UNIVERSITY OF CAPE TOWN

Individual Document Management Techniques:

an Explorative Study

A dissertation submitted to the Department Of Computer Science, Faculty of Science at the University Of Cape Town in partial fulfilment of the requirements for the degree of Master of Philosophy (in Information Technology).

By Mpho Sello

February 2007

Supervised by Dr Hussein Suleman

Abstract

Individuals are generating, storing and accessing more information than ever before. The information comes from a variety of sources such as the World Wide Web, email and books. Storage media is becoming larger and cheaper. This makes accumulation of information easy. When information is kept in large volumes, retrieving it becomes a problem unless there is a system in place for managing this.

This study examined the techniques that users have devised to make retrieval of their documents easy and timely. A survey of user document management techniques was done through interviews. The uncovered techniques were then used to build an expert system that provides assistance with document management decision-making. The system provides recommendations on file naming and organization, document backup and archiving as well as suitable storage media. The system poses a series of questions to the user and offers recommendations on the basis of the responses given.

The system was evaluated by two categories of users: those who had been interviewed during data collection and those who had not been interviewed. Both categories of users found the recommendations made by the system to be reasonable and indicated that the system was easy to use. Some users thought the system could be of great benefit to people new to computers.

Acknowledgements

Table of Contents

Chapter 1 – Introduction and Motivation

1.1 Introduction

Information plays a major role in the activities of organisations. People in organisations need to exchange information while carrying out their duties. In some organisations, information is the sole product, for example libraries and publishers, while in other organisations it serves as support to organisational products, for example user manuals. Regardless of the nature of organisation, information is used to facilitate the undertaking of all organisational activities. The information used is stored and retrieved as necessary.

Before the advent of information technology, information was contained mainly on paper. The paper was kept in indexed files to make it easy for users to access it. When the filed information was no longer used regularly, it was put into storage as records. The records were either stored on the organisational premises or entrusted to an organisation that dealt specifically with storage of records.

Storing information on paper posed a number of problems to users of the information. Files were sometimes inappropriately filed or misplaced by some users and this made it difficult or impossible to find them. Sometimes files were lost or stolen. Storage of old files also posed a problem because it required organisations to have physical storage space or pay for storage by other organisations.

With the advent of information technology, some of the information storage problems were solved. Large amounts of information could be stored electronically on small storage media that do not require large physical storage spaces. However, problems like inappropriate filing were not solved entirely and other new problems came about. With the wide choice of storage media available, users tend to store more information. Due to the increase in stored information, retrieving it is not always easy. Users forget where they have stored information or the names with which they have stored the information. Sometimes the medium on which information is stored fails, gets misplaced or becomes obsolete.

Information stored electronically is usually contained in documents. It is these documents that users often have problems accessing. These problems have inspired an area of research called document management. Document management research is concerned with devising ways of making information storage and retrieval easy for users. Document management is a relatively new research area, but it follows in the steps of earlier research efforts that focused on file organisation.

Doing a search on document management on the Internet does not yield many results. Even books on document management are hard to find. From the little research that has been conducted in this area, it seems that researchers approach the problem from varying angles. Some researchers have focused on document storage and retrieval and others have focused on presentation of stored documents on retrieval.

1.2   Description of the problem

As storage media become larger and cheaper, users are able to store more information than ever before. This sometimes results in users keeping information that they would otherwise not keep, for instance, information that can be easily put together when required or information that is not likely to be needed again in future. Keeping this extra information then results in problems when trying to locate other, more important, information.

The most common problem encountered by users when trying to retrieve information, is forgetting document names or the storage medium on which the information is stored. This is often the case when the document being sought has not been accessed for some time.

Specific document management problems encountered by users include:
- Users lose stored information because they cannot retrieve it.
- Users struggle to locate old documents.
- Users forget the names with which they store documents.
- Users are unable to retrieve information because they cannot remember where they stored it.

- Users have no systematic way of storing documents.
- Users lose information because the media containing it fails.
- Users cannot keep track of documents stored over a period of years.
- Users have devised their own methods of managing documents, but these methods are not fail-safe.

1.3　Methodology and evaluation

The aim of this project was to study expert users' habits and techniques for managing documents and use this information to build an expert system that would help average users to manage their documents better. The study examined the techniques employed by users for storing and retrieving information and these were incorporated into the expert system.

The expert system can be used to make decisions about how to store documents, what documents to backup, what documents to archive and how to do so in a way that will make retrieval of the stored documents easy. Through the system, the users can also make decisions about the best media to use for document storage, backup and archiving.

A study of users' habits and techniques for managing documents was carried out. This was done using a questionnaire that attempted to discover users' habits and techniques for managing documents. A sample of typical information workers was chosen randomly from the UCT community; lecturers, administrative assistants and postgraduates. These were people who were believed to deal with large volumes of documents in their day-to-day activities.

The questionnaire contained mainly open-ended questions as it was meant to collect information about users' practices. Instead of distributing questionnaires to users, to fill in during their free time, respondents were interviewed by the researcher. This was done with the aim of attaining a high response rate. Due to the open-ended nature of the questions, it was feared that users might not respond well to the questions if left to fill it in on their own. When all the respondents had been interviewed, their responses were compiled and used to build the document management expert system.

Expert systems are usually evaluated at three levels. They are first evaluated for their performance, that is, their reasoning abilities and the quality of decisions they put forward to the user. The first form of evaluation is carried out by the expert system builders/developers. They are also evaluated by the experts whose knowledge was used to build the system, to establish whether they are able to give the correct information. Lastly they are evaluated by users - the people for whom the system was built. The users' evaluation is meant to establish the system's usability and efficiency.

On completion of its development, the expert system was tested to establish its ability to assist users with making document management decisions. The system was tested on users who had been interviewed during data collection and users who had not been interviewed. As providers of the information used to build the system, and the intended users of the completed system, the users who had been interviewed tested the system on two levels. They tested it for the correctness and utility of the information it contains and also for its usability and efficiency.

1.4 Dissertation Outline

Chapter 2: Background on Document Management
The chapter gives an introduction to document management. Practices, problems and developments in the area of document management are outlined.

Chapter 3: Expert Systems
The chapter introduces the reader to expert systems. It explains what expert systems are, their makeup and how they are developed. Examples of classical expert systems are given.

Chapter 4: Data Collection and Analysis
The chapter outlines the data collection process and gives an analysis of the collected data.

Chapter 5: Development of the expert system

The chapter gives an outline of how the expert system for managing documents was developed.

Chapter 6: Evaluation of expert system and results

The chapter outlines the process of evaluating the developed expert system. It also discusses the results of the evaluation.

Chapter 7: Conclusions and Recommendations

The chapter outlines conclusions drawn from the study and makes suggestions for future work.

Chapter 2 – Background on Document Management

2.1 Introduction

Computers provide an easy and convenient way of storing documents. However, retrieval of the stored documents is not always easy when users deal with vast amounts of information and have a variety of locations where they can store their documents. The volume of information and the variety of storage locations end up becoming hindrances when users want to retrieve stored documents. The problem has become so widespread that it inspired several areas of research in information management.

Information management research has focused on areas such as file organisation, information retrieval, personal information management and document management. Despite their different approaches, these areas of research are all concerned with storage and retrieval of information. They look at the problems encountered by users when working with information and attempt to devise solutions to these problems.

The focus of this study is document management. Most of the information that users deal with is contained in documents. Users create and store documents for later retrieval. Users often have a problem storing documents in a way that will help them retrieve the documents later. This results in delays and frustrations when trying to retrieve the stored information. Users forget the names with which they stored their documents or the locations where the documents were stored. Researchers in the area of document management focus on devising ways of overcoming these challenges to make storage and retrieval of documents timely and easy.

This chapter focuses on document management and the user. It looks at users' document management practices, the problems they encounter and the efforts undertaken by researchers to help users overcome these problems.

2.2 Document management practices

A number of studies have been undertaken to examine users' document management practices with the aim of developing tools that will assist the users to improve these practices. The following has been identified as major practices employed by users when dealing with information: acquiring, organising, maintaining and retrieving [Jones 2005].

Users get information from a variety of sources and in a variety of ways. They get information that they source themselves from activities such as reading or searching the Web and they also get information sent to them by other people, such as email. Information that is considered important is usually kept for future reference or use. The kept information is only useful if it can be found when required. To ensure that the information is retrievable in future, users employ different strategies to organise and maintain the information.

Studies of user information management strategies have focused on management of paper documents and electronic documents. Findings from these studies indicate that there are many similarities in the way people manage paper documents and electronic documents. This is mainly due to the fact that computers derive their file organisation features from the organisation of paper documents.

Malone [1983] undertook a study of how professionals and office clerks physically organise information in their offices. He did this with the aim of using the observed practices to inform the design of computer information systems. He found that some people kept their information neatly organised in files, while others kept it in piles on their desktops. For those people who had their information filed, retrieval was quick and easy, but those who kept piles struggled to locate the information they were looking for.

The respondents had reasons for the ways they organised their files. Neat files were kept to make location of information easy. Piles, on the other hand, were kept to serve as reminders, to keep frequently used information within reach and because people had difficulties classifying information for filing.

Other researchers [Barreau & Nardi 1995] looked at how users organise and retrieve electronic documents. Their research was inspired by the realisation that users spend a great deal of time storing and retrieving information. They believed that the large volumes of information that users find themselves having to deal with, as a result of computer networks, further complicates these activities. The researchers believed that research in this area was lacking and as a result nothing was being done to make these activities easier for the users.

The study discovered that users prefer to search for documents by location rather than using computer search facilities. To locate a file, a user would go to a specific folder or diskette and browse through the list of documents in there, to locate the one they were searching for. They found it easier to look for a document where they had placed it instead of relying on a computer search that would return a long list of file names that could not help users to remember the file they were trying to locate. The search facility was used only if the location search was not successful and to locate documents stored on other computers, on the network.

Barreau and Nardi [1995] identified three types of information that users deal with. The type of information determined the way it was organised. The three types were ephemeral, working and archived information. Ephemeral information was used within a short time and it was usually organised loosely on the desktop for users to remember. Working information was used frequently in carrying out day-to-day activities and it was organised in folders and categorised. Archived information was old and no longer considered important. As a result, not much effort was made to explicitly organise it. Most of it became archived in the place where it was initially saved.

