the final subsection, we develop a framework for information management in research collaboration, based on our cognitive findings.

5.1. KNOWLEDGE ELEMENTS

A qualitative analysis of the log files and protocols revealed a range of knowledge elements that were used by the analysts when indexing their documents. These knowledge elements are organized around four major categories. The first concerns knowledge about the classification scheme that the system uses to organize the documents. The second category is about knowledge in the subject area of a document. The third type of knowledge is related to the functioning of the system. These three categories of knowledge have been reported in research by Chen and Dhar (1990c) that investigated the searcher’s information retrieval process using an on-line catalog system. The last category, which we call collaboration knowledge, appears to be unique to the task of collaborative information management.

5.1.1. Classification scheme knowledge

Knowledge about the classification scheme the system employs can be divided into two types: knowledge about the file assignment process and knowledge about the controlled vocabulary used.

1. Knowledge about file assignment

The system includes a subject directory tree. This tree represents topics that are of interest to the group at different levels of specificity. At the bottom level, there are “files” (or “folders”) that cluster texts having similar contents. New documents need to be assigned to the proper files. (A document can be assigned to multiple files.) Knowledge for file assignment has been identified as being of two types:

- (i) *Knowledge of file contents.* Due to the complexity of the Mosaic directory tree, which encompasses many levels of sub-directories and files, analysts are often faced with the task of making file assignments to an area of the
-

directory tree. Many analysts spent time browsing the documents in the files in an attempt to identify the correct files to which to assign the new document. As one analyst remarked: "... the directory tree has to be known, inside and out" (before proper file assignment can be achieved).

- (ii) *Knowledge about the file assignment principle.* Along with understanding the contents of a particular file, analysts had to recognize the principle used for file assignment. Documents should be assigned to files that are specific to the content of the document (the "whole document indexing principle" described earlier). Some novice analysts assigned their documents to files that were broader in content than those of the actual documents.

2. Knowledge of controlled vocabulary

The system employs a controlled vocabulary for entering certain token fields for a document, such as reference identification (REFID), organization identification (ORGID) and journal identification (JID). Two types of knowledge were noted:

- (i) *Ability to apply correct REFID, JID or ORGID.* An analyst often needed to ascertain the full name of an organization, journal or author from the information provided in the text. Confusion sometimes arose as to which piece of information belonged to which token. For example, when a text came to the analyst from a news or translating service, there was confusion, especially among novice analysts, about whether the primary source (in the original language) or the secondary source (in English) should be placed in the journal identification field.
- (ii) *Ability to recognize the correct acronym.* When searching for the correct REFID, JID or ORGID for the text, analysts may encounter incomplete information associated with these identifiers. Experts were better at disambiguating these incomplete identifiers than novices.

5.1.2. Subject area knowledge

The second category of knowledge is domain-related. Some analysts work almost exclusively in one subject domain, while others deal with material concerning many topics. Two types of subject area knowledge were identified in this study: keyword assignment and interpretation of text.

1. Keywording knowledge

Analysts assigned keywords to represent the contents of the documents. Two types of knowledge were essential for keyword assignment.

- (i) *Choosing terms as keywords.* The database adopts an uncontrolled vocabulary for keywords. Analysts had to choose semantically correct terms to represent the contents of a text. Analysts needed to have a proper grasp of the essential ideas of a text. This may require extensive subject area training.
 - (ii) *Knowledge about the keywording principle.* Analysts had their own principles for assigning keywords. For example, one analyst remarked that: "software is a 'dead' keyword. It is too broad to be useful". Some
-

principles were also suggested in the training manual, such as keywording concepts in their abbreviated form (e.g. "IC" for integrated circuits).

2. Interpretation of text

Information in the text presents a number of challenges to the analyst; some of which relate to content and some the expertise of the analyst. Two types of text-related knowledge were identified.

- (i) *Text translation.* Most expert analysts were able to translate the source documents, enter the translated version and index the documents while using the system. This process often required extensive foreign language understanding. The process of reading and translating the documents has been an important part of the Mosaic research cycle. Analyst's knowledge in the subject area increases during the process of "digesting" the information. This information is stored in the analyst's long-term memory (within the brain). The system, however, provided an external memory (the database) for the analysts by storing all the details.
- (ii) *Synthesizing a text.* Some information comes in the form of long articles, which need to be filtered and synthesized into abstracts of acceptable length in order to be entered as a text. This process requires specific subject area knowledge to identify key concepts in the text.

5.1.3. System knowledge

Detailed knowledge about the functionality of the system was necessary for successful information management. We identified two types of system-related knowledge.

1. *Selecting appropriate search options.* Detailed knowledge about the applicability of the various search options was necessary for conducting an efficient and fruitful text entry process. Expert analysts, in particular, used the system's various search options in an efficient way. For example, expert analysts sometimes used the system's underlying QUEL, query language (the underlying file structure of AAIS is based on the INGRES DBMS), to search for documents.
2. *Knowledge about assigning token values.* Analysts were often aware of the importance of filling in all token values in a document template. Assigning complete and appropriate token values ensured the correctness of information stored.

5.1.4. Collaboration knowledge

Knowledge about other researchers' expertise and current research activities is essential to information management in research collaboration. We identified two types of collaboration knowledge in our study.

1. Knowledge about HOT file assignments

Some analysts had problems sending texts to special files of the database known as "HOT" files. These files contain important or timely information on topics that are the subject of current research, by one or a group of analysts. HOT file assignment is particularly important for research projects that

examine current, time-critical issues. Considering the rapidly changing nature of international politics (especially in the eastern-bloc countries recently) and the advancement in computer technologies world-wide, it is crucial for the researchers to collectively filter information and send it to the appropriate HOT files for further analysis and research. HOT files are also used as "mailboxes" for analyst-specific communication.

