Measuring the Effectiveness of Personal Database Structures

Darrell R. Raymond
Alberto J. Cañas
Frank Wm. Tompa
Frank R. Safayeni

University of Waterloo
Waterloo, Ontario, Canada
N2L 3G1

ABSTRACT

The increasing proliferation of electronic billboards, hypertexts, and other informal electronic databases necessitates effective tools for personal data structuring. An experiment was conducted to investigate subjective processes involved during structuring an online database. Ten subjects organized two hundred proverbs into hierarchical structures over four sessions and used their structures to solve queries. Structuring and retrieval activity in the online environment was markedly poorer than in a previous manual experiment. In both experiments retrieval performance was correlated to the level of distinction employed in the construction of categories.

Measuring the Effectiveness of Personal Database Structures

Darrell R. Raymond
Alberto J. Cañas
Frank Wm. Tompa
Frank R. Safayeni

University of Waterloo
Waterloo, Ontario, Canada
N2L 3G1

1. Introduction.

Large public databases need tools for personal information structuring. The paradox of such databases is that the more they increase in size and accessibility, the more they tend to hinder effective access to information. This is because indexing schemes are relatively static while large public databases tend to be highly dynamic. Subscribers to large electronic bulletin boards, for example, are constantly confronted with topics and postings that cross group boundaries or define new groups. An individual user has a fixed amount of time available to process information. If the increasing size of a database is not countered with increasing ability to retrieve, then more time will be consumed in discarding useless mail or postings than in reading and absorbing relevant items. Furthermore, as the user extracts information from the database the problem repeats itself in the small; often the result of this extraction is a collection of obscurely named files and little evidence as to their contents. Better personal information structuring tools would improve the user's ability to deal with the incessant flow of electronic data.

If automatic indexing and classification methods were more powerful than they are today, it might be argued that good centralized indexes would be sufficient for retrieval. Even so, it is clear that people habitually restructure information simply because existing structures are so often unsatisfactory. Paper documents, for example, are subject to underlining, photocopying, clipping, ‘‘dog-ears’’, highlighting, and so on — each a means of restructuring the document to provide quicker access to sections of interest. A more modern example of this kind of activity is found in programmable videocassette recorders and the phenomenon of time-shifting (see CIT (1984)). The ability to restructure the television networks' broadcast schedules is a significant factor in the popularity of the videocassette recorder.

Structuring tools of one form or another are found in several advanced systems for computer-based information manipulation; examples of such systems are reported by Englebart, Watson, and Norton (1973), Feiner, Nagy, and Van Dam (1982), Akscyn, McCracken, and Yoder (1988), and Halasz (1988). Rather than supporting traditional formal databases, these systems are a first step toward informal databases. Informal databases exhibit several interesting characteristics, of which we emphasize two. The first is the subjective nature of informal categories and organizations. A subjective organization is an information structure based on personal estimates of the use, value, or meaning of information. Thus a personal library is typically not organized according to Library of Congress cataloging rules, but instead according to criteria such as cost, format, age, or status (e.g., ‘‘borrowed’’). The second characteristic is the flexibility with which informal databases are manipulated. A flexible organization is one which can be changed (or whose interpretation can be changed) with minimal effort. Flexibility is facilitated by a lack of formalism and consistency, since these factors interfere with multiple interpretations of a structure. A file folder labelled *Accounts Receivable* has a well-defined (hence inflexible) content, but an unlabelled stack of papers can be thought of as ‘‘unimportant work’’ at one point in time, and ‘‘my overdue assignments’’ somewhat later. Similar issues are discussed by Malone (1983).

Proper evaluation of the effectiveness of informal database systems depends on an adequate understanding of the structuring behaviour that leads to their organization. Accordingly, we have investigated a measure of structure which reflects the subjective and flexible nature of personal databases. Our results show that this measure is also a useful predictor of retrieval performance on such databases, and hence can guide the design of better structuring tools.

2. A model of structuring.

To understand structuring implies both an knowledge of the activities common in a structuring task, and a model of the internal mechanisms governing these activities. In an online environment such as an electronic billboard, users submit and receive large quantities of information in small packages or units. These units are processed in many sessions over a long period, and during that time the needs and activities of users may change. In each session with the online system, new information is processed and either rejected or integrated with an existing personal database. Typically only a small fraction of the total information is entered into the personal database. The position of any given unit in the structure is often determined by the same mechanism that would be used to retrieve it. This mechanism should be the focus of any attempt to explain the structuring task.

There are few studies which consider the subjective organization of data in an online environment, but none that we are aware of have a quantitative component. As a result, we must first decide what to measure. Information structuring is a relatively high-level activity, so it is probably insufficient to measure keystrokes as done by Card, Moran, and Newell (1980). Gross aspects of structure such as the number of categories or the average size of a category are at a high level, but also seem to suffer from an indirect relationship to the subjective judgements of the structurer. Another possibility is to analyze the labels chosen by subjects for their categories, perhaps similar to the method used by Jones and Landauer (1985). While such a study might be indicative of subjective judgements, it is hard to produce a uniform quantifiable comparison of labels. Furthermore, the assignment of labels to categories is distinct from the activity involved in generating the category. It is desirable to measure category generation as directly as possible.

Structuring items involves distinguishing them from one another, a process which can be carried out at several levels of precision. Fermented grape beverages, for example, may be classified into a single category (‘‘wines’’), or they may be split into a few major sets (‘‘red’’, ‘‘white’’, ‘‘rose’’). Further distinction can be obtained by considering the type of grape, age, bouquet, country of origin, vintner, container, cost, ownership, method of storage, or many other factors. The organization of an informal database requires distinctions to be drawn between units of information, and these distinctions are almost always highly subjective. Our maxim is that good information structuring requires the establishment of an appropriate level of distinction.

We will assume that online databases are organized around information in compact, self-contained units known as *items*; clusters of similar items will be called *categories*. A collection of such categories over a set of items will be called a *structure*.

The choice of a given level of distinction for a structure depends upon several factors. One factor is the limited knowledge of the person doing the structuring; for example, people who are unfamiliar with wine may not know the difference between bordeaux and burgundy, and hence they are incapable of including such a distinction within their structure. A second factor is limited resources; people do not often select the maximal level of distinction of which they are capable because of the cost of doing so. Generally there is an implicit task or purpose perceived for the structure, which affects the selection of a level of distinction. The marginal cost of potential extra distinction is balanced against the marginal value of that distinction, in order to arrive at an acceptable solution.