A study of personal information organisation by Raviso [2004], identified practices not very different from those identified by Barreau and Nardi [1995]. The users she observed were experienced and average users of computers. These users used the computer desktop to organise their files. They used it in the way they used physical desktops. They divided the desktops into different sections where they kept different

types of documents. There was a section for system resources, current documents, documents to be attended to and short cuts to applications used on a daily basis. These users searched for documents in the same way that users in the Barreau and Nardi study did.

A study undertaken by Boardman and Sasse [2004] examined the way users managed their files, e-mail and bookmarks for websites. This was done with the aim of developing a tool that would integrate the three interfaces to make their management uniform. The researchers identified different management strategies employed by users to manage their files, email and bookmarks.

There were two categories of users identified: those who filed their documents and those who did not. Users who filed their documents did so at different stages of their work. Some filed documents on creation while others filed documents when they had finished working on them. These users occasionally conducted routine maintenance of their documents and filed any documents that had been left unfiled. The other category of users did not file their documents at all. They just had a long list of files saved in one location.

Email was managed in a way not very different from files. Some users put their email in files and others just left it in the inbox. Some users filed their email immediately after reading it, while others left it in the inbox until the time when they would organise their inboxes. With bookmarks, the practices were also not very different. Some users did put bookmarks in files while others did not. Bookmarks were filed as they were created or during routine maintenance.

Retrieval of documents was done mainly by browsing through documents stored in folders. This was because users said they knew where the documents were stored. When they were not sure of the exact location, they said that while browsing through documents they were assured to come across the file they were looking for. The users depended more on location for retrieval than file names and as a result did not use the computer search facility much because it required them to know file names. The search facility was used only when a location-based search failed to locate the

documents sought. Similarly, email was retrieved by scanning the contents of the inbox. The search tool was used as a last resort. With bookmarks, users referred to them mainly for websites that had been recently accessed. For most of the websites, they conducted a new search instead of using the bookmarks. This was because they said website locations changed frequently and using a search engine was quick and easy.

These methods of retrieval were effective. Users hardly ever failed to locate the files they were looking for. The few times when documents could not be found was due to the document having been deleted, put in the wrong file or left "unfiled".

2.3   Document management problems

File hierarchies have been identified by many researchers as a hindrance to file organisation and retrieval [Lansdale 1988]. File hierarchies require strict categorisation of information and this poses a problem to users. Users usually want to view information in a way different from the way it was stored and file hierarchies do not support this. For instance, a user may want to view reports created within a specified period of time, but fail to achieve this because the reports have been categorised by subject.

Users also struggle to categorise documents because some documents fit into more than one category. They do not find it easy to assign names to categories and subsequently forget the category names. When the name of a category cannot be remembered, it becomes difficult to remember files stored in the category. This results in reluctance to categorise documents and when there is no categorisation documents become cluttered in one location, making it even more difficult to locate them.

Henderson [2004] undertook a study to establish how useful file hierarchies are to users' document management needs. This was done to identify strengths and weaknesses of software system with the aim of improving them to suit users' file organisation requirements. She found out that users have a problem keeping track of different versions of one document. The users consider these different versions as one

document, but the system treats them as separate files. They wanted a system that would support document organisation rather than file organisation.

A system that allows for maintenance of different versions of a document was developed by Ramike et al. [2004]. The system integrated document management into the Windows operating system and it could display different versions of a document in one view. When a version is selected from the list it is displayed with all the differences from the original document.

2.4 Document management research

Research in this area has been aimed at improving the way computers currently handle document management, as it does not seem to be in line with users' needs. Traditionally, storage of files in computers has been in the form of hierarchical structures. Similar files are grouped together in folders or directories that are arranged hierarchically. Dourish et al. [2000] attempted to steer away from the hierarchical system of storing files. Their study explored the use of document properties for document management applications. They developed a system that uses document properties as the basis of interaction, rather than location. They called the system Placeless Documents.

The developers of Placeless Documents see this way of interaction as being more user-oriented than the hierarchical system, because it does not require users to know the location of documents when retrieving them. The system also frees users from having to categorise files when storing them, as this can become complicated when dealing with files that belong to more than one category. Placeless Documents allows users to arrange files according to their own needs, and not the way the system dictates.

The Lifestreams model [Freeman 1996] was another attempt at simplifying document management for users. This model uses time as the basis for document management. Documents are kept in a stream according to the time when they were created. The stream has three distinct portions; the past, the present and the future. The past is found at the end of the stream, the present in the middle and the future at the beginning. The past contains old documents, the present contains current documents

and the future contains documents to be attended to in future, for example reminders and lists of things to do. When the date for documents in the future section arrives, the documents are moved to the present section automatically.

The developers of the model found the use of file names and locations not to be useful to the user as they are often forgotten. To overcome this problem, their model allows the user to store files without assigning names to them. When a user creates a new document using the "New" feature, a new document is placed on the stream. To search for documents, users make use of the find operation. This operation allows a user to enter a query stating the nature of documents to be retrieved, for example, "Find documents created last week".

Search results are presented in a virtual substream that contains all documents fitting the search criteria. Substreams can be created and destroyed without affecting the main stream. Substreams can be used to organise and filter documents. If a substream is not deleted, it continues to gather new documents that fit the original search criteria. Old documents in the lifestream become read-only and they are subsequently archived. As new documents are added to the stream, older ones fade from the user's view.

Spatial memory has also been used as a way of managing documents. Data Mountain is a system designed to help users manage documents using spatial memory [Robertson et al. 1998]. The system provides a 3D (three-dimensional) inclining environment with passive landmarks where users can store documents. Users interact with documents using 2D (two dimensional) techniques. Using the mouse, users can drag documents onto the incline and place them in any way they want using the landmarks to help them remember where they have placed the documents.

When a document is placed on the incline, the system produces cues that help the user to remember where the document has been placed. As a document is dragged to a specific location on the incline, documents already on the incline move aside to make way for the new document. The pages make a sound as they move away. The placement of a document is also accompanied by a sound that mimics arrangement of

objects in the real world. The sound changes according to the speed with which the document is moving.

Documents are stacked on top of each other on the incline in a way that documents placed underneath others are still visible from the top of the stack. Documents can be identified with thumbnail presentations and pop-up titles. As the mouse pointer hovers above a document, the title of the document is displayed. Clicking once on a document brings it to the fore, so that it can be viewed. Clicking on the document again returns it to its original location.

A document management study that was carried out by researchers from Politecnico di Milano took an approach that deviated from the studies discussed above. Instead of focusing on storage and retrieval of documents, it focused on presentation of stored documents on retrieval. Zelig document management system [Celentano et al. 1992] was developed with the aim of overcoming problems encountered when the same document is presented to people of different backgrounds and knowledge levels.

The system allows for different presentations of the same documents depending on users' profiles. Users get different views of the same document and the depth of the content is also displayed to suit the knowledge level of the user. This helps to ensure that the documents presented to users contain content that is suitable to them. The Zelig document management system operates within the Microsoft Windows environment.

Some researchers have looked at how hierarchical file systems can be improved to make it easy for users to locate files. Marsden and Cairns [2003] looked at how the principle of relational databases can be used to improve the users' interaction with the hierarchical file system. Marsden and Cairns wanted to build a model that operates more or less like email. Email is stored as a list of messages arranged in order of receipt date, but it can be re-arranged according to an attribute specified by user, such as sender or subject.

They wanted to create a model that would enable users to view files according to a specified attribute, despite the way the files are organised in the file store. They created a prototype interface of a visualisation system that could be used to view files according to different attributes. Some of the attributes were access date, file size and creation date.

Researchers from Apple Computers [Mander et al. 1992] undertook a study to examine how people manage the flow of information in their offices. This was done with the aim of developing a system that would enable users to manage their electronic documents in the same way that they manage their physical documents. The findings of the study indicated that users informally managed information by creating piles. Incoming information was kept in piles until it could be attended to. Sometimes piles were created with working information divided by subject. Piles were sometimes created to delay the process of categorisation. The documents from the piles were eventually filed after they had been acted upon.

The researchers developed a prototype using the pile metaphor. The prototype was tested on users to establish how useful such a system would be in information organisation. The system had features that allowed users to create piles of mixed content and data types. Piles could be created by stacking documents on top of each other. An item in the pile could be identified by minute depictions of the document's first page and depth of the document. The pile retained the unkempt look of a physical pile.

Other features of the system included browsing of items in a pile, system created piles and pile management. Users could flick through the pile using a mouse or fingers on a touch screen. Users could issue a command to the system to create piles based on a specified property. The system also allowed for pile management activities such as ordering of piles and creation of sub-piles. Users could order piles according to data type, subject or date. They could also use colour to distinguish items of a similar type from the others.

2.5 Concluding remarks

Many studies have been undertaken with aim of devising systems in a way that would be helpful to the user in managing their documents. Different approaches have been used to attain this. Some researchers have studied user document management practices alone while others have gone on to develop systems based on findings from user observations. Most of these studies have identified file hierarchies as a major constraint to document management, to most users. However, computers still continue to use hierarchical structures for file organisation.

Instead of working separately, researchers in document management could come together to advocate the development of systems that are beneficial to the users depending on their different information needs.

Chapter 3 – Expert Systems

3.1 Introduction

Expert systems are a branch of Artificial Intelligence that deals with knowledge processing, as opposed to data processing. Artificial Intelligence is a branch of computer science that attempts to emulate human actions using computers. Expert systems operate like human experts to give advice and solutions to problems on a specific subject area. According to Feigenbaum [1982], an expert system is "an intelligent computer program that uses knowledge and inference procedures to solve problems that are difficult enough to require significant human expertise for their solution."

Expert systems, also known as knowledge-based systems, are programmed with knowledge obtained from human experts, and sometimes written sources of expertise, on a specific subject [Giarrantano & Riley 1989]. Users can access this knowledge by giving input to the expert system and/or answering questions posed by the expert system. Expert systems give reasons and make inferences, like human experts do when providing solutions to users. They can give explanations for the conclusions they reach. Expert systems usually contain knowledge in one area of expertise. They have been used to diagnose diseases, diagnose problems with machinery, configure computers, make financial decisions and train employees new to a job, among other uses.

Expert systems were first developed in the 1970s [Giarrantano & Riley 1989]. They were developed to provide expert knowledge in the absence of a human expert. When specialised human knowledge is costly or rare, expert systems can be used to make decisions or to support decision-making. The first expert systems contained only specialised knowledge, but over time they came to contain other types of knowledge. In addition to knowledge attained from human experts, expert systems now contain knowledge from sources such as books or people whose knowledge of a subject has been acquired through experience, who may not be considered experts in that area.