2. Knowledge about other researchers' expertise

Often analysts may encounter information which is not within their areas of expertise. They can either send the text to the appropriate analyst or enter the text themselves. Each analyst may act as a dispatcher, routing information to the appropriate expert in the area. Knowledge about other analysts' subject area expertise was crucial in this process. Many junior analysts also adopted the "follow-the-leader" method, in which they browsed the texts that were entered by other analysts and used similar index terms for their texts. They often relied strongly on texts entered by analysts who were considered experts in the areas.

Collaboration knowledge appears to be unique and crucial to the success of information management in a changing research environment. Only through effective information sharing and communication can the group function as a unit in response to its external environment.

In Figure 1 we present a taxonomy of information management-related knowledge.

5.1.5. Novice/expert comparison

As mentioned before, each analyst was observed during three document entry and indexing session, drawing from text material falling into three classes of difficulty: easy, medium and difficult. While this distinction cannot represent a unilateral

- A TAXONOMY OF KNOWLEDGE:
- I. Classification Scheme Knowledge:
 - A. Knowledge about file assignment
 - A.1. Knowledge about the contents of files
 - A.2. Knowledge for file assignments
 - B. Understanding the controlled vocabulary
 - B.1. Ability to supply correct REFID, JID, or ORGID
 - B.2. Ability to recognize correct REFID, etc.
 - II. Subject Area Knowledge:
 - A. Keywording knowledge
 - A.1. Choosing uncontrolled terms
 - A.2. Keywording principles/heuristics
 - B. Interpreting text
 - B.1. Text translation
 - B.2. Synthesizing/filtering text
 - III. System Knowledge:
 - A. Selecting appropriate search options
 - B. Assigning token values
 - IV. Collaboration Knowledge
 - A. Knowledge about HOT file assignments
 - B. Knowledge about other researchers' expertise

FIGURE 1. A taxonomy of indexing knowledge.

classification among all analysts, it offered a fair method of evaluating differences between groups.

Even though all the analysts had been trained to use the system and had exhibited essential knowledge for their work when they were recruited to the group, knowledge still needed to be accumulated through years of experience working in a specific domain, using a specific classification scheme, interacting with a particular system and collaborating with other analysts. Our study showed differences in the users' subject area knowledge, classification scheme knowledge, system knowledge and collaboration knowledge.

In general, experts exploited the system's capabilities much better than the novices. They used the system's functionality in an efficient way. They often spent less time for document entry and indexing than the novices. The AAIS provides a flexible research collaboration environment for users. Different users can select tools they felt most comfortable with. Although the use of these tools may be sub-optimal (compared with what an expert would do), it has provided a system that users can operate and learn while using it. This "tool-box" approach was very important for a research collaboration environment in which the group members had different training and experience.

Subject area knowledge was one area in which the experts demonstrated their superiority. It usually took about three to four years' constant exposure to the material collected by the group to warrant the analyst's being viewed as having subject area competency (considering the difficulty and scarcity of the type of information collected—information technology in the Communist countries). As we discussed earlier, the text reading, entry and indexing process was extremely important to the intellectual growth of the analysts, especially in subject area knowledge. Analysts accumulated their expertise in a subject area by reading previously entered documents and the documents they collected and stored. Productive analysts may enter more than 2000 documents in a year. (The amount of information an expert filtered through was usually three to four times that number.)

Expert's classification scheme knowledge assisted them in choosing better indexes for the documents than the novices. Their indexes were typically more complete and accurate. Once the documents had been entered, the only way to retrieve them was to use appropriate keywords, files and other controlled vocabularies (e.g. organizational identifiers, journal identifiers etc.). The indexes were as important, if not more so, than the text itself. Compared with subject area and system knowledge, classification scheme knowledge appeared to be the one type of knowledge that requires less training and experience to acquire.

Experts tended to have a better understanding of other analysts' expertise and ongoing research areas. They often had no difficulty "routing" the appropriate information to the right analysts at the most opportune time via the "HOT file" mechanism. They performed HOT file assignment more frequently than novices. This ensured the timeliness of access to information on current research activities. Novices, however, were often unaware of current research activities and they sometimes expressed hesitation in sending texts to other expert analysts. Collaboration knowledge was crucial for inter-researcher communication and cooperative information management activities.

Lack of system knowledge, subject area knowledge, classification scheme knowl-

edge and collaboration knowledge may cause inefficient actions or lead to incorrect results. The latter is the most damaging in a research collaboration environment because inexact synthesis of source texts, inappropriate keyword assignment or incorrect file assignment may often cause the documents to be lost or misrepresented. A "polluted" knowledge base may be more damaging to group collaboration than an incomplete one. The knowledge exhibited by the analysts affected their process of performing information indexing and management. In the next sub-section, we present this process, based on our qualitative analysis of their behaviors. We used PBG as a tool for data analysis. The relationships between the knowledge elements and the process also will be examined in detail.

5.2. PROCESS OF INFORMATION ENTRY AND INDEXING

We constructed one PBG for each task to indicate the states involved and the actions taken during the task. Then we categorized each PBG according to the stages involved. Five distinct stages were identified and this global process was the same for both novices and experts. However, during each stage, their behaviors were quite different. In this section, we first present this global process along with an example and we then compare the novices' and the experts' behaviors at each of these stages in detail.