It is important to note that the level of distinction employed in producing a structure cannot be inferred from physical properties of that structure. Intuitively, the number of categories and the average category size seem indicative of the level of distinction, since well-defined categories often contain few members, and a high level of distinction tends to produce many categories. However, for a given set of physical properties there are many possible structures, not all of which are semantically appropriate. For example, there are more than 11 billion ways of choosing 4 categories of 5 items from a set of 20, but most of these are not suitable for any task. We must measure the level of distinction employed in the creation of the structure in such a way that it involves the subjective component directly.

Given two specific items, a measurement of the level of distinction can be made by asking for a spatial approximation. The person responsible for the distinction is asked to place the items close together if they seem similar, and far apart if they seem different. If a scale is provided, the relative distance can be given a numerical value; we will call this the *subjective distance* between the two items.

A category inherits a level of distinction based on the accumulation of the pairwise subjective distances between its members. The closer its members seem to be to each other, the more well-defined is the category, and the lower is the *variability within the category*, denoted as V. Determining the subjective distance between several items simultaneously is somewhat difficult. A spatial indication of the level of distinction may be clumsy or impractical if it includes many items. However, it is possible to approximate V for a category by measuring the subjective distance between the *most representative* item of the category and *least representative* item. The determination of which items are most and least representative is made by the same person who provides the subjective distance.

Similarly, a set of categories inherits a level of distinction based on the subjective distance between its members (which are categories, rather than items). The more dissimilar the member categories are relative to each other, the more well-defined is each individual category. We refer to such a set as having high *variability between categories*, denoted D. We can approximate D by measuring the subjective distance between the most representative elements of each category in the structure.

V and D can be combined to arrive at a measure of the overall level of distinction used to construct the structure, which is called R or *variability ratio*. R is defined as $R = \frac{\bar{V}}{D}$, where $\bar{V}$ is the mean of V for the component categories of the structure. Small values of R correspond to succinct, well-defined categories which are quite distinguishable from one another. Large values of R correspond to loose, ambiguous categories that are less distinguishable from one another. We expect R to be less than one for good structures, since the average variability within categories should be less than the variability between categories. We conjecture that for a given task (i.e., class of data and class of queries) there is a range of R that will result in the best retrieval performance. Structures with a smaller R than optimal will generally have categories that are more discriminating than the queries. Structures with a larger R than optimal consist of categories with many irrelevant or unrelated items. In either case retrieval performance will be reduced.

3. The structuring experiment.

We wanted to observe people performing a structuring task that closely simulated the processing of information into an online database. Several factors were important:

• conduct task in an online environment
• storage for the purpose of retrieval
• avoid memory effects
• facilitate evolution of structures
• emphasize subjective characteristics

Our first decision was that subjects should perform their tasks in a working online system. Though paper-based simulations are useful indicators and important for comparison, we considered the use of an online system to be essential in capturing unknown variables and problems in the online structuring task. The system employed should also be representative of an existing class of systems so that test results would have some relevance to system designers. We next decided that subjects should not just create a database, but should be required to use their database to perform a non-trivial retrieval task. Their effectiveness in solving the retrieval task would reflect the effectiveness of their structuring. It would also permit us to test the correlation of R with structuring effectiveness. Furthermore, a known task would encourage subjects to work at producing useful structures. To avoid the possibility that subjects might use memory rather than their structures to solve the retrieval task, the number of stimuli to be structured would be large. A large stimulus set entails several experimental sessions per subject, but this would have the advantage of simulating the repetitive access common in online situations. Multiple sessions would also permit us to observe the structures as they evolved. In addition, a large number of stimuli would encourage subjects to budget their structuring time, also a common feature of realistic situations. The need to budget time would emphasize the tradeoffs involved in the choice of a level of distinction.

Given these criteria, the next most important issue was the choice of stimuli to be structured. We wished to discourage structures based on simple mechanical classifications (i.e., chronological, alphabetic, or functional), and wanted to select stimuli that encouraged flexible, subjective distinctions. At the same time, it was necessary to employ concise stimuli so that a large number could be accommodated without overly taxing the subjects. It was also necessary to be reasonably confident that subjects had equal knowledge of the stimuli. We rejected *recipes* and *office documents* because they tend to be organized along simple, previously learned dimensions. Alternatively, pilot studies showed that *famous quotations*, while being short, were so thought-provoking that subjects had difficulty in choosing satisfactory categories. *Newspaper articles* require a significant amount of reading and are susceptible to classification by key words or phrases.

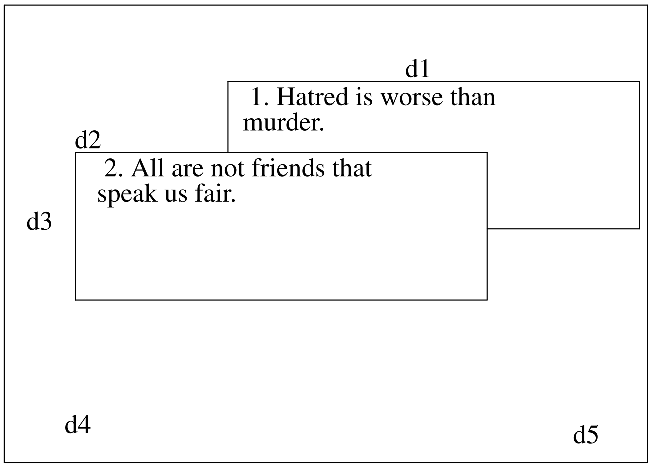
We decided that the subjects should organize *proverbs*. Pilot studies showed that proverbs are easily comprehensible during a session, but are flexible enough to permit various categorizations. For example our subjects interpreted *He laughs best who laughs last* as belonging to categories labelled *silence*, *triumph*, *winning*, and *wisdom*. Subjects were asked to play the role of ‘‘proverb manager’’ for a hypothetical newspaper. In each of four sessions they would receive a set of online proverbs, add them to an existing organization, and then find solutions to queries such as *Find a proverb which points out that hindsight is always better than foresight*.