3.2 Expert system structure

Expert systems are built by Knowledge Engineers. These are the people who get expertise from experts and incorporate it into the expert system. Expert systems are made up of a number of components. Even though the components may not be uniform in all expert systems, there are some components that are common to all expert systems. The following are components that are typically found in an expert system [Castillo & Alvarez 1991]:

1. **Knowledge base:** contains factual and heuristic knowledge acquired from a human expert. Factual knowledge is shared by all experts in the same field while heuristic knowledge is individualistic. Heuristic knowledge refers to personal approaches taken by experts to solve problems and therefore differs from person to person. The knowledge may be in the form of inference rules or just facts arranged in a way that they can be compared to determine the one relevant to the case at hand.

2. **Inference engine:** manipulates the facts and rules in the knowledge base to reach conclusions and provide solutions to problems. The inference engine chains inference rules to provide a line of reasoning. There are two methods of reasoning that can be used by an inference engine. These are forward chaining and backward chaining.

   ▪ Forward chaining starts from a set of conditions, moving towards a conclusion. This method of reasoning enables drawing of conclusions from existing data.

   ▪ Backward chaining starts with a conclusion and tries to establish whether or not there is data to support the conclusion.

3. **Knowledge acquisition subsystem:** allows for addition of knowledge that is not already contained in the expert system.

4. **Explanation subsystem:** analyses the processes followed by the inference engine and shows the user how conclusions were reached.

5. **Working memory:** stores data that is specific to a problem being solved as well as procedures for subsystems of the expert system.

6. **User interface:** enables a user to interact with the expert system.

3.3 Expert system shells

When building an expert system, the Knowledge Engineer has a choice of either developing the knowledge base from scratch or using an expert system shell. An expert system shell is a software package that contains an inference engine, a knowledge representation scheme and sometimes a user interface, but it does not contain any knowledge. Different types of expert system shells are available commercially for different knowledge domains. Knowledge engineers choose shells that are appropriate for their needs and incorporate knowledge. Using a shell cuts down the process and cost of building the system from scratch.

Expert system shells were first introduced by removing the knowledge base from fully developed expert systems [Nebendhal 1988]. EMYCIN is an example of an early expert system shell. It was made by removing the knowledge base from an early expert system called MYCIN. With the MYCIN knowledge base removed, EMYCIN could be used for a different expert system by incorporating a new knowledge base. The inference engine and the user interface remained the same [Giarrantano & Riley 1989]. The making of EMYCIN paved the way for expert system shells.

Since EMYCIN, different kinds of expert system shells have been made. Some are commercial while others are freely available. An example of a popular expert system shell is CLIPS (C Language Integrated Production System). CLIPS is a public domain expert system shell that was developed in 1985 and it is written in the C programming language [http://www.ghg.net/clips/WhatIsCLIPS.html 2006]. It supports rule-based and object-oriented knowledge representation schemes. The rule-based scheme allows for representation of knowledge as rules. This object-oriented representation scheme allows for modeling of components that can be re-used in different applications. Since its conception, CLIPS has been continually improved and is available in different versions.

3.4   Knowledge representation

Expert systems differ in the way knowledge is represented in their knowledge bases. A Knowledge Engineer can encode knowledge using any one or a combination of the following knowledge representation schemes:

1. **Semantic Networks:** This is the first knowledge representation scheme that was used in Artificial Intelligence [Woods 1975] and it is the one that is mostly used in expert systems. It is a network of nodes that are connected by arcs [Castillo & Alvarez 1991]. The nodes represent objects, while the arcs represent relationships between objects. The nodes are labeled with nouns and the arcs are labeled with verbs.

    Example:

    A semantic network
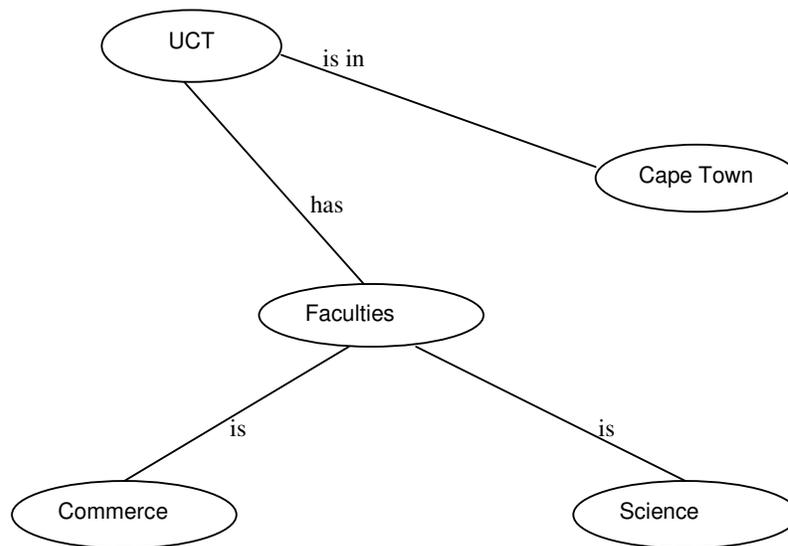
2. **Object-Attribute-value triplets**

    This is a special case of a semantic network [Harmon 1985]. It has three elements: object, attribute and value. Objects are represented by a circle, attributes by an arrow and values by a rectangle.

Example:
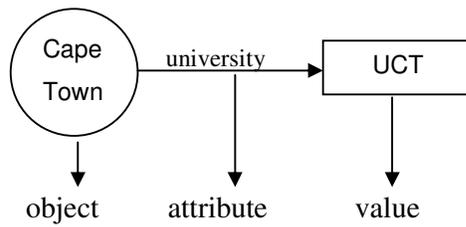
The above example of an object-attribute-value triplet would appear as follows in a semantic network:

3. **Rules**

Rules can be used to represent relations [Davis 1977]. They are made of two parts: a premise and a conclusion. A premise consists of an "IF" statement and a logical expression with one or more object-oriented triplets, connected by logical operators "and", "or" or "not". A conclusion is composed of the adverb "then" and a logical expression.

Example: *If* graduated from UCT *then* UCT alumnus.

4. **Frames**

These are a particular case of semantic networks [Castillo & Alvarez 1991]. Frames can contain information that directly describes an object or information that leads to a description of an object.

Example:

UCT	
Attributes	**Values**
Vice Chancellor	N. Ndebele
Location	See map of Cape Town
Student population	22 000
Courses offered	See list of courses
Campuses	See list of campuses

5. **Logical Expressions**

   Logic is used to determine the truthfulness of successive statements [Davis 1977]. If all the statements are true, then the argument is true. When using logic to prove something, facts are compared with arguments to determine whether or not they have the same structure. There are two common logic systems.

a) **Propositional logic:** in this logic system, propositions are either true or false expressions [Castillo & Alvarez 1991]. When the propositions are linked by logical connectors such as "and", "or", "implies", "equivalent" or "not", they are called compound expressions. The truthfulness of propositions can be determined using a set of rules that are selected on the basis of the connectors of the propositions.

   Example: the ⟶ arrow is a connective that represents conditionality. It can mean any of the following:
   - if…then
   - implies
   - only if

b) **Predicate calculus:** this was developed to make up for the shortcomings of propositional logic [Castillo & Alvarez 1991]. Unlike propositional logic that deals with complete statements, predicate calculus deals with individual objects. It is used to describe objects and their relationships. In predicate calculus, semantic rules are used to generate logical expressions from other logical expressions.

   Example:

Statements	Predicate calculus
Cape Town is a city	City (Cape Town)
UCT is in Cape Town	is_in (UCT, Cape Town)
John works at UCT	works_at (John,UCT)

The three statements are facts. From these statements, the following statement can be generated:

Statements	Predicate calculus
John is in Cape Town	is_in (John, Cape Town)

The selection of a knowledge representation scheme is dependent on the type of problem an expert system addresses.

3.5   Developing an expert system

After a problem to be solved by the expert system has been established, a knowledge engineer identifies the human expert that can provide the knowledge required for the expert system. The knowledge engineer can acquire the knowledge from the expert by means of interviews, observing the expert at work and evaluating the tools used by the expert. When collecting knowledge from the expert, the knowledge engineer establishes what the expert knows and how they reach the conclusions they make. This means that the knowledge engineer should be able to differentiate between facts that would be known to all experts in the subject area and personal ways in which the individual expert solves problems.

The knowledge acquired from the expert is then encoded into the knowledge base, according to the knowledge representation scheme selected. As knowledge is being encoded into the knowledge base, there may be a need to modify the inference engine to suit the type of knowledge being encoded.

When knowledge has been encoded into the knowledge base, a user interface is designed to enable users to interact with the expert system. The user interface should enable the user to give input and respond to questions posed by the system. The sophistication of the user interface would depend on the type of problem that the expert system solves. Some problems may require presence of graphics while for some problems text menus may suffice. Since the user interface is the point through which a user interacts with the system, it is important to make it appealing to the user because many users tend to judge the quality of an expert system largely by its user interface.

3.6 Examples of expert systems

There are many expert systems that have been developed in different subject areas. Some of them are well known and they have been used as benchmarks for other expert systems. An example of such an expert system is MYCIN, which was developed in the 1970s at Stanford University [Nebendhal 1988].

MYCIN was developed to aid physicians in diagnosing and treating infectious blood diseases and meningitis. Given symptoms of the patient, MYCIN can identify the most likely disease and give a prescription of antibiotics for treating the disease. The expert system got its name from the suffix common to most antibiotics - "mycin". MYCIN used a rule-based knowledge representation and it was the first expert system to have an inference mechanism. It was this inference mechanism that led to the development of an inference engine that came to be used in subsequent expert systems.

MYCIN was derived from an earlier expert system called DENDRAL [Giarrantano & Riley 1989]. DENDRAL was also developed at Stanford University, in 1965. DENDRAL was developed to help organic chemists analyse the molecular structure of organic chemical compounds. DENDRAL used a rule-based knowledge representation, with no inference mechanism.

PROSPECTOR [Castillo & Alvarez 1991] is another expert system developed in the late 1970s by the Stanford Research Institute, inspired by the success of MYCIN. It was developed to aid geologists in the early stages of exploring sites for minerals. Geologists fed data about a site to PROSPECTOR and it used the data to determine the possibilities of finding minerals at the site. Its knowledge base contained geological rules and a database of known rocks and minerals.

Chapter 4 – Data Collection and Analysis

4.1   Introduction

Data for building the expert system for document management was collected by interviewing respondents. The purpose of the interviews was to find out the techniques that respondents have devised for managing their documents. The techniques would then be incorporated into the knowledge base of the expert system.