The data entry task on the Mosaic database is facilitated by the AAIS tool for data entry (which is referred to as the ADD procedure). The ADD procedure makes use of different frame-like templates for various information sources (e.g. article template, book template etc.). Each template contains fields (or attributes) to be filled by the analysts. An article template example is shown in Figure 2. Some fields are based on the controlled vocabulary indexing principle. Analysts need to search for existing identifiers before they create their own. AAIS provides various search facilities that include high-level menu-driven search functions (e.g. VIEW, the primary AAIS query mechanism) and the DBMS-specific search options such as Query-By-Form (QBF) or QUEL. Analysts used these tools extensively when entering text.

5.2.1. Global process

The document entry and indexing procedure in AAIS can be modeled as a process of five stages: *template entry and checking*, *text entry*, *keyword assignment*, *file assignment* and *text parsing*. Each stage lasted from 3 min to 1 h.

1. *Template entry and checking*: As described above, the system templates contain fields that must be completed. These fields provide search keys for future retrieval of text items. These fields are: REFID (reference identifier, a unique code assigned to each text entry), ORGID (organization identifier, an acronym used by the database to identify organizations that appear in the text), JID (journal identifier, a standard name for each periodical source in the database) and a series of bibliographic fields, such as author, editor, publisher, date of publication, number of pages etc. Analysts complete these fields with information from the text item. This is the first stage of the document indexing process. The template checking requirement helps maintain indexing consistency among analysts.

```

ARTICLE TEMPLATE *****
&REFID: Gelb86
&JRNID: EKO
&ENGTIT: Software Begs Protection
&NONENGTIT: Programnoye obespecheniye prosit okhrany
&NUMBER: 3
&DATE_PUBLISHED: 1986
&ART_PAGES: 159-165
&AUTHORS: A B Gel'b
&PAGE_OF_REF: 159
&TEXT:
|A B Gel'b| is with the #IKANESSR#. [an:: This article is from one of the
EKO publications. From the hard copy, we are not able to tell which city
it is from. However, since we have many more citations from EKO
(Novosibirsk) than from the Kiev and Minsk versions, and because the
IKANESSR has close ties to the Computer Center in Novosibirsk, EKO
(Novosibirsk) seems a good guess.] The questions of legal treatment of
software was first raised in a 1966 publication. Unfortunately, there has
been only presentation of one view or another, without direct discussion
and certainly without widespread discussion by specialists in ^law^,
programming, etc. At present, the legal status of software is specified in
two documents:: the Interpretation of #Goskomizobreteniy# from Nov. 13,
1975, "On the recognition as ^invention^s computer technology objects
characterized by systems software" and "Decree on the state fund of
algorithms and programs (^GosFAP^) with the corresponding instructions for
the formulation, ^registration^ and software expertise."
&COUNTRY_CODES: UR
&IDNUM: FF/UR/D/SOFTWARE/LEGAL_ISSUES
&ENTRY_PERSON: pw
&SEND_TO_FILES: softlaw.dat; softdstrb.gen

```

FIGURE 2. An article template.

2. *Text entry*: After the template fields have been completed, text can be entered. These data capture the semantics of the document. Analysts often abstract information from a long text source or translate and synthesize from a foreign language text. For simple or short documents, analysts may choose to enter the text verbatim.
3. *Keyword assignment*: Content-bearing terms that appear in the text are often marked by analysts as keywords. These terms become important search keys for future information retrieval. Analysts may employ a variety of heuristics in assigning keywords. This process often requires a good deal of subject area knowledge. Good keyword assignment practice can capture the essence of a document and ease the process of information retrieval in a collaborative environment. Bad keyword assignment (such as assigning overly broad keywords to a document) may cause future retrieval difficulty.
4. *File assignment*: The strength of the AAIS database is based on the clarity and organization of its domain-specific directory tree structure. This directory tree represents topics that are of interest to the group at different levels of specificity. At the bottom level, there are "files" (or "folders") that cluster texts having similar contents. New documents need to be assigned to the proper files. (A document can be assigned to multiple files.) Within Mosaic, the directory tree plays a major role in helping analysts assign text to the

proper files. Proper file assignment is critical to the success of future retrieval of a text. Analysts can use both top-down and bottom-up approaches to find the proper files for each text, as we discuss in detail in the coming sub-sections.

5. *Text parsing*: After document entry and indexing are completed, the system parses the text entries and performs both syntactic and semantic checking. It matches the text with the controlled vocabulary used by the database. If an error occurs (e.g. an unrecognized entry), a menu is presented to the analyst to correct the information. After editing, the analysts can then re-submit the text to the parser. If no errors are found, the entry is written to the database.

We present an actual document entry and indexing example using a PBG. It illustrates the five-stage process model we have just described. Two conventions are used in this graph: nodes and labeled links. Nodes represent knowledge elements that analysts used when performing document entry and indexing. Labeled links represent the actions performed by the analysts. This PBG (Figure 3) presents an example in which an expert analyst entered a text about "software protection issues in the Soviet Union" from a Russian journal article.

On the left-hand side of the figure, the operators used during data entry are represented as arrows. They moved the document indexing process through five distinct stages. These operators took certain analyst's knowledge elements as input and produced results as output. The knowledge elements (or knowledge states) are shown as the shaded round-cornered rectangles. The result of each operation is represented in an oval adjacent to the knowledge element. The analyst's verbal protocol appears on the right-hand side of the figure.