4. The online system.

We required an online system with three characteristics: it should be capable of presenting unstructured stimuli; it should support flexible structuring; and it should maintain a detailed record of structuring activity. Several existing systems were rejected because they concentrated on aspects other than structuring or because it was difficult to add the necessary experimental features. Instead, we implemented a simple but complete hierarchical *structure editor* in order to retain close control over the system.

Previous pilot experiments and a full-scale manual experiment conducted by Cañas, Safayeni, and Conrath (1985) had shown that people rely heavily on spatial strategies to organize proverbs. As subjects processed proverbs, they arranged them on the desks or floor, clustering related proverbs and categories via spatial proximity. Large categories were often overlapped so that important items were more visible than less important ones. The online analogue to this situation is commonly called the "desktop" metaphor. Here windows represent desktops and icons represent items; example systems are described by Negroponte (1979), Negroponte (1981), Herot (1980), Smith, Irby, Kimball, and Harslem (1982), and Shniederman (1983). Because of the growing popularity of desktop interfaces, we chose to construct our editor in this style.† Our icons were short strings of text, with proverbs represented by strings of the form **d***i*, where *i* ranged from 1 to 200. Proverbs could be spatially arranged by moving the appropriate icon with a mouse. The subject could examine the proverb by pressing a button on the mouse; this would open a small window and display the proverb's content. Figure 1 shows the initial display employed to familiarize subjects with the editor. Proverbs *d1* and *d2* are visible in windows below their icons.

Figure 1. Demonstration Session Display
(containing five proverb icons and two windows)



The desktop also contained categories created by the subject. Each category was represented by a short string of the subject's choice, and was spatially manipulated just as the proverbs were. Subjects could move proverbs (or other categories) into a category by positioning them on top of the destination category's icon. Subjects could view the contents of the category by "entering" it (moving the cursor to the icon and pressing a mouse button); this action would display a new desktop in which proverbs could be organized and more categories could be created. We refer to the initial desktop as the *root* category of the structure. By permitting nesting of desktops, the editor facilitated construction of arbitrary depth and breadth hierarchies which were spatially organized at each node. A maximum of 1700 characters could be displayed on any one desktop.

All categories created by the subject contained initially the system-supplied category *back* which enabled the subject to return to the parent category (i.e., towards the root). *back* also served as a "tunnel" through which categories and proverbs could be moved to other parts of the hierarchy. *back* always appeared in the lower lefthand corner of the desktop, enabling users to

---

While there is significant intuitive weight to such designs, it should be noted that work by Jones and Dumais (1986) and Dumais (1985) throws doubt on the efficacy of spatial organizations.

move quickly to the root with repeated clicks of the appropriate mouse button.

Each desktop also contained a cyclic list of objects which were waiting to be positioned on the desktop. Proverbs that were moved to a category were not displayed directly on the desktop, but appended to the category's cyclic list. Only one member of this list was visible at any one time, in the lower righthand corner of the desktop. Before each session, the proverbs to be organized were appended to the cyclic list of the root desktop by the experimenter. Subjects would then obtain their proverbs from the root list. In Figure 1, *d5* is the current member of the list. Using function keys, subjects could rotate the list forward or backward, or view the content of the current member without removing it from the list. The root's cyclic list enabled us to present the experimental stimuli with minimal spatial bias; it also simulated the essential operation of electronic bulletin board programs.

In addition to these features, subjects could make an unlimited number of copies of each proverb (but not copies of categories) at any time. Copies had the same label as the original. Subjects could also close all open proverb windows and leave just the icons visible with a special function key.

One of the issues in designing the structure editor was how to decide when adequate functionality was provided. In particular, it seemed reasonable to provide subjects with a ‘‘trash can’’ or other means by which unwanted objects could be removed. Similarly, the ability to re-label proverbs is a natural one. The trash can facility is merely a specific instance of an existing capability, since users could easily create a category called ‘‘junk’’ and move unwanted objects there. Relabelling a proverb can be simulated by moving the proverb to a singleton category with the desired label. We were not sure how to balance augmented functionality against the extra training time, the more complicated experimental measures, and the extra development time for the editor, so we decided to keep the editor to a minimal set of features and look for evidence that these were insufficient.

Two types of data were automatically collected in addition to recording the subject's structure. First, the editor maintained a detailed log of the subject's activity which enabled us to examine each session in detail. The log consisted of timestamped records of the invocation of every function other than simple cursor motion. Second, special facilities enabled the experimenter to insert data about performance in the subject's session log during retrieval.

The editor was developed on an IBM PC/XT running Waterloo PORT, a multi-process message-passing operating system. The display was produced with an Electrohome QUICKPEL board generating NAPLPS graphics displayed on a 19" Sony KX1901-A monitor. A three-button Hawley mouse was used as a pointing device.

5. The experiment.

Ten undergraduate students at the University of Waterloo were paid for their participation in the experiment. All subjects had English as mother tongue; none had expertise in computer or library science. Each subject played the role of ‘‘proverb manager’’ for a newspaper, organizing a set of proverbs over four sessions and then solving queries. Two hundred proverbs were extracted randomly from Fergusson (1983) and Tripp (1970) and split into sets of 50, 75, and 75 for classification in the first three sessions.

Session 1 began with a short training session to familiarize the subject with the features of the editor. The training session included examples of structuring activity on a small set of proverbs not included in the experimental stimuli, as well as examples of the retrieval task. Subjects were allowed to practise until they felt confident in using the editor. The remainder of Session 1 was spent organizing the first 50 proverbs. Subjects were allowed to keep notes on paper if they wished during the sessions, and were free to ask questions about the use of the editor at any time.

Session 2 began with a retrieval task performed on the structure created during Session 1. The experimenter asked 10 queries one at a time; for each query the subject located any and all proverbs thought to be useful answers. The retrieval task of Session 2 was followed by classification of 75 new proverbs.

Session 3 was identical to Session 2 except that 15 new queries were solved (on the structure as created in Session 1 and modified in Session 2) and 75 new proverbs were given for further classification. Session 4 consisted of 30 new queries for solution and measurements of subjective distances for randomly selected categories. At the end of Session 4 subjects answered a general questionnaire about the editor. The duration of each session was controlled by the subject, typically requiring two to three hours.