This chapter outlines the data collection process and analysis of the collected data. It starts off with an outline of the sampling procedure and the data collection method used. This is followed by a presentation of data collected. An analysis of the collected data is then given and finally intermediate conclusions are drawn from the findings.

4.2   Sampling procedure and data collection method

A questionnaire was drawn up for collecting data (see Appendix A). In order to capture as much detail from the respondents as possible, the questionnaire consisted mainly of open-ended questions. Initially the questionnaires were meant to be distributed to respondents to fill in, in their own time. However, due to the open-ended nature of the majority of questions, it was decided that it would be better to interview the respondents in person, as this would ensure a high response rate and allow for clarifications.

The target population was knowledge workers within the UCT community: postgraduate students, lecturers and administrative staff. The respondents were selected because they were believed to deal with large volumes of documents in their daily activities. A sample of 21 respondents was selected randomly from the target population: 11 students, 8 lecturers and 2 administrative staff members.

Generally the questionnaire was designed to find out the following:
- how often respondents create documents,
- what procedures and media they use to store their documents and
- what techniques they use to ensure easy retrieval of their stored documents.

The questionnaire was initially tried with a small number of respondents (5), before being administered to a larger sample. After interviewing the first five respondents, some questions were modified and others were substituted with new ones. Questions that respondents found not to be clear were modified. Questions that did not seem to capture information relevant to the document management practices of the respondents were removed and relevant ones put in their place. Other questions were added to complement existing ones.

After modifying the first version of the questionnaire, interviews were conducted with the selected respondents. Interview times were set up with respondents depending on their availability. Respondents were asked questions from the questionnaire and their responses were recorded in writing. Some respondents opted to fill in the questionnaires by themselves.

4.3 Summary of collected data

a) Document editing Frequency

Early in the questionnaire, respondents were asked about the frequency with which they create or edit documents on a computer. This was to establish whether or not respondents met the sampling requirements. The study was aimed at people who deal with documents in their day-to-day activities. All the respondents met this requirement as they all said they created or edited documents on a daily basis.

b) Type of storage media

Users were asked about the type of media they use for storing their documents and the reasons for using the specified media. A list of commonly used storage media was provided and users had to choose the ones they used. The table below shows the storage media used by the users.

Table 1. Storage media usage

Storage media	Number of users
CD – Rom	18
Magnetic tape	2
Flash disk	19
Hard disk	21
DVD	11
Server hard drive	13
NAS (Network Attached Storage)	4
Floppy disk	6
iPod	3
External hard drive	3
e-mail	2
Web Drive (Gmail)	1

Users gave a number of reasons for the choice of media storage they used. The following were reasons given for using the different storage media:

CD-Rom and DVD: These have been grouped together because of the similarity of reasons given for their usage. They were used for their portability, compatibility, convenience, durability, capacity, fast access, reliability, availability, easy storage of CD/DVD itself, "re-writability" and cheap cost. They were used for the purposes of transferring documents to other computers, for backup and storage of large amounts of data.

Flash disk: The flash disk was used mainly for transferring documents to different computers. Reasons given for its usage were convenience, portability, reliability and fast access. It was also used for temporary backup, work in progress, security and personal files. Another reason cited was additional features such as an MP3 player that is available in some flash disks.

Hard disk: This is the one medium that was used by all respondents, the main reason being that it is the default storage. Other reasons cited were ease of use, capacity,

26

reliability, durability, fast access, convenience and accessibility. The hard disk was used for storing work in progress, important documents, sharing files over a network and backup.

Server hard-drive: This was used mainly for backup. Other reasons included remote access, seamless access, sharing, convenience, capacity, security and redundancy.

Magnetic tape: This was not commonly used. Out of 21 respondents, only two users made use of this storage medium. This was mainly due to its high cost and that it requires a tape drive which is not commonly available in personal computers. It was used mainly for archiving and for its large storage capacity.

Network Attached Storage (NAS): Only four respondents out of 21 indicated that they use NAS for storing their documents. The number was probably higher, because NAS was used in the Systems Manager's office, but users only knew that they were saving their documents on a network server. It was used mainly for backup. Other reasons given were redundancy, seamless access, security and privacy.

Floppy disk: The floppy was used for backup and transferring documents to other computers. It was preferred for its low cost, portability, ease of use and ease of storage.

Web Drive: Only one person used it. It was used for its seamless access.

iPod: It was used for emergency backup and for zipped files. It was preferred for its capacity, convenience and portability.

External hard drive: This was used mainly for transferring documents to other computers, backup and large files. It was preferred for its convenience, capacity and portability.

E-mail: Email was used for backup because it can be easily accessed from any computer connected to the Internet.

c) File naming and organisation

On the question of how users name and organise their files (see Appendix B1), users gave the following as the criteria they used:

i.   File names

Files were assigned meaningful and descriptive names. They were named according to content or function, by date, by title, by chapter name or given obscure names for security reasons. Some users used structured names, such as AdminReport_05_09_2006 or 20040712-219. The first name was used for reports and it was based on content and date of creation. The second name was used for data files. The format for the second name was year, month, date and day of year.

ii.   File organisation

Files were mainly organised in folders created for similar files. Folders were created on the basis of function, year and category, content or versions. Some users used the desktop for organising their files. Frequently used files were stored on the desktop for quick access. Other users used the hierarchical file structure to organise their files. Some users organised files according to format. They kept files in different formats and numbered them.

The main reason for having a system in place for naming and organising files was to aid retrieval. Some users mentioned that the system was also useful for quick storage and backup of files.

d) Backing up documents

Out of 21 respondents, 18 backed up their documents, one did not backup documents at all and two only did so rarely. Respondents used different criteria for backing up documents (see Appendix B2). Some users backed up all their documents, while others backed up only selected files. Among the selected files were important files, files used regularly, files that are not easy to recreate, files that may be required in future and documents that may have legal implications.

e) Archiving documents

As opposed to backing-up files, users did not seem to be in the habit of archiving files, at least not explicitly. For most users backed-up files ended up as archives. Out of 21 respondents, 11 explicitly archived their documents, eight did not and two only did it occasionally. Respondents used different criteria for archiving their documents (see Appendix B3). Some users archived all the contents of their hard drives, while others archived only selected documents. The selection was made on the basis of time or importance. One user archived anything that had not been used in three months and another one archived documents that were likely to be required in future. The following are documents that were archived: research data, important emails, zipped files, publications, text-based documents at the end of a project, academic work and computer programs.

f) Long-term document storage

Users had different ideas of what long-term storage is. This ranged from one month to 20 years. They also had different ways of storing documents in the long term. In addition to the storage media previously mentioned, some users kept printed copies of documents as a way of avoiding loss through storage media failure or virus.

g) Document retrieval strategies

In order to be able to retrieve documents in the long term, users employed a variety of strategies. The following are the strategies stated by users:
- Storage media is labelled and stored in a safe place.
- Files are kept on the hard drive.
- Files are stored in standardised formats, such as doc, jpeg and pdf.
- Files are stored on a network drive.
- A catalogue of stored files is kept.
- Documents are stored alphabetically.
- Durable storage media is used, such as CDs and DVDs.

h) Guarding against document format obsolescence

Files in storage can become inaccessible in the long-term, due to media or format obsolescence. Some users did take measures to ensure that their documents remain

accessible in the long-term, while others did not. Those who did gave the following as the strategies they employed to guard against format obsolescence:

- Storing files in common formats.
- Using standard Windows backup programs.
- Using standard UNIX formats, such as *.tar.
- Zipping files.
- Using Latex.
- Using applications that are backwards compatible.
- Keeping old computers.
- Saving a copy of a reader program along with documents.
- Saving files in formats that are accessible on different platforms, such as jpeg.

i) Guarding against storage media obsolescence

The following were given as strategies to guard against storage media obsolescence:

- Migrating documents to newer media, for example, from floppy to CD.
- Keeping copies of documents on removable media on the hard drive.
- Storing documents on a server.
- Using backwards compatible media such as CDs and DVDs.

j) Additional document management strategies

Some users had additional document management strategies that they used. These were as follows:

- Version control.
- Using only CD-R, to avoid erasure.
- Using Google desktop.
- Emailing documents to own inbox.
- Creating folders for old documents. The folders are named "old".
- Using a UNIX program called rsync for synchronising contents of folders.
- Temporarily storing documents on the desktop before categorising them.
- Creating a folder for new documents and moving documents to the relevant folders once attended to.
- Saving documents on different drives.

4.4   Data interpretation and Analysis

The respondents who took part in the study all had some system in place for managing their documents, even though the sophistication of these systems varied from user to user. Some users had well thought out ways of managing their documents, while others just did what worked for them without having put much thought into it. One respondent used a system inherited from a previous office incumbent.

The techniques and storage media used seemed to have a relationship with the number of documents users dealt with. Users who dealt with a large number of documents seemed to have put some effort into managing their documents, while users with a small number of documents did not seem to give much thought into how their documents were managed. They just did whatever worked and in some instances could not give reasons why they managed their documents the way they did.

The choice of storage media used by respondents differed from user to user, with some media being more widely used than others. Hard drive, flash disk and CD were the most widely used storage media. Storage media such as NAS (Network Attached Storage), magnetic tapes, iPods and external hard drives were not as commonly used. They were used by people who worked with large amounts of data, who seemed to be well versed with computer storage media.

There was a great deal of similarity in the way users named and organised their files. Most users named files using meaningful names that give an indication of what is contained in the file. These files were then saved in folders according to content or function. All respondents had folders created for similar files. This was done to aid retrieval of the stored documents.

Backing up documents seemed to be more popular with respondents than archiving as out of 21 respondents, 18 backed-up their documents while only 11 archived their documents. For most users archiving was not done explicitly. Documents that had been backed up eventually became archives. Some users backed up all their documents, while others backed up only selected documents. Archiving also followed a similar pattern. Some users archived everything while others archived selectively.

Out of the 21 respondents, only one did not keep documents in the long term because the documents they dealt with became redundant with time. Of the respondents who kept documents in the long term, some had in place strategies for ensuring that their documents remained accessible in the long term, while others did not. The strategies were meant to guard against format and media obsolescence as well as loss and damage to the storage media.

4.5 Concluding remarks

While respondents all had some document management system in place, the sophistication of these systems seemed to vary according to respondents' exposure to computers and amount of data. Users in the field of Computer Science had more sophisticated document management systems as compared to user in other fields. The length of computer usage did not seem to play a large role in the sophistication of document management systems, as some respondents had systems that were more sophisticated than those of respondents who had been using computers for a longer time.