In this example, the analyst began with the first stage of the process model, *template entry and checking*. He searched on the JID table (using QBF) for the proper journal name using "EKO*" (a wildcard search), the name which appeared in the Russian article sent by the Soviet scientist. This produced seven matches, or seven journals beginning with the letters "EKO" (first darkened box on the upper left hand corner). The analyst decided to search further. He used the INGRES query language, QUEL (which is more powerful than QBF), to retrieve all texts that came from the journal "EKO" (one of the seven journals that appeared in the original JID search) and retrieved 56 texts. He performed the same search on "EKOKIEV" (another journal that appeared in the original JID search) and retrieved 13 texts. After comparing the results of the two QUEL searches, "EKO" (56 texts) vs "EKOKIEV" (13 texts), he chose "EKO" as the JID based on his previous experience working in this subject area. As he indicated in his protocol: "I would guess this came from NOVOSIBIRSK (the source location of the journal "EKO"). They do a lot of software development". Then, he went on to check the REFID using "GELB86" (REFID convention: the first four characters of the author name plus the year of publication), the unique identifier for each text in the database. No matches resulted from that search (i.e. the REFID was unique), so the entry "GELB86" became the new REFID for this article. The analyst performed one more search, on the ORGID table, to retrieve the full acronym for a Russian organization "GOSKOMIZOBRETENIYE", which appeared in the source article. The first stage took 8 min.

FIGURE 3. A problem behavior graph example.

The analyst then spent the next 25 min translating and synthesizing the article. He frequently consulted an English–Russian dictionary during this *text entry* process.

The third process involved assigning keywords. The analyst proof-read the text entered and marked significant words in the text or added additional terms. The analyst chose keywords based on his knowledge of the subject without consulting any other external sources or tables in the database. This stage lasted for 5 min.

The file assignment process, which novices found most difficult, was easily done by this expert. He consulted his personal file directory (hardcopy) for this subject

area. Since he knew this area very well, he had no difficulty in sending the text to two files that he had previously created, "SOFTLAW.GEN" (about software legal issues) and "SOFTDSTRB.GEN" (about software distribution). As he remarked: "... This ADD corresponds to my file system on software protection and legal issues. ...".

He then sent the completed article template to the parser, which performed syntactic and semantic checking of the text. This entry was parsed without error and was written to the database. The entire document entry and indexing procedure took 50 min.

5.2.2. Novice/expert comparison

Detailed examination of the stages involved in the global process model revealed overwhelmingly different expert and novice behaviors. In general, experts were able to apply more specific and powerful keyword and file assignment rules. They used the system's search options more effectively, and their subject area knowledge permitted them to synthesize and abstract the source documents.

1. *Template checking*: During the *template checking* process, novices made marginal use of the system's search tools such as VIEW, QBF and QUEL; experts, on the other hand, made extensive use of these tools to reduce uncertainty about any template field entry. We illustrate this difference graphically in Figure 4 (the center part of the figure indicates the novices' process; the right part indicates the experts'). The templates entered by the experts were more accurate and consistent than those entered by the novices. This helped guarantee the accuracy of the knowledge base for research collaboration.
2. *Text entry*: When performing *text entry*, novices usually entered the text verbatim, without synthesizing or editing extensively. They never translated text from a foreign language while using the system. Experts exhibited confidence in both translating and synthesizing while interacting with the system.
3. *Keyword assignments*: Both experts and novices browsed the text to elicit terms for keywording. Novices, however, frequently exhibited a lack of consistent behavior when assigning keywords. They even expressed regret about the lack of keywording guidelines. Experts seemed to have precise rules of thumb for selecting keywords, although these rules varied widely among individuals. For example, one expert analyst keyworded sparingly in order to avoid useless or "dead" keywords (keywords that are too general to be useful, such as "hardware" or "information systems"). Another expert keyworded liberally to ensure that a text cannot be "lost".
4. *File assignment*: The most overwhelming differences between the novices and experts appeared with file assignment practices (Figure 4). Novices used the Mosaic directory tree to find files with similar texts. They then browsed these previously-entered texts extensively and identified other files that these texts had been sent to. This bottom-up method (texts are stored at the bottom level of the directory tree) traversed links across files through texts. This method was often very time consuming. Novices also browsed texts already in the

FIGURE 4. A framework for information management.

database to find keyword assignments. They rarely assigned a text to a product-specific file or sent a text to HOT files. (HOT files contain important or timely information on topics that are the subject of current research by one or a group of analysts in the group.)

Experts also browsed the directory tree, but using a top-down approach. They looked at sub-directories to identify the correct files for the text, rather than going down to the actual texts. They did not traverse links between files and

texts, rather, they relied on a personal copy of directory or files, a subset in their particular area of interest and expertise, compiled after years of experience. Since they were aware of the current research activities among the analysts, experts often sent texts to HOT files without difficulty. Because they were more knowledgeable about specific products in the subject area, they also frequently sent texts to product specific files, a practice rarely adopted by novices.

5. *Text parsing*: When preparing text for the parser, novices performed a great deal of checking in the first stage, *template checking*; while some experts relied upon the system's parser to discover their errors. Experts used error menus for correcting and refining text.

5.3. A FRAMEWORK FOR COLLABORATIVE INFORMATION MANAGEMENT

We summarize the differences between experts and novices in their processes on the right hand side of Figure 4. The left hand side of the figure indicates the knowledge elements that contributed to the actions performed in the process. This picture presents a framework for understanding of information entry and indexing behavior in a collaborative research environment.

As described earlier, four areas of knowledge are considered essential for performing document indexing in a ICRS. These areas include classification scheme knowledge, subject area knowledge, system knowledge and collaboration knowledge. These knowledge elements affect an analyst's document indexing process.

Template entry and checking require understanding of the system and knowledge of the classification scheme. Entries must use correct controlled vocabulary adopted by the classification scheme. Effective use of search tools requires a good understanding of the system's functionalities. This is indicated by the links from the "System Knowledge" and "Classification Scheme Knowledge" boxes to the "Template Checking" box.