During Session 1, the experimenter suggested to each subject that a category named *junk* be created so that any errors could be placed there. The experimenter added categories *1-50* and *1-125* to each subject's structure before Sessions 2 and 3, respectively. These categories contained only cyclic lists with the proverbs encountered up to (but not including) the respective session. The subjects were told that these categories need not be examined, but would enable a quick look at previously categorized proverbs if it was thought that some previous proverbs might belong in newly created categories.

Queries and solutions were developed by a person not otherwise participating in the experiment. Some example queries and their solutions are shown in Table 1. *Rewording* queries were derived from a single proverb whose words were slightly modified to produce the query. *Situation* queries were derived from a single proverb and presented a situation for which that proverb seemed most appropriate. *Multiple response* queries were situation queries that admitted several proverbs as solutions. *Non-existent* queries were derived from proverbs not contained in the stimulus set.

The measurement of *R* was carried out during the last session. Subjects were asked to choose the most and least representative proverb for a set of randomly selected categories. Subjects were provided with copies of the proverbs on 3 by 5 cards, and a 1-meter scale with ten gradations. The experimenter placed one of the proverbs at the extreme left of the scale; the subject placed the other at a point that would indicate the relative similarity of the two proverbs, and the experimenter noted this value.

During the retrieval part of the sessions, the experimenter logged the time at which the query started, the times at which solutions were located, and the time that subjects indicated that no solution existed or no further solution could be located. Retrieval performance was calculated as the hit rate (percentage of correct answers) multiplied by 100 divided by elapsed time in seconds.

6. Results.

Variability and performance measures. Table 2 gives variability and performance measures for each subject. The variability within categories was less than the variability between categories for all except one subject. This exception had very poor retrieval performance as might be expected when intra-category differences are larger than the differences between categories.

With the exclusion of subject 4, variability ratio seems to be a good predictor of retrieval performance. This subject reported headaches during the last session, and the experimenter observed that she was not able to concentrate while providing results. Both her retrieval performance and measurements of subjective distance are suspect. Exclusion of this subject results in a strong inverse linear relationship *($r=0.86$, $F(1,7)=19.23$, $p<0.01$)* shown in Figure 2. Inclusion of this subject would result in a correlation which is not significant *(linear: $r=0.55$, $F(1,8)=3.39$,*

Table 1. Example Queries (**bold**) and Responses (*italic*)

1. Rewording queries:

Find the proverb that says something like: You shouldn't judge a man until you have tried walking in his shoes.
Don't judge any man until you have walked two moons in his moccasins.

Find a proverb that says something like: If one keeps one's mouth shut, one won't say anything wrong.
Silence never makes mistakes.

2. Situation queries:

A proverb is needed which addresses the importance of desire or want in the accomplishment of goals.
Where there is a will, there is a way.

Your editor is writing an article about overeating and wants a proverb which stresses its serious consequences.
The glutton digs his grave with his teeth.

3. Multiple-response queries:

Find all proverbs about old age.
Even if we study to old age we shall not finish learning.
Age is a bad traveling companion.

Find all proverbs about the importance of sleeping.
Sleep is a priceless treasure; the more one has of it the better it is.
The beginning of health is sleep.

p=0.10).

Table 3 gives some simple objective measures of structure pertaining to the physical size of structures and categories. We did not include the category *junk* or the categories *1-50* and *1-125* in our totals. The most interesting result is the large range of all three measures. The total number of categories ranged from 240 to 20, mean category size ranged from 1.02 to 17.60, and the number of root categories ranged from 7 to 54. Our subjects clearly had differing ideas about the physical composition of an appropriate structure. No subject created a structure more than three levels deep.

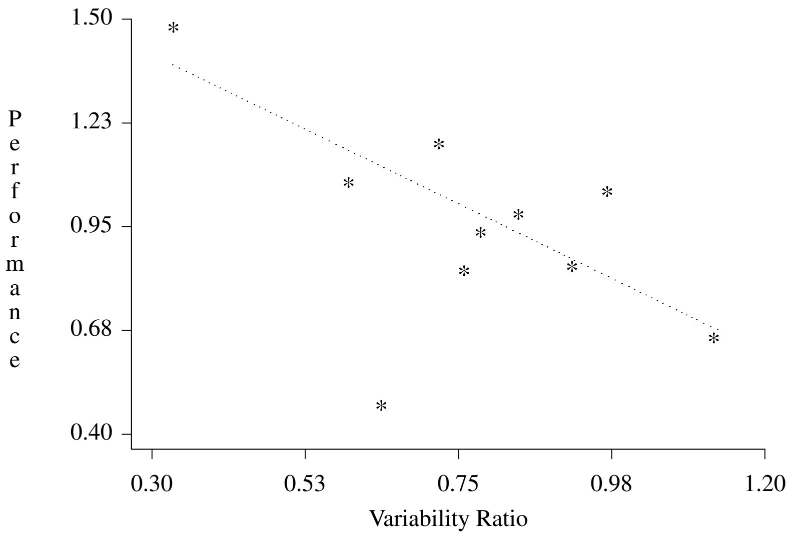
Use of space. At the root level, eight subjects organized their categories in column order, starting at the top left corner. One subject organized in row order starting at the top left; one appeared to place categories randomly. Perhaps the most interesting result is that a mean of 93.8% (*s=7.27*) of subjects' categories occupied the same desktop position in Session 4 as in Session 1. A mean of 85.7% (*s=14.45*) of all categories and proverbs were left on the cyclic list of the category to which they belonged. These results suggest that subjects generally did not attempt to vary the spatial organization of their structures.

Comparison with manual systems. Table 4 contrasts the current experiment with the earlier manual experiment conducted by Cañas (1985). All comparisons in Table 4 are significant at the

Table 2. Variability and Performance Measures

Subject #	Variability V	D	R	Retrieval Hit %	Time	Performance
1	6.8	7.02	0.97	0.85	82.9	1.03
2	6.5	7.76	0.84	0.69	71.5	0.97
3	6.6	7.20	0.92	0.65	78.6	0.83
4	4.5	7.07	0.64	0.61	133.0	0.46
5	4.6	6.07	0.76	0.61	74.8	0.82
6	5.8	7.41	0.78	0.71	77.4	0.92
7	4.6	6.38	0.72	0.79	68.6	1.15
8	4.5	7.64	0.59	0.68	64.8	1.05
9	2.6	7.84	0.33	0.77	52.7	1.46
10	7.5	6.67	1.12	0.74	116.1	0.64
Mean	5.4	7.12	0.74	0.70	82.1	0.93