From the interviews it became apparent that some users did not give much thought to how they managed their documents. Even though the strategies they had in place were working well for them, they seemed to believe that there was a better way for them to manage their documents.

From the data collected, there are enough document management techniques to build a small expert system. The techniques were compiled and categorised so that they could be used to address specific document management problems. The next chapter looks at how the data collected was used to build the expert system for managing documents.

# Chapter 5 – Development of the expert system

5.1 Introduction

The chapter gives an outline of how the expert system for managing documents was developed. It explains how the information gathered from interviews was processed for inclusion into the expert system. The chapter also gives an explanation of how a specification for the system was drawn up and how it was used to build the system. The chapter ends off with an explanation of the system design and operation.

5.2 System development

After data was collected, the successful document management techniques put forward by respondents were identified and compiled. The techniques were then categorised as follows:

- Choice of storage media
- File naming and organisation
- Backup of documents
- Archiving of documents
- Retrieval strategies for long-term document storage
- General document management strategies

In order to use the techniques as solutions to specific document management activities, they were matched with the relevant document management problems. The problem areas were: storage media; file naming and organisation; backup; and archiving of documents. A hierarchy of problems and solutions was then drawn. This was done by formulating questions that a user would have to answer in sequence to converge on a specific solution. The hierarchy of questions and solutions was thus constructed as a decision tree (see Figure 1).
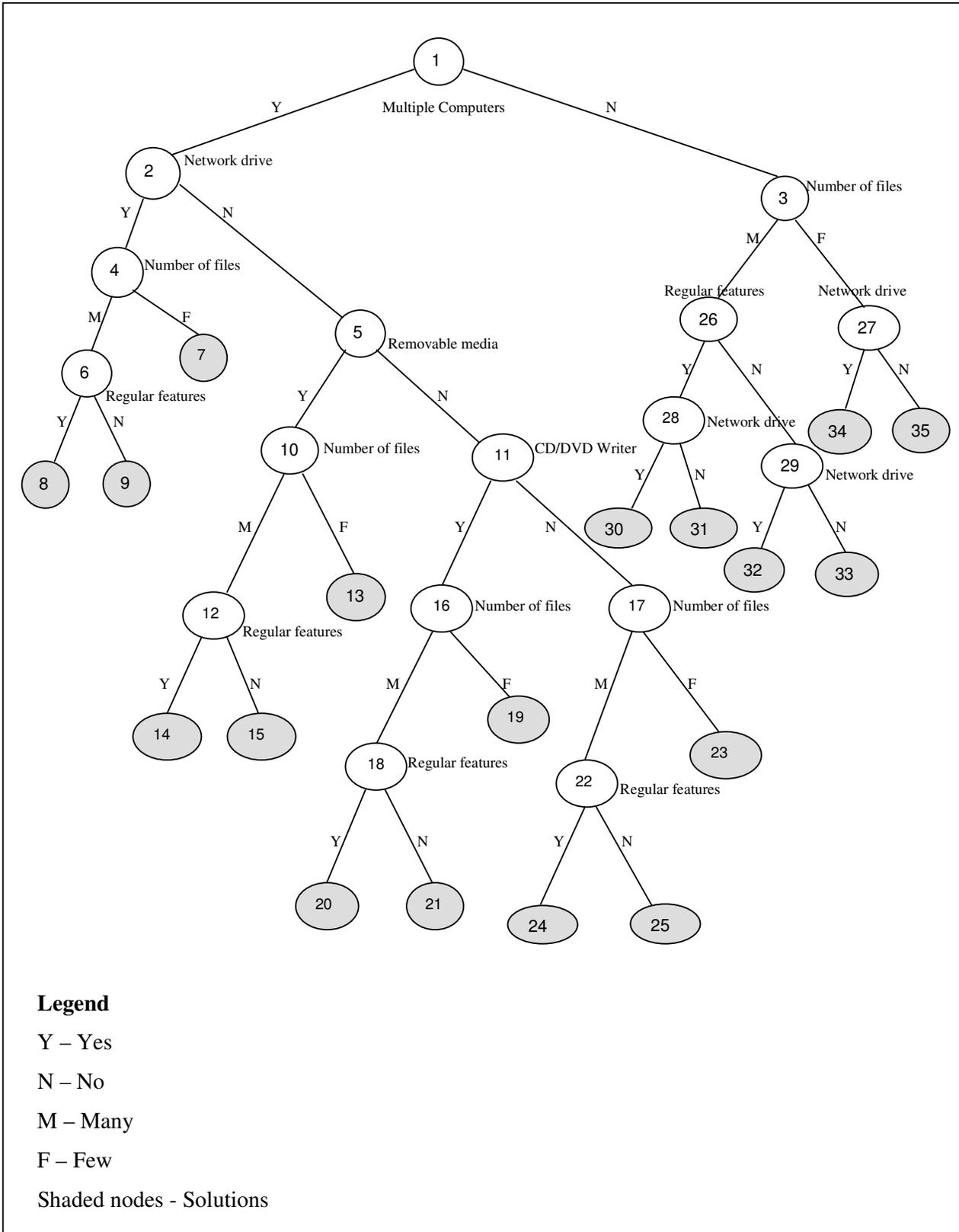
Figure 1: Decision Tree

The tree consists of branches of questions and answers that eventually lead to a solution. The tree is made up of nodes and arcs. Nodes represent questions, while arcs represent answers. The last node of each branch is a solution. The decision tree was constructed manually to organize questions and solutions hierarchically, as the system does not learn automatically. Some questions are repeated on different branches of the tree because the only way to encode the system is via branches of the tree i.e., there are no state variables. From the decision tree, a specification table for building the expert system was drawn, as shown in Table 2.

The system was initially meant to be developed using an expert system shell. However, efforts at acquiring a suitable expert system shell were not successful. Some shells were costly while others were written in unfamiliar languages. It was then decided that the expert system would be developed using a simple rule-based engine that generates static HTML pages using a specification file as input.

Table 2: Expert System Specification table

Page	Question	Answer	Page	Answer	Page
1	Do you need to work on more than one computer?	Yes	2	No	3
2	Do you have a network drive?	Yes	4	No	5
4	How do you describe the files you deal with?	Many	6	Few	7
6	Are the files of a similar type?	Yes	8	No	9
5	Do you have removable media?	Yes	10	No	11
10	How do you describe the files you deal with?	Many	12	Few	13
12	Are the files of a similar type?	Yes	14	No	15
11	Do you have a CD/DVD Writer?	Yes	16	No	17
16	How do you describe the files you deal with?	Many	18	Few	19
18	Are the files of a similar type?	Yes	20	No	21
17	How many files do you deal with?	Many	22	Few	23
22	Are the files of a similar type?	Yes	24	No	25
3	How do you describe the files you deal with?	Many	26	Few	27
26	Do the files have regular features?	Yes	28	No	29
28	Do you have a network drive?	Yes	30	No	31
17	Do you have a network drive?	Yes	32	No	33
7	Do you have a network drive?	Yes	34	No	35

The specification table (Table 2) shows how questions and solutions would be arranged in the expert system and it sets rules for arriving at a solution. It shows the pages on which questions and solutions are contained, arranged in rows and columns. The first column shows the page on which a question is contained. The second column contains the question. Each question has two responses. These are contained in columns three and five respectively. Columns four and six contain the pages to which the selected responses would lead.

In the specification table, the first page contains a question "Do you need to work on multiple computers?". There are two responses to this question: Yes and No. If the response is "Yes", the user is meant to be taken to page 2 and if the response is "No", the user is meant to be taken to page 3. Pages 2 and 3 contain different questions and

responses to these questions lead to other questions on different pages, until a solution is eventually reached. The shaded cells are solutions (see Appendix C).

The specification table was converted to a Comma Separated Value (CSV) file that could be read using a Java program. Using the CSV file as input, a Java program that generates HTML pages of questions and solutions was then written. The program generates the pages along with the links to connect them, as specified in the specification table. All pages include a "breadcrumb trail" to help the user keep track of how they got to the page they are currently viewing and also to make navigation through the system easy. A "breadcrumb trail" is a list of links to pages that have already been visited before the current page. Users can move directly to a page they have visited before by clicking on its link in the trail.

5.3 Using the system

The Java program that generates HTML pages is saved in a folder named DocMert together with the CSV specification file. In the same folder, there is another folder called "output" to which the HTML pages generated when the program is run are written. The program generates the Questions and the Solutions pages. The output folder contains an Images folder and two HTML pages that are not generated by the program. These are the Index (Home page) and the General Recommendations pages. The pages are part of the system and they have links to the pages generated by the program.

To use the system, the program is compiled and executed. After execution the index page in the output folder is opened. The index page links to the page that contains the first question in the system. From the first question page the user can use links to get to other pages. This is done by clicking on the relevant response to each question asked until a page containing solutions is reached. From the solutions page, the user can go back to any of the previous pages by clicking on the relevant link in the breadcrumb trail.