Entering text mainly requires knowledge of the subject area such that synthesis and translation can be employed. It seems appropriate for experts to focus their text entry around basic facts and for novices to enter text verbatim for fear of excluding important information in the sources of the text.

Both keyword and file assignment rely upon classification scheme knowledge, subject area knowledge and collaboration knowledge. For example, one novice analyst failed to keyword a text properly due to lack of understanding of the term and its relevance to the text. Some novices searched extensively in the part of the directory tree that was unknown to them, trying to verify a file assignment. Experts with a personal directory combined both classification scheme and subject area knowledge in one effective representation. Collaboration knowledge, in particular, facilitated accurate keyword and file assignment and enhanced effective group communication.

Text parsing demands system knowledge in order to deal effectively with error messages and corrections. Novices with less system experience seemed reluctant to deal with error menus.

From this discussion, it is clear that a correlation can be established between the knowledge elements the analysts possess and their document indexing process. The

greater the expertise in these four areas—classification scheme, system, subject area and collaboration—the more effective and less uncertain the process of indexing. This finding confirms our prior study that investigated the information retrieval process (Chen & Dhar, 1990c). We believe our framework has added to the understanding of the information indexing and retrieving process in general. Our findings also identify the impacts of knowledge-related individual differences on the group research collaboration effort. In order to let a research collaboration process function effectively and efficiently, ICRS design needs to consider these individual differences and provide on-line assistance to different types of users. We discuss the implications of our study in detail in the next section.

6. Implications for research collaboration

The process of research collaboration occurs over time, with people who have different interests and varying levels of expertise, and it frequently involves the integration of past research efforts. Ideally, the process should also provide the basis for future research efforts. In a collaborative setting, effective indexing is critical to providing information sharing across researchers and over time. The indexing process facilitates sharing and communication of information. The indexing problem is made especially difficult by changes over time in a number of areas. In this section, we outline evolutionary indexing problems in the areas of people and tasks, the underlying system, and the information base, and then propose a means of dealing with these problems in a collaborative context.

6.1. EVOLUTIONARY PROBLEMS

Time changes everything. The people we work with, the tasks we work on, the system and information we use to support our research may all change over time. In a collaborative research environment, these changes interact with one another to compound the information indexing and sharing problem.

1. *People and tasks:*

Research collaborators may have different levels of experience, abilities and interests, and these change over time. Task focus may change, both for individuals and the group. The composition of the collaborating group can change considerably over time, changing the mix of experience and interests, and affecting the behavior of the individuals (McHenry *et al.*, 1990). Individually as indexers, researchers display inconsistent indexing behavior over time. They use their own vocabularies and personalized information organization methods. Their conceptual models of the information base vary considerably. As researchers become more expert and as tasks change, their indexing vocabulary and behavior changes.

Much of the inconsistency outlined here is possible because of the uncontrolled nature of the indexing terms. In a collaborative research environment, this is an absolute necessity because researchers conceptualize their information very differently. While they share a common area of interest, researchers work on related topics but frequently for different reasons. Task-specificity may demand indexing specificity. This requires flexibility in

indexing vocabularies. The very nature of research requires extensible indexing vocabularies, as new terms are added frequently. Controlled vocabularies that are extended only by group agreement limit flexibility, and often force researchers to use terms that are less semantically rich. With little index standardization across researchers and over time, however, the information sharing problem becomes more difficult.

2. *Information base:*

Research frequently builds upon past efforts. This requires an information base of writings, ideas, communications and annotation that is added to constantly. The information base is constantly growing, and only rarely is information retired. The group information base discussed here offers significant benefits to users, as, on average, 48% of the documents used to support research citations were entered by someone other than the author(s). (McHenry *et al.*, 1988). Information entered into the system at a given point in time by an individual for a specific reason may be used at another point in time by someone else for an entirely different reason. It is extremely difficult to anticipate these future situations and to index accordingly.

In the collaborative research environment discussed here, both information and indexes have grown at an almost constant rate over the last eight years (see McHenry *et al.*, 1990 for a complete description). Currently, there are over 36 000 pieces of text, 35 000 unique keywords, 800 directory nodes and 10 000 files. As index vocabularies grow, it becomes increasingly difficult to assign index terms correctly. Retrieval becomes more difficult and recall and precision are reduced. While full text search capabilities are available, they are rarely used because the time necessary to complete a full text search is prohibitively long, thereby placing additional importance on the indexing process.

Index maintenance of uncontrolled vocabularies is made difficult by the sheer size of the vocabularies. For instance, an initial manual attempt to eliminate incorrect index terms failed because, at the time, there were over 30 000 unique terms and over 200 000 assignments. A heuristic algorithm was necessary to identify suspected incorrect terms. After a group inspection of the identified terms, this effort resulted in over 1800 being eliminated. Still, it was a significant undertaking, even using automated support. A similar problem exists with controlled vocabularies when generating and using unique identifiers for references and organizations. Unique identifiers facilitate communication of information between researchers and ease retrieval. An additional burden is placed upon researchers at data entry time, however, as they are required to assign the previously entered identifier correctly. Heuristics are in place to assist this process, but require the researcher to wade through candidates. Many times even this is problematic, especially in translation work, because of incomplete or inconsistent information. A failure to index correctly in this situation results in a proliferation of terms for the exact same entity, and increases the semantic redundancy in the information base, resulting in confusion and poorer retrieval.

These problems are particularly perplexing for the novice, who encounters an increasingly steep learning curve as the information base and index vocabularies continue to grow.