Figure 2. Retrieval Performance vs. Variability Ratio



0.01 level with the exception of mean category size. Retrieval performance was significantly better in the manual experiment, with percentage of correct answers higher and elapsed time smaller. The difference in elapsed time is reflected in the number of categories visited. Subjects in the manual experiment typically went directly to the subcategory containing the desired proverb without looking at intermediate categories, a procedure not permitted by the editor's design. Categories were better defined in the manual experiment, as reflected in the smaller mean V and larger mean D. It is interesting to note that the mean category size was somewhat larger in the manual experiment, suggesting that category size is not directly related to either retrieval performance or variability. More copies of proverbs were used in the manual system, despite the ease with which copies could be generated in the editor. This may suggest that a need for copies was not

Table 3. Objective Measures

Subject	Mean Category Size	Number of Categories (*total*)	(*root*)
1	9.23	22	22
2	6.79	29	7
3	7.03	32	32
4	3.71	107	33
5	3.13	62	54
6	7.85	27	19
7	1.02	240	27
8	6.11	35	20
9	15.25	60	30
10	17.60	20	19
Mean	6.77	63.4	26.3

perceived in the online environment, or perhaps that subjects found it more difficult to keep track of copies in the editor's structures.

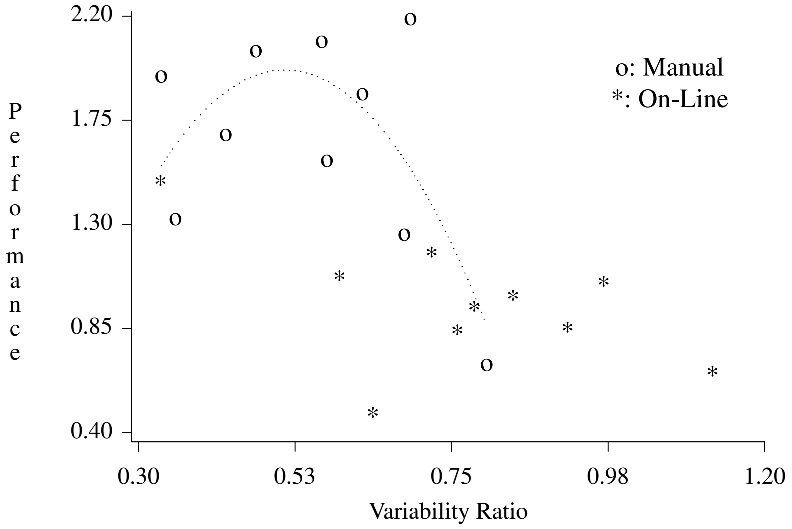
Table 4. Comparison of Manual and Editor Experiments

	Experiment (Mean, *s*) *Manual*		*On-line*
% Hits	0.81 (*0.20*)	>	0.71 (*0.22*)
Elapsed time	54.16 (*28.48*)	<	82.05 (*34.24*)
Categories visited	1.58 (*0.46*)	<	2.56 (*0.90*)
V	4.01 (*2.09*)	<	5.40 (*2.54*)
D	7.45 (*2.18*)	>	7.11 (*2.08*)
Number of categories	52.80 (*45.89*)	<	63.40 (*67.50*)
Category size	6.51 (*6.25*)	>	4.08 (*5.47*)
Number of copies	17.31 (*2.71*)	>	13.15 (*1.71*)

Figure 3 contrasts retrieval performance and R for both experiments, showing the optimum range of R in the inverse quadratic relationship obtained in the manual experiment. It is not possible to treat subjects in both experiments with the same correlation because of differences in experimental procedure. In particular, subjects in the manual experiment were asked to provide subjective distances in each of the four sessions, and were also asked to give short descriptions of each of their categories. The experimenter observed that as subjects performed these tasks, they realized that their structures could be improved and proceeded to make the necessary changes.

Another important difference in procedure was that subjects in the manual experiment often ordered the proverbs in their categories from most to least typical. Such an ordering is not possible in the editor-based structures without extensive reorganization on the desktop. Furthermore, this ordering is not captured by variability measures; an ordered category has the same value for V as an unordered one. We observed that ordering resulted in better retrieval, as subjects often knew the approximate position of the solution proverb within the category if it was ordered. These observations lead us to believe that the better performance of the manually-produced

Figure 3. Retrieval Performance Comparison



structures was at least partly a result of the subject's greater knowledge about the *R* of their structures.

Function usage. Table 5 shows the subjects' usage of the editor's functions. These functions can be organized into three groups: *list* functions (forward, back, display current proverb in list), *spatial* functions (position on desktop, enter a new category, and show a proverb), and *categorization* functions (move object to a category, create a category, copy a proverb). The table shows the normalized mean number of invocations, standard deviation and percentage. The normalized mean is the mean number of invocations per proverb; it gives some indication of the effort expended to organize a single proverb independent of the session. Normalized figures indicate that use of list functions decreased over the sessions, use of the spatial functions remained relatively constant, and use of categorization functions increased.

The total number of function invocations and their distribution becomes more meaningful if one considers a hypothetical ‘‘lazy’’ categorizer, who would expend minimal effort. Such an organizer would merely look at a proverb (forward and display), occasionally make a category (make), and move the proverb to the category (move). The lazy categorizer would invoke a maximum of 4 functions per proverb, of which two would be list functions and two categorization functions.

Subjects averaged more than three times as many function invocations as the maximum effort of the lazy categorizer, showing that they were investing significant effort in categorizing. However, the distribution of function usage was similar to that of the ‘‘lazy categorizer’’; most effort was concentrated in list functions and moving objects to categories, with spatial manipulation used very infrequently. This reinforces our earlier observation that subjects did not experiment with various types of spatial organization while developing categories.