5.4 Snapshots of the system's pages
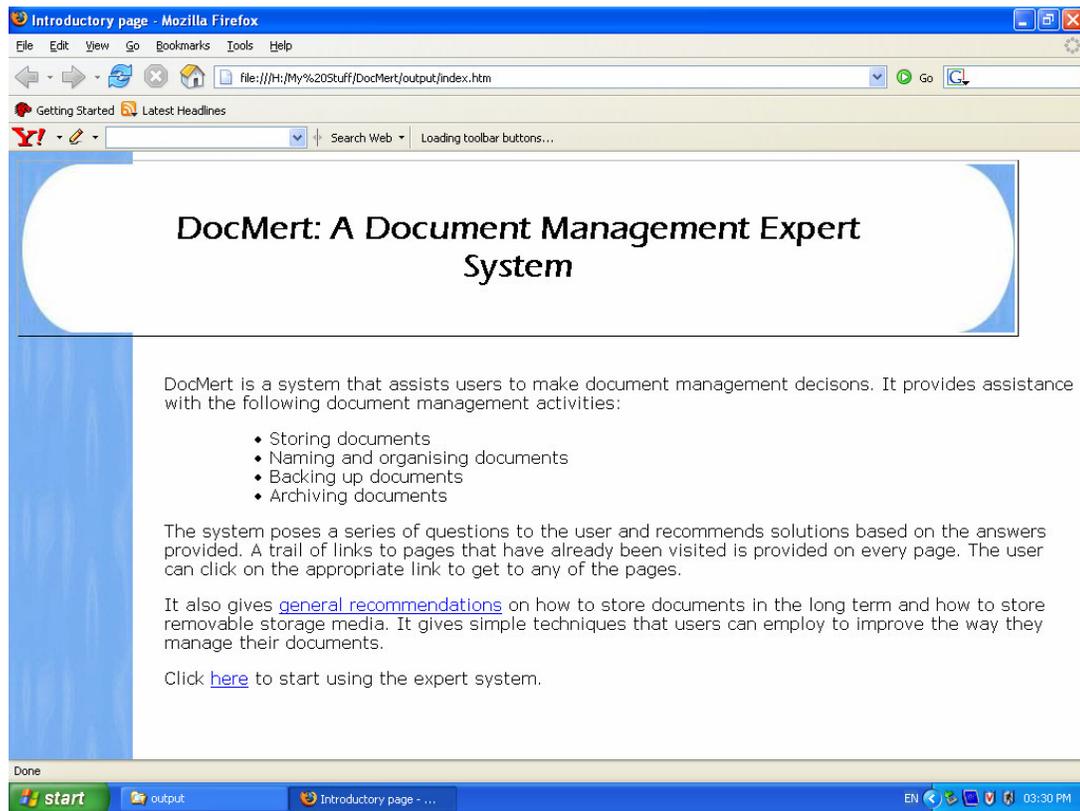
Figure 2: Home Page

The Home page (see Figure 2) is the first page of the system. It gives an introduction to the system and explains what it does. It has links to the general recommendations page and the first question page of the system.
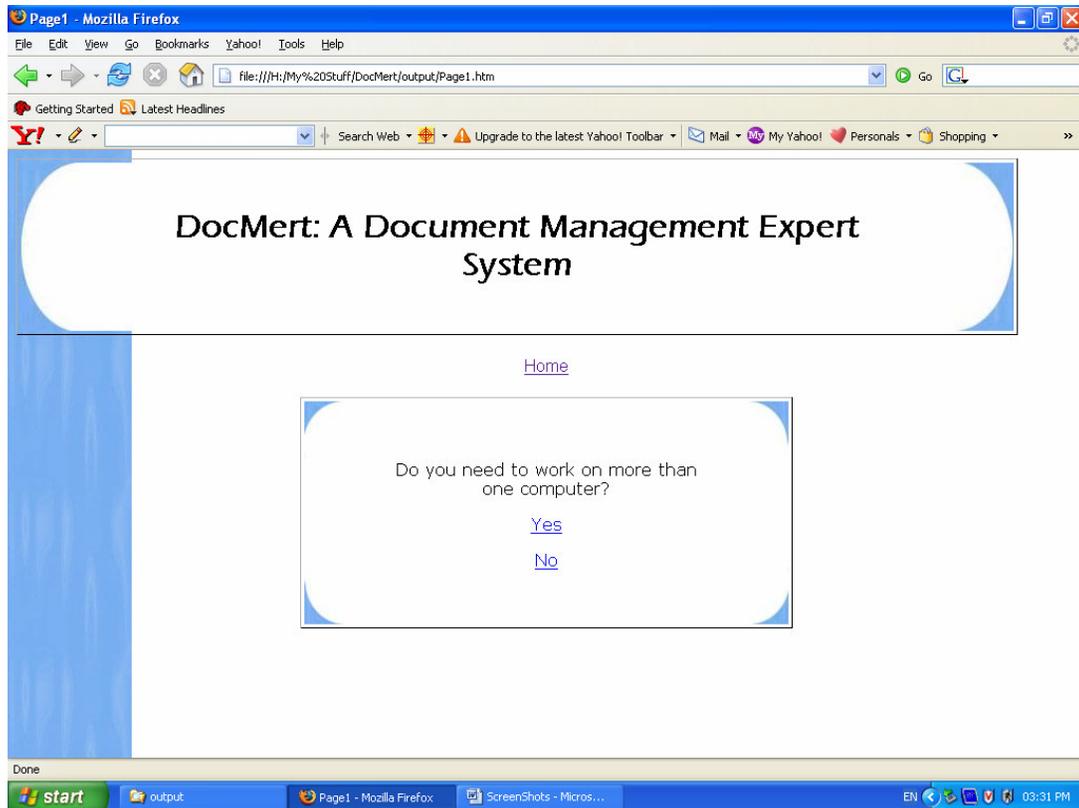
Figure 3: Question page

Figure 3 illustrates the first question of the system. It has a breadcrumb showing the page that was visited previously, in this case the "Home" page. The first question is meant to establish whether a user works on a single or multiple computers. Depending on the response selected, the user is taken to other pages containing questions relevant to the number of computers they use. These questions eventually lead to a page containing solutions specific to the user's problem.
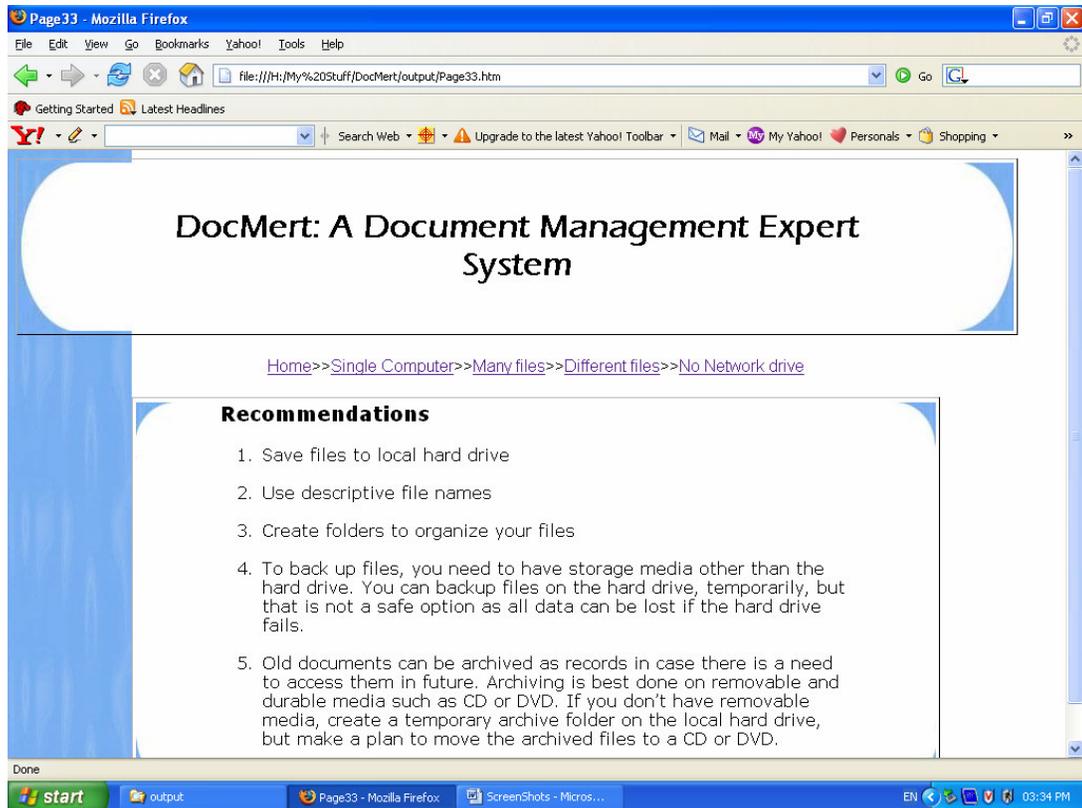
Figure 4: Solutions page

Figure 4 illustrates the page that contains solutions specific to the user's problem. It gives advice on the following aspects of document management:

- storage media to use
- file naming and organisation
- document backup and media to use
- document archiving and media to use

It has a breadcrumb trail of all the pages visited prior to the solutions page. The user can backtrack to any of these pages by clicking on the relevant link.
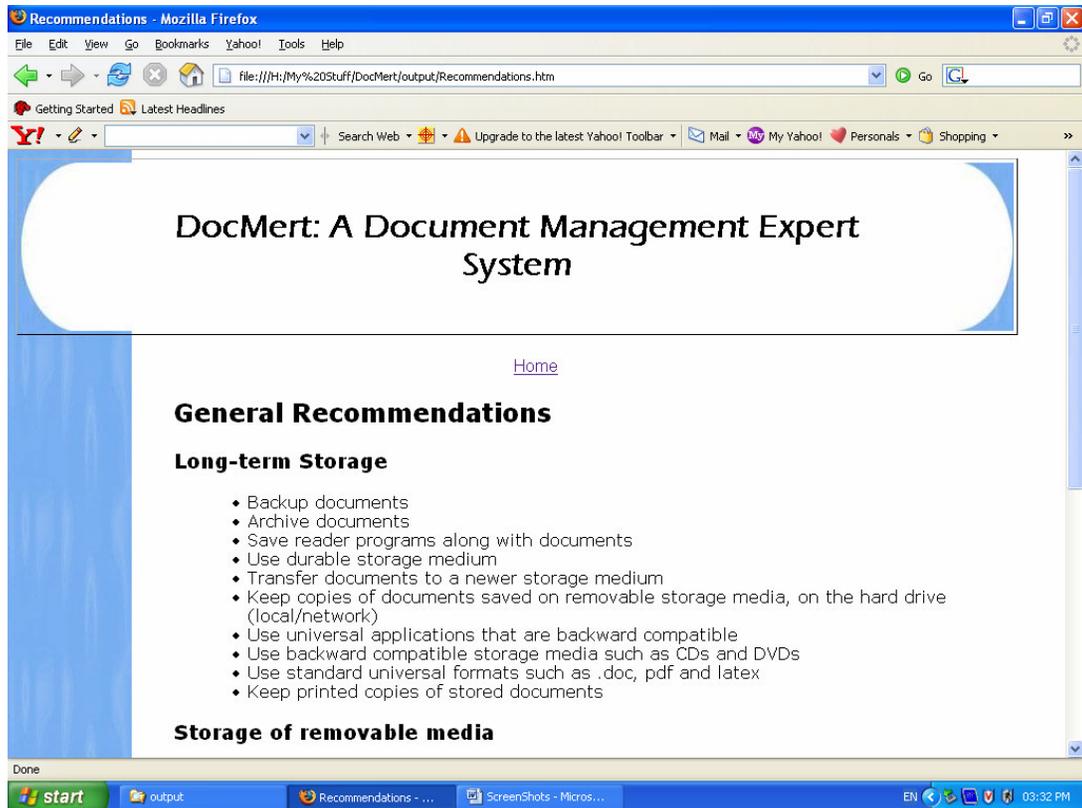
Figure 5: General recommendations page

Figure 5, the general recommendations page, contains general recommendations on document management. It outlines retrieval strategies for long-term document storage and other general document management techniques. It links with the "Home" page only and it has a breadcrumb showing the "Home" page as the page previously visited.