3. *System:*

As people become more expert in their subject areas and in the use of the system, they place new demands on it to support their efforts. As tasks of the group and individuals change, the underlying system must accommodate those changes. While these are aspects of many types of systems, it is especially significant in group applications, as they tend to be more communication-oriented and less structured than automation applications. Communication between group members facilitates learning, both of the subject area and the system, and learning improves indexing behavior and retrieval performance. System extensibility supporting group communication is important and difficult to provide.

Researchers desire interfaces that map to their conceptual models of the information base. Their conceptual models differ widely and require considerable flexibility in the user interface, which ranges over the entire novice-to-expert user spectrum. Providing this kind of flexibility across a wide range of user expertise is extremely difficult. A researcher typically wants to model information using his or her personal view, and to operate on it using a personal process model. Integrating both of these at the system level across users usually results in a level of complexity that makes information sharing more difficult and makes the system much harder to learn. Often, parts of the system built for a specific person or task will fall into disuse, but will not be eliminated in a timely manner, which adds unnecessary complexity. Index terms fall into disuse as tasks and research focus change.

The flexibility required by researchers working both individually and in collaboration, over time on different tasks, forces the system to be as unobtrusive as possible. This, in turn, allows the information base to be compromised because indexing is largely uncontrolled. The system must strike a balance between being too restrictive or not restrictive enough (Silver, 1988). This inherent flexibility forces the group to maintain policies and procedures which maintain indexing consistency. However, groups may suffer from gradual deterioration of data and procedures over time, as personnel changes, and as the size of the information base continues to grow.

6.2. ADDRESSING THE PROBLEMS

Indexing indeterminism in a collaborative research environment evolves over time as a function of the people, the tasks, the system and the information base. In order for a research group to survive in this volatile environment, information must be indexed and managed correctly to ensure proper retrieval. But how should this occur, given the interactions among all these factors? To address the multiplicity of problems encountered in indexing in a collaborative environment, we propose a set of solutions that:

1. aids in the indexing of concepts in multiple contexts,
 2. promotes communication and sharing of information,
 3. captures and distributes group expertise,
 4. facilitates learning of the information base, and
 5. captures and maintains an organizational or group memory.
-

Researchers require extensive flexibility in modeling information. Higher-order indexes that can abstract beyond structural differences in the information and conceptual models of researchers to the underlying semantics of information are necessary. In the collaborative research environment discussed here, the directory and file hierarchy comprises a higher-order index. By organizing logically related pieces of text into files, which are then organized as sets into directories, which are then organized hierarchically, an index is created that can be employed and navigated at multiple levels of detail. Texts are placed in an average of three files (McHenry *et al.*, 1988), enabling the association of a text with multiple concepts embodied by files. Files and directories comprise the highest-level knowledge structuring activity of the group. The directory hierarchy can be navigated to facilitate learning, and used to both broaden and narrow searches according to topics of interest at different levels of detail. Still, more automated support is necessary as the information base grows, and as files and directories proliferate and fall into disuse. Specifically, support for automatic assignment of new information into existing files, and support for using higher-order indexes to increase retrieval effectiveness, is necessary. Automatic file assignment has been accomplished initially by creating file "characterizations" based on Boolean combinations of words which appear in the text of entered information. Whenever a "characterization match" occurs, the system suggests to the researcher that the entered information be associated with the file specified in the characterization. The researcher makes the final determination as to whether a text should be assigned to a file. Our current research involves statistical- and knowledge-based means for deriving characterizations automatically from existing files, in order to relieve the expert researchers from having to create Boolean characterizations. This is a means of both capturing and distributing the expertise of the researchers who created and maintained files, in an effort to preserve their expertise in a form that will be helpful to other group members over time. Automatic association of text with files provides a context that adds meaningful information (Motro, 1986), enhancing the indexing process in its most problematic area—file assignment.[†]

File characterizations can also be used to enhance the retrieval process. Given the inconsistent indexing of group members, file characterizations may be used at retrieval time to suggest files (logically related pieces of text) as candidates for answers to a query. Since many file assignments are made without system direction, files and their characterizations represent the "answer" to queries that are defined by the files themselves. The suggestion of files as answers to queries facilitates learning the information base, and helps distribute knowledge embodied in files; their placement in the directory adds information, as do the pieces of text that comprise them. This kind of information can be invaluable as researchers try to understand a subject area that has already been researched by a former group member.

A "HOT" file is a special kind of file that increases communication of information between researchers by providing a repository for researchers to "send" text to, when they believe it to be of interest to someone. This can ensure the timely receipt of information by someone who did not actually enter it. "HOT" files can also have

[†] This statistically based capability exists and has been tested, but is not integrated into the current system as of this writing.

characterizations associated with them, which enables the automatic routing of pieces of information pertinent to a given researcher, akin to Malone's Information Lens (Malone *et al.*, 1987). Group policies which define current research interests can also result in the creation of "HOT" files and characterizations, thereby creating a portion of the organizational memory that is necessary for effective, on-going operation of the group members.

Specific index reduction techniques must be used to identify semantically equivalent indexes, minimize redundancy, and identify indexes no longer being used. This must be accomplished at each index level, for instance, for keywords as well as files. The system should attempt to identify semantic equivalence and suggest this equivalence to researchers, who will then be able to make a determination. As part of our current research efforts, we are trying to identify semantically equivalent directories and files to eliminate redundancy at those levels.

Good group communication facilitates consistent indexing behavior. Automated support in indexing and retrieval can also enhance consistent information indexing and retrieval practices. As the system begins to "suggest" to researchers which indexes to assign, including higher-order indexes, researchers can learn more about what constitutes a "good" index. This is particularly useful for less experienced researchers. More importantly, they can also learn about the knowledge structure of the "group memory". Extensive system-supported indexing principles are necessary to augment group policies and procedures to help ensure consistent information management over time.