Questionnaire results. The subjects, none of whom were computer specialists, rated themselves average in computer experience. They found the editor easy to learn and gave it a high overall rating. Display of proverbs seemed the easiest function to use, with list manipulation, copying, and spatial positioning about equal. Category creation was rated the most difficult activity. Subjects claimed they almost never wanted to remove categories. Subjects thought they

Table 5. Function Usage

Function	Session1 norm†	Session1 %	Session2 norm†	Session2 %	Session3 norm†	Session3 %
list	*11.67(8.11)*	*73.11*	*8.64(3.83)*	*64.94*	*7.82(6.18)*	*59.56*
forward	5.71(5.46)	35.75	4.40(2.75)	33.05	3.36(3.49)	25.63
display	5.17(2.57)	32.36	3.55(1.25)	26.65	3.71(2.03)	28.28
back	0.80(0.73)	5.00	0.70(0.36)	5.24	0.74(0.87)	5.65
spatial	*2.05(2.43)*	*12.87*	*1.83(2.36)*	*13.72*	*2.10(2.95)*	*15.97*
position	0.59(1.25)	3.68	0.29(0.47)	2.18	0.44(1.21)	3.34
enter	1.29(0.99)	8.11	1.40(1.68)	10.54	1.64(1.85)	12.47
show	0.17(0.36)	1.08	0.13(0.29)	1.00	0.02(0.04)	0.15
category	*2.24(0.89)*	*14.02*	*2.84(1.70)*	*21.34*	*3.21(1.69)*	*24.47*
move	1.63(0.58)	10.20	2.15(1.00)	16.13	2.52(1.13)	19.23
create	0.39(0.15)	2.43	0.39(0.65)	2.97	0.31(0.43)	2.40
copy	0.22(0.28)	1.39	0.30(0.40)	2.24	0.37(0.41)	2.84
total	15.60(7.75)	100.00	12.70(3.40)	100.00	12.60(7.43)	100.00

† Invocations per proverb (Mean, $s/\sqrt{R}$)

spent equal amounts of time exploring categories and looking at the list, with half as much time in spatial positioning and fixing mistakes.

7. Discussion.

The complexity and duration of the experiment meant that we could not test a large number of subjects, as would have been desirable. Our subjects provided interesting and consistent results which clearly indicated a correlation between variability and performance. However, there were not enough subjects to fix this relationship more precisely.

The editor contained three different types of structuring tools. These were the desktop or spatial dimension provided at each node, the cyclic list at each node, and the hierarchy of nodes. We expected that the desktop would be used for experimenting with temporary categories which would eventually become explicit members of the hierarchy, since we had observed this type of behaviour in the manual experiment. In particular, we expected subjects to group related proverbs spatially without explicit categorization until groupings exceeded a threshold size or complexity. At this point the group would coalesce into an explicit, labelled category. The cyclic list was intended merely as a convenient means with which to present stimuli and as a holding place for objects that were being moved around the hierarchy. Our subjects, however, had other ideas.

The training session included examples of overlapping and clustering strategies, since we were convinced that these were the best structuring possibilities within the limitations of our simple editor. Despite this training bias, subjects made very little use of either clustering or overlapping strategies, and as previously noted, made very little use of spatial functions in general. One counterexample is subject 7, who showed subjective clustering at internal structure nodes. His was the largest structure, with the root organized in alphabetic columnar order as shown in Figure 4. This subject's performance in the first retrieval session was quite dismal; recognizing this, he spent a great deal of time re-organizing his structure. After re-organizing, his retrieval improved dramatically and was second-best overall. Figure 5 shows the subcategory *judgement*; note that *notbylooks* and *appearance* are close together and separate from *inhisshoes* and *experience*.

Figure 4. Root Categories for Subject 7

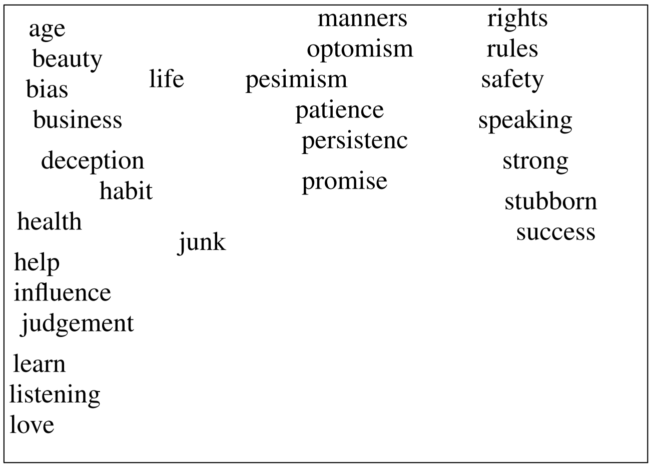
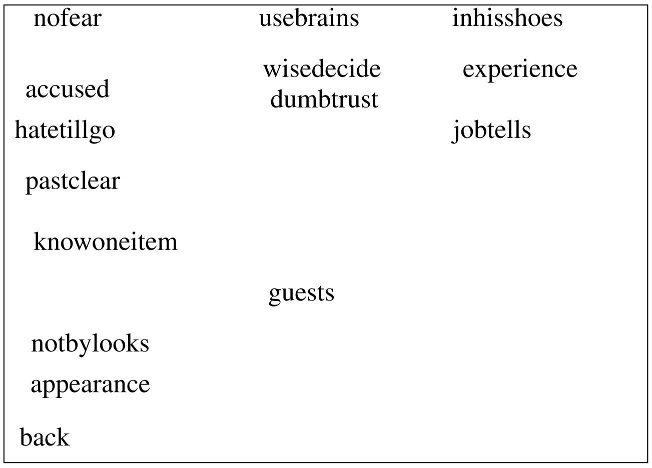


Figure 5. Subcategories of *judgement* for Subject 7



Another indicatation of the spatial dimension can be found in the structures of subject 8 and subject 3. Subject 8's root level contained two categories labelled *home*, while Subject 3's root level contained two categories labelled *senses*. These identically labelled categories had no proverb in common, and subject 8 in particular was unaware of the collision until the experimenter pointed it out. The spatial position of the identically labelled categories must have been an important index into a non-trivial memory pattern of the structure.