5.5   Concluding remarks

The system has 37 pages in total for the set of derived rules: 17 question pages, 18 solutions pages, the Home page and the General Recommendations page. Generating HTML pages simultaneously greatly lessened the work of having to code each page individually. Another advantage of generating pages this way was that editing could be done in one place and be effected to all the pages at once.

Chapter 6 - Expert system evaluation and results

6.1 Introduction

After the expert system was developed, it was tested on users to evaluate the quality of its recommendations and its usability. This chapter outlines the evaluation process and the results attained. Firstly, it outlines the preparations undertaken for user testing. An outline of how the user tests were carried out is then given, followed by the presentation and discussion of evaluation results.

6.2 Preparations for system testing

Before the system could be tested with a large sample of users, a pilot test was carried out with three users. One user had been a respondent in the data collection stage and the other two had not. The three users were asked to use the system and comment on it. After using the system, the users pointed out aspects of the system that needed to be improved. The following are areas that were pointed out:

- Some questions were unclear.
- The recommendations page needed a heading.
- The Home page did not sufficiently explain what the system does.
- A "top of page" link was required at the bottom of the Recommendations page.

The recommended improvements were made before testing the system on a larger sample of users. The questions that were pointed out were rephrased to make them clear. The recommendations page was given a heading. More information was added on the Home page to explain what the system does and how it operates and a "top of page" link was added at the bottom of the Recommendations page.

An evaluation questionnaire was designed (see Appendix D). The questionnaire was meant to collect information about how the users felt about the system. The questions covered the following aspects of the system:

- the adequacy of information presented on the Home page,

- the clarity of questions posed by the system,
- the quality of recommended solutions,
- the ease of using the system and
- general comments about the system.

6.3 System testing

A sample of 14 users was selected randomly from the UCT community, for system evaluation. The sample was made up of seven people who had been interviewed during data collection and another seven who had not been interviewed. Respondents were selected on the basis of availability for user evaluation. Some of the people approached could not make the time arranged for evaluations and they were substituted by others who could. Users were asked to use the system and fill in an evaluation questionnaire after using the system.

Before users could start using the system they were given a brief verbal overview of what the system does. The Home page of the system was used as a starting point and users were requested to read it for guidelines on how to use the system. Users were observed as they used the system. After using the system, all the users filled out an evaluation questionnaire.

6.4 Presentation of results

The following tables show a summary of responses of users to the evaluation questions. The responses have been split among three tables depending on the type of question asked. The first table (Table 3) shows responses to Likert-scale statements. The second table and the third table (Table 4 and 5) show responses to multiple-choice questions. The responses have been split over two tables because the response options offered are not similar. General users' comments are included below the tables.

Table 3: Likert scale responses

Statement	Responses				
	Strongly disagree	Disagree	Undecided	Agree	Strongly agree
The first page sufficiently describes what the system does.	0	0	0	8	6
The questions posed by the system are easy to understand.	0	0	0	4	10
The solutions recommended by the system are reasonable.	0	0	1	10	3
The system is easy to use.	0	0	0	5	9

Table 4: Multiple-choice responses

Question	Responses		
	None	Some	All
Do you currently apply any of the recommended solutions to manage your documents?	0	14	0
Would you apply any of the recommended solutions to manage your documents?	0	9	5

Table 5: Multiple-choice responses

Question	Responses		
	No	Some	Yes
Given the response you provided, did you expect the solutions recommended?	0	5	9

General Comments

- *The recommendations seem simple, elegant and efficient to implement. The document/file structure is effective and intuitive to follow.*

- *The responses provided could be more detailed. A link could be provided that users can follow to get detailed responses. For instance, details about how to backup on CD or DVD, in case users are not familiar with these technologies.*

- *Overall, really nice system. Consider your end-point though and make it clear that these are system recommendations. Maybe, also provide a "Start again" function.*

- *The system guides well and gives descriptions very well. I would recommend that more detailed explanations behind the processes be made, especially for people who do not know much about document management.*

- *The system seems clear. It gives examples when there is uncertainty, which I like. I like the way you can see your path and go back to any step to change information. Makes the process clear.*

- *Include back and forth navigation arrows. Bigger font on the question boxes. There is lots of space.*

- *It was unclear at the start whether the system would make recommendations or actually manage a set of documents. (Just a bit ambiguous). This could make a good online reference.*

- *The system would be good for educational purposes. It could be helpful to people new to computers.*

- *Some of the recommendations seem like things the user would already apply, e.g., "Use descriptive file names". I'm not sure if the recommendation will be helpful. It may seem "obvious" for a user to already do so without the system to tell them.*

- *The system is missing a type where a user may have more than one computer, but not connected to the same network.*

- *The system seems to be usable by all irrespective of computer literacy level.*

- *Very user-friendly, but it also might help to clarify at the start (Do you need to work on more than one computer?) whether the question is referring to at-home or at-work use. Also would it make a difference if there were multiple users accessing the same files? Maybe describe "network drive". The system would be helpful to people taking introductory computer classes.*

- *Perhaps an "I'm not sure" option could be added, so that more explanation can be given if needed. There are recommendations under general, which might also apply to specific recommendations. These could be linked to the related recommendation under general, e.g., when recommending archiving should be done with CD/DVD.*

6.5 Discussion of results

During user testing, some users took time to read the contents of the Home page. Others just rushed through it and came back to it if there was something they wanted to check. Overall the users were able to get around the system without assistance.

From the observations of users while they were using the system and the responses they gave to evaluation questions, it can be concluded that the system is easy to use. When users seemed to be at a loss as to what the next step was, they were able to use the links provided on the breadcrumb trail to go back to previous pages and start afresh.

Users made comments about how the recommendations and navigation could be improved. Some comments were about what the system could be used for. Comments about the usability of the system were also given. Some users thought the system could be useful as a form of reference on document management. One user thought it could be a good online reference and two others thought it could be helpful to people taking introductory computer courses. On the other hand, there were users who thought the system needed to be more detailed for people who may not be familiar with document management and its technologies. One user, however, thought that some of the recommendations were obvious and did not need to be made.

Generally, users agreed that the recommendations made by the system were reasonable. They indicated that they expected the recommendations given and were already applying some of them. They were also willing to apply some of the recommended solutions that they had not been using before.

6.6 Conclusions

From the evaluation the following can be concluded:
- The system is easy to use as users managed to navigate around it without any help from the researcher.
- The system makes reasonable recommendations as both the interviewed respondents and those who had not been interviewed agreed with them.

- As the techniques recommended by the system came from expert users, the system embodies good document management practices that can be tailored and adopted by non-expert users.

Chapter 7 – Conclusions and Recommendations

7.1 Conclusions

The aim of the project was to study users' document management techniques and use them to build an expert system that recommends best practices in managing documents. From the initial interviews conducted, a number of useful techniques that users employed for managing their documents were discovered. These techniques formed the basis of an expert system that users subsequently found to be useful during system evaluation. Users who tested the system came across techniques that they were already using and others that were new to them.

From the study it seems that there are similarities in the way users manage their documents. One area that demonstrates this is file naming and organisation. All the users interviewed indicated that they named their files using descriptive names and created folders for similar files. While there were some similarities in the way users managed their documents, some users had in place techniques that were more sophisticated than those employed by others. This seemed to be the case for users who dealt with a large number of documents.

The solutions recommended by the systems were found to be equally reasonable to those users who had been interviewed during data collection and those who had not been interviewed. All the users who took part in this evaluation indicated that they would use some or all of the solutions recommended by the system.

Document management seems to play a big role in the activities of computer users. This is evident in the techniques that were uncovered by the study. Users do take time to think of strategies that can make retrieval of their stored information easy. These strategies can be quite helpful to people new to computers.

7.2 Future Work

1. System Recommendations

As recommended during user evaluation, the system could be extended to include detailed explanations about how to implement the recommended solutions. This would be helpful to users not familiar with document management activities and technologies. A "how" link could be added next to each recommendation. Users who do not need detailed explanations could ignore the link and implement the recommended solution. A link to the general recommendations page could also be included in each of the solutions pages.

2. Document managing functionality

Functionality could be added to the system to manage documents for users. In addition to giving recommendations, the added functionality would enable users to use the system to implement the recommended solutions. This way, users would not have to switch to a different workspace to implement the recommended solutions.

3. System deployment

The system could be used to aid users new to computers in managing their documents. It could be used in introductory computer classes or be placed online for users to access at their convenience.

Appendices

Appendix A

Data collection questionnaire

Individual Document Management Techniques: an explorative study

This research is conducted as a requirement by the University of Cape Town for a Masters degree in Information Technology. The aim of the research is to develop an expert system that will help users make decisions about document storage, backup, archiving and retrieval. The system will be developed on the basis of responses given by the respondents to the questionnaire. Please take some time out of your busy schedule to answer the following questions.

1. What is your occupation?
    _____
    _____

2. If you are a student, please indicate your field and year of study.
    _____
    _____
    _____

3. Do you use computers to carry out your work?

    ☐ Yes   ☐ No   ☐ Sometimes

4. If you do use computers, how long have you used them?
    _____
    _____

5. How often do you type or edit documents on a computer?
    _____
    _____
    _____
    _____

6. Which of the following media do you use to store your documents? *(Please tick the appropriate boxes.)*

    ☐ CD-ROM

☐ Magnetic Tapes

☐ Flash Disk

☐ Hard Disk

☐ DVD

☐ Server hard drive

☐ Network Attached Storage (NAS)

☐ Storage Area Network (SAN)

☐ Other [_____] (please fill it in)

7. Why do you use the storage media that you have indicated in the previous question? Please give reasons for each of the media you have indicated.

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

8. How do you name and organise your files?

_____
_____
_____

9. Why do you use this particular criterion?

_____
_____
_____
_____
_____

10. Do you ever backup your documents? (*Save copies of documents on a different media or location, as a way of ensuring that the data can be recovered if the original file is lost or destroyed*)

☐ Yes ☐ No ☐ Sometimes

51

11. How do you decide on what documents to back up?

_____
_____
_____
_____

12. On what medium do you store the documents that you back-up?

_____
_____
_____

13. Do you ever archive your documents? (*Archiving refers to storing, as records, data that are not accessed regularly*).