To support over time, a set of users who differ widely in expertise, interests, and abilities and who are working together on different tasks with a changing information base requires a system that is flexible, extensible, and largely non-restrictive. At the same time, to provide continuity, the system should provide support for consistent indexing over time across researchers. To encompass the varying levels of system experience, the system should be easy to learn. Our next-generation successor to our current collaborative research environment addresses these requirements and is named CARAT, for Computer Assisted Research and Analysis Tool. (CARAT requirements appear in Lynch *et al.* (1990).)

We are making extensive use of our current research in our CARAT design efforts. CARAT provides enhanced support for information modelling to map more easily to the conceptual models of researchers. Sophisticated, high-level indexing mechanisms (including rules) capture group expertise, and facilitate the filtering, organization, and sharing of information. Snyder & Lynch (1990) describe the underlying technologies which support CARAT. An organizational memory is built and maintained as part of the base operation of the system. System flexibility and extensibility is maintained while providing dictionary, thesaurus, glossary and translation support, as well as semantic index identification and reduction. Icon-based interfaces and a streamlined approach to entering information are being used to increase the ease of use of the system.

7. Conclusion

In this paper, we report a cognitive psychology-based empirical study that investigated the individual differences in responding to the CSCW technology. Our

findings provided important insights to the information management process during research collaboration.

Four types of knowledge elements were shown to be crucial to the success of information management in research collaboration: classification scheme knowledge, subject area knowledge, system knowledge and collaboration knowledge. These knowledge elements affected the process of information entry and indexing. Subjects who were classified as experts in this research were more knowledgeable in all four areas than novices. Their subject area knowledge, classification scheme knowledge and collaboration knowledge, in particular, played an important role in shaping the correctness and completeness of the group's information structure. Both novices and experts exhibited similar processes of information entry and indexing. However, experts were able to apply more specific and powerful keyword and file assignment rules. They used the system's search options more efficiently, and their subject area knowledge permitted them to synthesize and abstract source documents in a concise and precise format. Knowledge appeared to be the major factor distinguishing group members in a research collaboration setting. For a research collaboration system to be useful, it is important to create an environment that recognizes differences among group members and provides interfaces that are most appropriate to individual users.

Our research validates use of the cognitive modeling technique for the evaluation of a CSCW application. Based on our research, we were able to pin-point design criteria for creating a more useful and "intelligent" integrated collaborative research environment.

References

- ASHBY, W. R. (1973). *An Introduction to Cybernetics*. London: Methuen.
- BAIR, J. H. (1985). The need for collaboration tools in offices. *Proceedings of the 1985 Office Automation Conference (AFIPS)*, pp. 59-68, Atlanta, GA.
- BATES, M. J. (1986). Subject access in online catalog: a design model. *Journal of the American Society of Information Science*, **37**, 357-376.
- BATRA, D. & DAVIS, J. G. (1989). A study of conceptual data modeling in database design: similarities and differences between expert and novice designers. *Proceedings of the 10th International Conference on Information Systems (ICIS-89)*, pp. 91-100, Boston, MA.
- BOUWMAN, J. M. (1983). Human diagnostic reasoning by computers: an illustration from financial analysis. *Management Science*, **29**, 653-672.
- CHAN, L. M. (1986). *Library of Congress Subject Headings: Principles and Application*. 2nd Edn. Littleton, CO: Libraries Unlimited.
- CHEN, H. & DHAR, V. (1987). Reducing indeterminism in consultation: a cognitive model of user/librarian interaction. *Proceedings of the 6th National Conference on Artificial/Intelligence (AAAI-87)*, pp. 285-289, Seattle, Washington.
- CHEN, H. & DHAR, V. (1990a). Cognitive process as a basis for document-based retrieval systems design. (under revision) *Information Processing and Management*.
- CHEN, H. & DHAR, V. (1990b). A knowledge-based approach to the design of document-based retrieval systems. *Proceedings of the 5th Conference on Office Information Systems*, pp. 281-290, Cambridge, MA.
- CHEN, H. & DHAR, V. (1990c). User misconceptions of on-line information retrieval systems. *International Journal of Man-Machine Studies*, **32**, 673-692.
- CHI, M. T., FELTOVICH, P. J. & GLASER, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, **5**, 121-152.
-