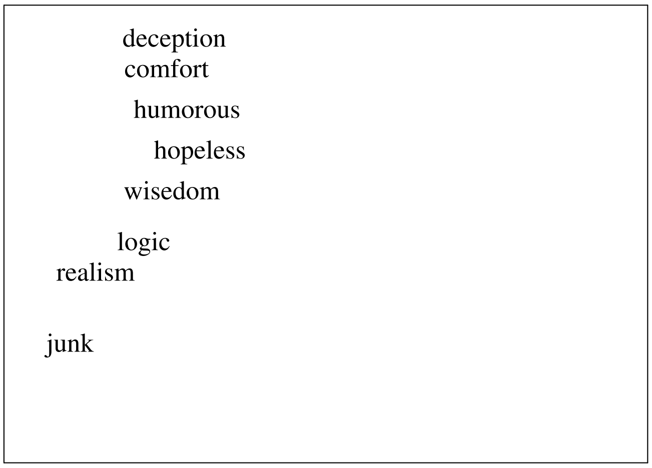
People unfamiliar with the structures of subject 8 or subject 3 would not be able to distinguish between the identically labelled categories without extensive investigation of their contents. Such an investigation would serve to create a connection between memory and space similar to that originally established by the subjects themselves. Our model does not explicitly address the role of the structure as a cue to memory. We conjecture that space is a significant aid to memory in both the manual and on-line environments.

Subjects typically categorized by following an interesting procedure that we call *hierarchical extraction*. This method consists of refining a category by choosing some closely-knit subset of its members and defining it as a subcategory. This process was carried out iteratively on the initial category as long as is deemed necessary, and then recursively on the new subcategories. The root's cyclic list was employed as an initial ''temporary'' category for the hierarchical extraction; subjects would examine this list without moving its contents to the desktop. After some small number of passes through the list, the subject would create one or more categories and move proverbs directly from the list to the category: in effect, directly from one cyclic list to another. Subjects continued to reduce the root list until it contained only miscellaneous, hard-to-categorize proverbs.

The heavy use made of the cyclic lists is evidenced by the fact that 85% of the proverbs and categories remained in some list and were not moved to the desktop. We did not expect that such a large fraction of objects would be considered miscellaneous at some level of distinction, or that the lists would so facilitate hierarchical extraction that they would replace the use of temporary categories in the form of spatially clustered proverbs. We conjecture that the driving motive behind hierarchical extraction is to avoid structuring ambiguous objects.

Some studies of menu hierarchies have focused on objective measures such as mean category size and ''depth/breadth'' parameters; see Raymond (1986) for a survey. We were curious to know if these objective measures could be of some use in predicting the performance of our subjects. Our method differs from other work in that we encouraged the use of copies and did not provide the subjects with pre-existing structures. Nevertheless, we did attempt to find correlations between performance and mean category size, total number of categories, and number of root categories. No significant correlation was found. Subjects 2 and 5 provide an illuminating example of the extreme range in objective measures; the roots of their structures are shown in Figures 6 and 7, respectively.

Figure 6. Root Categories for Subject 2



Each subject's root desktop was essentially a menu to a hierarchical data base. The root menus for subjects 2 and 5 are the most extreme ones constructed, in the sense that all other users had menus that contained more items than subject 2 but fewer than subject 5. The great difference in the appearance of their root menus might lead us to predict that performance would also be quite different, yet these subjects had essentially equal performance. Their structures were also quite close in R value.

Figure 7. Root Categories for Subject 5

While these results do not invalidate work on "depth/breadth tradeoff" or studies of other objective measures, they do indicate a limited range of applicability for such results. Objective models evaluate the mechanical effort involved in traversing a structure. We would not be surprised to learn that this effort is in some cases less important than the mental effort required.

What kind of fundamental limitations are faced when using a system that employs a desktop metaphor? Contrasting the results of this experiment with the results of the manual experiment indicates that online structures tend to a larger variability ratio. We must therefore explain how the editor interfered with our subjects' ability to make adequate variability judgements.

Subjects had limited ability to see and evaluate their environment compared to the manual experiment. The editor permitted subjects to display at most two or three proverbs simultaneously without overlapping. If the desktop contained several objects, subjects would avoid covering them with windows, further reducing the amount of space available for structuring. Subjects could see only the immediate contents of a category unless they navigated through the structure, a time-consuming process.

The subjects' ability to manipulate the environment was also greatly limited compared to the manual experiment. Since subjects could only manipulate what was on the screen, reduction in vision also constitutes a reduction in manipulation capability. Furthermore, subjects were effectively limited to manipulation of single items. In the manual experiment, a simple sweep of the hand would suffice to move a spatially contiguous temporary category to a new location. A similar task in the editor would require a tedious process of moving objects one by one to the new location. As one pilot subject observed, moving proverbs on the screen is similar to using a magnet to move objects kept under glass.

Since subjects could only evaluate a small part of their structure, the subjective quality of their structures would tend to a local rather than global optimum. Since subjects could manipulate their structures only with difficulty, the cost of temporary categories exceeded their perceived marginal value, and hence they were not often employed.

Our implementation does not employ the most advanced hardware. While we expect that a higher resolution display or faster processor would make the interface more pleasant, we do not think such modifications would result in a fundamental difference unless improvement by orders of magnitude were attained. A real desktop provides a space continuum that is qualitatively distinct from a discrete display device employing several virtual screens for presentation of one or more dimensions. The subject's perception of the continuum undergoes continual visual refresh as the subject scans the structure. By contrast, a discrete display device requires explicit, conscious action for refresh. The real desktop permits arbitrarily fine adjustments to be made to the spatial contiguity of various parts of the structure so that it matches the subjective contiguity; however, information that is on different screens in a discrete display seems separate no matter how closely the screens may be linked in the overall hierarchy.

Perhaps more importantly, the manual environment also includes highly-developed manipulative tools (i.e., hands) with powerful group-oriented functions. Using one's hand to push some proverbs to the side of the table is a simple manual activity, but it has complex structuring implications. Its most immediate purpose is to render the moved set less important by moving it out of the centre of vision, but it often also results in increased clustering of the items in the set. This clustering reinforces both the increased variability between categories (the set is spatially more distinct from its neighbours) and the decreased variability of the category (the members are seen as more alike in their unimportance). At the same time, the clustering preserves much of the relative spatial organization within the category and thus it can be reconstituted at a later date if the decision to move it was too hasty. Finally, the clustering increases the amount of overlap in the set and hence reduces the amount of information that must be evaluated when considering further structuring moves.

8. Conclusions and future directions.

Our subjects learned the editor very quickly and gave it a high rating for ''user-friendliness'', confirming the general notion that desktop interfaces are pleasant, fun to use, and quickly learned. However, we have identified a significant, quantifiable distinction between such interfaces and the real desktops they attempt to emulate, namely the added interference in making and preserving variability judgements. This result has important implications for design: improvements in the interface will not result in improvements in task performance unless they directly address the assessment of the subjective quality of the personal database.

We suggest two general approaches to more effective structuring, which we refer to as the *manipulation* and *evaluation* approaches.

The manipulation approach concentrates on augmenting existing tools with capabilities that encourage the use of temporary categories and permit a wide base of comparison. Group-oriented manipulation analogous to the activity of the hand might be a first step, though Halasz and Moran (1982) recommend care when using analogy. Other means of improving manipulation might reduce the number of discrete steps in navigating a structure. For example, Raymond (1984) suggests that simple menus be replaced by *multi-menus*, which permit multiple (rather than single) choices at each node and show more than one level of descendants. Multi-menus reduce the the number of discrete steps needed to explore the environment by permitting larger steps and by displaying a structured view of the environment. Proper implementation of multi-menus requires intelligent automatic management of limited display space, but is completely independent of analogies to a manual environment.

The evaluation approach concentrates on making the user more aware of the value of R, so that the structure can be appropriately adjusted. This approach is based on our observation that improved performance is partly a result of extra feedback about the variability of the structure. Evaluation might be carried out during the structuring process by automatic selection of appropriate elements of the structure for comparison to the element to be structured. Conversely, evaluation might be conducted by an off-line tool that would resemble an English style-checker — a *variability checker*. The user would indicate which parts of a structure were doubtful, and perhaps give some indication of the precision with which checking should be performed. The structure checker would then obtain subjective distance measurements from the user in order to compute R for the structure. The structure checker would indicate a relative measure of subjective goodness, and might also suggest where improvements are most necessary or could be most beneficial.

Finally, what of variability and the model of categorization? Our measures were incomplete pictures of the structures, since they did not include the effects of ordering within categories, nor were they expressive of the hierarchy of categories. These and various other inadequacies could be rectified in a future experiment in order to obtain a more precise correlation of R with retrieval performance. Even at the current level of investigation, however, we observe the importance of subjective quality of information structure, and this observation is invaluable in setting the direction for design of more advanced computer structuring tools.

9. Acknowledgements.

We would like to thank Bert Bonkowski of the Software Portability Group at the University of Waterloo for his unfailing courtesy and support during development of the structure editor. Our thanks also to Mert Cramer for his suggestions on an early draft of this paper, and to Dave Conrath for his involvement with the manual experiment. We gratefully acknowledge the financial support of the Natural Sciences and Engineering Research Council under grant G1154.

References

R.M. Akscyn, D.L. McCracken, and E.A. Yoder, "KMS: A Distributed Hypermedia System for Managing Knowledge in Organizations," *Communications of the ACM,* 31, 7, pp. 820-835 (1988).

A.J. Canas, "Variability as a Measure of Semantic Structure for Document Storage and Retrieval," *Ph.D. Thesis,* Department of Management Science, University of Waterloo, Waterloo, Ontario (1985).

A.J. Canas, F.R. Safayeni, and D.W. Conrath, "A Conceptual Model and Experiments on How People Classify and Retrieve Documents," *8th International ACM SIGIR,* Montreal, Quebec (1985).

S.K. Card, T.P. Moran, and A. Newell, "The Keystroke-Level Model for User Performance Time with Interactive Systems," *Communications of the ACM,* 23, 7, pp. 396-410 (1980).

CIT, *Videodisc Opportunites and Options,* Communications and Information Technology Research Ltd., London, England (1984).

S.T. Dumais, "A Comparison of Symbolic and Spatial Filing," *CHI '85 Conference Proceedings,* pp. 127-130, San Francisco (1985).

D.C. Engelbart, R.W. Watson, R.W., and J.C. Norton, J.C., "The Augmented Knowledge Workshop," *Proceedings of the 1973 AFIPS National Computer Conference,* pp. 9-20 (1973).

S. Feiner, S. Nagy, and A. Van Dam, "An Experimental System for Creating and Presenting Interactive Graphical Documents," *ACM Transactions on Graphics,* 1, 1, pp. 59-77, ACM (1982).

R. Fergusson, *The Penguin Dictionary of Proverbs,* Penguin Books Inc., Markham, Ontario (1983).

F.G. Halasz, "Reflections on Notecards: Seven Issues for the Next Generation of Hypertext Systems," *Communications of the ACM,* 31, 7, pp. 836-852 (1988).

F.G. Halasz and T.P. Moran, "Analogy Considered Harmful," *Proceedings of the CHI '82 Conference on Human Factors in Computing Systems,* pp. 383-386, Gaithersburg, Maryland (1982).

C.F. (1980) Herot, “Spatial Management of Data,” *ACM Transactions on Database Systems,* 5, 4, pp. 493-514.

W.P. Jones and Landauer, T.K. (1985), “Context and Self-Selection Effects in Name Learning,” *Behaviour and Information Technology,* 4, 1, pp. 3-17.

W.P. Jones and S.T. Dumais, “The Spatial Metaphor for User Interfaces: Experimental Tests of Reference by Location versus Name,” *ACM Transactions on Office Information Systems,* 4, 1, pp. 42-63 (1986).

T.W. Malone, “How Do People Organize Their Desks? Implications for the Design of Office Information Systems,” *ACM Transactions on Office Information Systems,* 1, 1, pp. 99-112 (1983).

N. Negroponte, “Books Without Pages,” *IEEE International Conference on Communications,* pp. 56.1.1-56.1.8, IEEE, Boston, Massachusetts (1979).

N. Negroponte, “Media Room,” *Proceedings of the Society for Information Display,* 22, 2, pp. 109-113 (1981).

D.R. Raymond, “Personal Data Structuring in Videotex,” CS-84-7, Department of Computer Science, University of Waterloo, Waterloo, Ontario (1984).

D.R. Raymond, “A Survey of Research in Computer-Based Menus,” CS-86-61, Department of Computer Science, University of Waterloo, Waterloo, Ontario (1986).

B. Shneiderman, “Direct Manipulation: A Step Beyond Programming Languages,” *Computer,* pp. 57-68, IEEE Computer Society (1983).

D.C. Smith, C. Irby, R. Kimball, and E. Harslem, “The Star User Interface: An Overview,” *Proceedings of the AFIPS National Computer Conference,* pp. 515-528, Houston, Texas (1982).

R.T. Tripp, *The International Thesaurus of Quotations,* Thomas Y. Crowell Co., New York, N.Y. (1970).