- DENNIS, A. R., GEORGE, J. F., JESSUP, L. M., NUNAMAKER, J. F. & VOGEL, D. R. (1988). Information technology to support electronic meetings. *The Management Information Systems Quarterly*, **12**, 591-624.
- DESANCTIS, G. L. & GALLUPE, R. B. (1985). Group decision support systems: a new frontier. *Data Base*, **16**, 3-9.
- ERICSSON, K. A. & SIMON, H. A. (1984). *Protocol Analysis: Verbal Report as Data*. Cambridge, MA: The MIT Press.
- FURNAS, G. W. (1982). Statistical semantics: how can a computer use what people name things to guess what things people mean when they name things? *Proceedings of the Human Factors in Computer Systems Conference*, pp. 251-253, Gaithersburg, MD.
- GOODMAN, G. O. & ABEL, M. J. (1986). Collaboration research in SCL. *Proceedings of the Conference on Computer-Supported Cooperative Work, (CSCW'86)*, pp. 246-251, Austin, TX.
- GOODMAN, S. E., MEHRER, A. L., LYNCH, K. J. & ROCHE, E. M. (1990). International developments in the information technologies: the Mosaic group at The University of Arizona. *Proceedings of the 23rd Annual HICSS Conference, Decision Support and Knowledge Based Systems Track*, pp. 605-611, Kailua-Kona, HI.
- GORRY, G. A., BURGER, A. M., CHANEY, R. J., LONG, K. B. & TAUSK, C. M. (1988). Computer support for biomedical work groups. *Proceedings of the Conference on Computer-Supported Cooperative Work (CSCW'88)*, pp. 39-51, Portland, OR.
- GREIF, I. (1986). Data sharing in group work. *Proceedings of the Conference on Computer-Supported Cooperative Work (CSCW'86)*, pp. 175-183, Austin, TX.
- GRUDIN, J. (1988). Why CSCW applications fail: problems in the design and evaluation of organizational interfaces. *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW'88)*, pp. 85-93, Portland, OR.
- GUGERTY, J. & OLSON, G. M. (1986). Debugging by skilled and novice programmers. *CHI'86 Conference Proceedings, Human Factors in Computing Systems*, pp. 171-175, Boston, MA.
- HUGUENARD, B., PRIETULA, M. J. & LERCH, F. J. (1989). Performance $\neq$ behavior: a study in the fragility of expertise. *Proceedings of the 10th International Conference on Information Systems (ICIS-89)*, pp. 101-117, Boston, MA.
- HYMOWITZ, C. (1988). A survival guide to the office meeting. *Wall Street Journal*, 35.
- IVES, B. & OLSON, M. H. (1981). Manager of technician? The nature of the information systems manager's job. *The Management Information Systems Quarterly*, **5**, 49-62.
- JACOBY, J. & SLAMECKA, V. (1962). *Indexer Consistency Under Minimal Conditions*. Bethesda, MD: Documentation, Inc.
- JOHNSON, B., WEAVER, G., OLSON, M. H. & DUNHAM, R. (1986). Using a computer-based tool to support collaboration: a field experiment. *Proceedings of the Conference on Computer-Supported Cooperative Work (CSCW'86)*, pp. 343-353, Austin, TX.
- KRAUT, R., GALEGHER, J. & EGIDO, C. (1986). Relationships and tasks in scientific research collaborations. *Proceedings of the Conference on Computer-Supported Cooperative Work (CSCW'86)*, pp. 229-245, Austin, TX.
- LARKIN, J. (1983). Problem representation in physics in mental models. *Mental Models*. D. R. GENTNER & STEVENS, Eds. Hillsdale, NJ: Lawrence Erlbaum Associates.
- LELAND, M. D., FISH, R. S. & KRAUT, R. E. (1988). Collaborative document production using Quilt. *Proceedings of the Conference on Computer-Supported Cooperative Work (CSCW'88)*, pp. 206-215, Portland, OR.
- LYNCH, K. J. (1989). *Data Manipulation in Collaborative Research Systems*. University of Arizona, PhD thesis.
- LYNCH, K. J., SNYDER, J. M., MCHENRY, W. K., GOODMAN, S. E. & HOOPES, L. M. (1990). Requirements for next-generation collaborative research systems. *Proceedings of the 23rd Annual HICSS Conference, Decision Support and Knowledge Based Systems Track*, pp. 72-83, Kailua-Kona, HI.
- MALONE, T. W., GRANT, K. R., LAI, K., RAO, R. & ROSENBLITT, D. (1987). Semistructured messages are surprisingly useful for computer-supported coordination. *ACM Transactions on Office Information Systems*, **5**, 115-131.

- McHENRY, W. K., LYNCH, K. J., CHEN, H., HOOPEs, L. M. & SNYDER, J. M. (1990). The evolution of an ICRS: What happens after several years of use. In press.
- McHENRY, W. K., LYNCH, K. J. & GOODMAN, S. E. (1988). Handling textual information in a GDSS database: experience with the Arizona Analyst Information System. *Proceedings of the 21st Annual HICSS Conference, Decision Support and Knowledge Based Systems Track*, pp. 232-239, Kailua-Kona, HI.
- MINTZBERG, H. (1983). *The Nature of Managerial Work*. New York: Harper and Row.
- MOSVICK, R. K. & NELSON, R. B. (1986). *We've Got to Start Meeting Like This: A Guide to Successful Business Meeting Management*. New York: Scott Foresman and Co.
- MOTRO, A. (1986). Assuring retrievability from unstructured databases by contexts. *International Conference on Data Engineering*, pp. 426-433, Los Angeles, CA.
- NEWELL, A. (1980). Reasoning, problem solving and decision processes: the problem space as a fundamental category. In R. NICKERSON, Eds. *Attention and Performance VIII*, Hillsdale, NJ: Lawrence Erlbaum Associates.
- NEWELL, A. & SIMON, H. A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.
- ORLIKOWSKI, W. & DHAR, V. (1986). Imposing structure on linear programming: an empirical analysis of expert and novice models. *Proceedings of the Fifth National Conference on Artificial Intelligence (AAAI-86)*, Philadelphia, PA.
- OVER, R. (1982). Collaborative research and publication in psychology. *American Psychologist*, **37**, 996-1001.
- SILVER, M. S. (1988). User perceptions of decision support system restrictiveness: an experiment. *Journal of Management Information Systems*, **5**, 51-65.
- SNYDER, J. M. & LYNCH, K. J. (1990). Technology to support groupware: the CARAT architecture. *Submitted for publication*.
- STEVENS, M. E. (1965). *Automatic Indexing: A State-of-the-art Report*. Washington, DC: U.S. Government.
- WATERMAN, D. & NEWELL, A. (1971). Protocol analysis as a task for artificial intelligence. *Artificial Intelligence*, **2**, 285-318.