☐ Yes ☐ No ☐ Sometimes

14. If your answer is "Yes" or "Sometimes", how do you decide on what documents to archive?

_____
_____
_____
_____

15. On what medium do you store the documents that you archive?

_____
_____
_____

16. In terms of years, what do you consider to be long-term storage of documents?

_____
_____
_____

17. How do you store documents in the long-term?

_____
_____
_____
_____

18. Are there any specific reasons why you store them that way? Please elaborate.

_____
_____

_____
_____

19. How do you ensure that the documents you store are easy to retrieve in the long-term?

_____
_____
_____
_____

20. How & where do you store removable storage media containing your documents, to ensure that they are accessible in the long-term?

_____
_____
_____
_____

21. How do you ensure that the format of your long-term documents is up-to-date (not obsolete)?

_____
_____
_____
_____

22. How do ensure that your long-term documents are not stored on obsolete media?

_____
_____
_____
_____

23. Please explain any other techniques that you have devised to help you manage your documents better.

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

Thank you for taking time to answer my questions.

Appendix B

Responses to the questionnaire

B1.   File naming and organisation

   i)   File naming

- Meaningful names that are easy to remember
- According to content
- By date
- According to function
- Content and date of creation, e.g. AdminReport_05_09_2006
- Use title as file name
- Descriptive name
- Obscure names for security

   ii)   File organisation

- Create folders for similar files
- Create folders in "My Documents" folder
- Save frequently used files on the desktop
- Create folders by year and category
- Logical directory structure
- Keep files in different formats and assign numbers according to format
- Hierarchical file structure
- Create folders according to function e.g. private, school-work
- Store some documents on the desktop for quick access

   iii)   Reasons for file naming and organisation

- For easy retrieval
- For fast retrieval
- For easy access
- Logical hierarchical organisation

- To differentiate between files according to function
- To differentiate between current and past work
- To recognise versions easily
- Easy to back-up
- Quick and easy
- Inherited from previous office incumbent and efficient
- To know contents without opening folder
- Habit
- Security

B2. Criteria for backing up documents

- Important files
- Files used regularly
- Files that are not easy to recreate e.g. thesis, assignments etc.
- To have more than one copy of documents (at home and at work)
- All the contents of "My Documents" folder
- All personal files
- Simulations
- Teaching material
- Documents that may have legal implications
- Templates e.g. letters to students
- All documents

B3. Criteria for archiving documents

- Stuff that hasn't been used in three months (since creation)
- Important emails
- Research data
- Zipped files
- Publications
- All documents (from early 90s)
- Ad hoc
- Document lifespan
- Old documents that may be required in future

- Computer programs
- All the contents of previous computers

Appendix C

Solutions

Solution 1 – Page 7 – (Multiple computers, have network drive.)

1. Save files to the network drive
2. Use descriptive file names
3. Save files on the network drive or removable media as back-up
4. Old documents can be archived as records in case there is a need to access them in future. Archiving is best done on removable and durable media such as CD or DVD.

Solution 2 – Page 8

1. Save files to the network drive
2. Use structured file names. For example, assignt1_CS_07, assignt1_ENG_07 etc.
3. Create folders to organize your file.
4. Save files on removable media as back-up
5. Old documents can be archived as records in case there is a need to access them in future. Archiving is best done on removable and durable media such as CD or DVD.

Solution 3 - Page 9

1. Save files to the network drive
2. Use descriptive file names
3. Create folders to organize your file
4. Save files on removable media as back-up
5. Old documents can be archived as records in case there is a need to access them in future. Archiving is best done on removable and durable media such as CD or DVD.

Solution 4 – Page 13 (Multiple computers, have no network drive, have removable media)

1. Save files to removable media
2. Use descriptive file names

3. Save copies of files on separate removable media or hard drive, as backup.

4. Old documents can be archived as records in case there is a need to access them in future. Archiving is best done on removable and durable media such as CD or DVD.

Solution 5 – Page 14

1. Save files to removable media

2. Use structured file names. For example, assignt1_CS_07, assignt1_ENG_07 etc.

3. Create folders to organize your files

4. Save copies of files on separate removable media or hard drive, as backup.

5. Old documents can be archived as records in case there is a need to access them in future. Archiving is best done on removable and durable media such as CD or DVD.

Solution 6 – Page 15

1. Save files to removable media

2. Use descriptive file names

3. Create folders to organize your files

4. Save copies of files on separate removable media or hard drive, as backup.

5. Old documents can be archived as records in case there is a need to access them in future. Archiving is best done on removable and durable media such as CD or DVD.

Solution 7 – Page 19 (Multiple computers, have no network drive, have CD/DVD Writer)

1. Save files to local hard drive

2. Use descriptive file names

3. Save copies of files on separate removable media or hard drive, as backup.

4. Old documents can be archived as records in case there is a need to access them in future. Archiving is best done on removable and durable media such as CD or DVD.

Solution 8 – Page 20

1. Save files to hard drive
2. Use structured file names. For example, assignt1_CS_07, assignt1_ENG_07 etc.
3. Create folders to organize your files
4. Save copies of files on separate removable media or hard drive, as backup.
5. Old documents can be archived as records in case there is a need to access them in future. Archiving is best done on removable and durable media such as CD or DVD.

Solution 9 – Page 21

1. Save files to hard drive
2. Use descriptive file names
3. Create folders to organize your files
4. Save copies of files on separate removable media or hard drive, as backup.
5. Old documents can be archived as records in case there is a need to access them in future. Archiving is best done on removable and durable media such as CD or DVD.

Solution 10 – Page 23 - (Muiltiple computers, no network drive, no removable media, no CD/DVD Writer.)

1. Save files to local hard drive
2. Use descriptive file names
3. To create a back up of files, you need to have other storage media other than the hard drive. You can backup on the hard drive temporarily, but that is not a safe option as all data can be lost if the hard drive fails.
4. Old documents can be archived as records in case there is a need to access them in future. Archiving is best done on removable and durable media such as CD or DVD.

Solution 11 – Page 24

1. Save files to local hard drive
2. Use structured file names. For example, assignt1_CS_07, assignt1_ENG_07 etc.
3. Create folders to organize your files

4. To create a back up of files, you need to have other storage media other than the hard drive. You can back up files on the hard temporarily, but that is not a safe option as all data can be lost if the hard drive fails.

5. Old documents can be archived as records in case there is a need to access them in future. Archiving is best done on removable and durable media such as CD or DVD.

Solution 12 – Page 25

1. Save files to local hard drive

2. Use descriptive file names

3. Create folders to organize your files

4. To create a back up of files, you need to have other storage media other than the hard drive. You can back up files on the hard drive temporarily, but that is not a safe option as all data can be lost if the hard drive fails.

5. Old documents can be archived as records in case there is a need to access them in future. Archiving is best done on removable and durable media such as CD or DVD.

Solution 13 – Page 30 - (Single computer, have network drive)

1. Save files to local hard drive

2. Use structured file names. For example, assignt1_CS_07, assignt1_ENG_07 etc.

3. Use the network drive to back up your files.

4. Old documents can be archived as records in case there is a need to access them in future. Archiving is best done on removable and durable media such as CD or DVD. If you don't have removable media, create a temporary archive folder on the network server or the local hard drive, but make a plan to move the archived files to a CD or DVD.

Solution 14 – Page 31

1. Save files to local hard drive

2. Use structured file names. For example, assignt1_CS_07, assignt1_ENG_07 etc.

3. To back up files, you need to have other storage media other than the hard drive. You can back up files on the hard drive temporarily, but that is not a safe option as all data can be lost if the hard drive fails.

4. Old documents can be archived as records in case there is a need to access them in future. Archiving is best done on removable and durable media such as CD or DVD. If you don't have removable media, create a temporary archive folder on the local hard drive, but make a plan to move the archived files to a CD or DVD.

Solution 15 – Page 32

1. Save files to local hard drive
2. Use descriptive file names
3. Create folders to organize your files
4. Use the network drive to back up your files.
5. Old documents can be archived as records in case there is a need to access them in future. Archiving is best done on removable and durable media such as CD or DVD. If you don't have removable media, create a temporary archive folder on the local hard drive, but make a plan to move the archived files to a CD or DVD.

Solution 16 – Page 33

1. Save files to local hard drive
2. Use descriptive file names
3. Create folders to organize your files
4. To back up files, you need to have other storage media other than the hard drive. You can temporarily back up files on the same hard drive, but that is not a safe option as all data can be lost if the hard drive fails.
5. Old documents can be archived as records in case there is a need to access them in future. Archiving is best done on removable and durable media such as CD or DVD. If you don't have removable media, create a temporary archive folder on the local hard drive, but make a plan to move the archived files to a CD or DVD.

Solution 17 – Page 34

1. Save files to local hard drive
2. Use descriptive file names

3. Use the network drive to back up your files.

4. Old documents can be archived as records in case there is a need to access them in future. Archiving is best done on removable and durable media such as CD or DVD. If you don't have removable media, create a temporary archive folder on the local hard drive, but make a plan to move the archived files to a CD or DVD.

Solution 18 – Page 35

1. Save files to local hard drive

2. Use descriptive file names

3. To back up files, you need to have other storage media other than the hard drive. You can back up files on the hard drive, temporarily, but that is not a safe option as all data can be lost if the hard drive fails.

4. Old documents can be archived as records in case there is a need to access them in future. Archiving is best done on removable and durable media such as CD or DVD. If you don't have removable media, create a temporary archive folder on the local hard drive, but make a plan to move the archived files to a CD or DVD.

Appendix D

Evaluation questionnaire

Evaluation of DocMert: an expert system for document management

DocMert is an expert system for managing documents. It poses a series of questions to the user and recommends document management solutions, based on the answers given by the user.

For each of the statements below, please indicate the extent of your agreement or disagreement by circling the relevant response.

a) The first page sufficiently describes what the system does.
1. Strongly disagree
2. Disagree
3. Undecided
4. Agree
5. Strongly agree

b) The questions posed by the system are easy to understand.
1. Strongly disagree
2. Disagree
3. Undecided
4. Agree
5. Strongly agree

c) The solutions recommended by the system are reasonable
1. Strongly disagree
2. Disagree
3. Undecided
4. Agree
5. Strongly agree

d) Given the responses you provided, did you expect the solutions that were recommended?

    1. No

    2. Some

    3. Yes

e) Do you currently apply any of the recommended solutions to manage your documents?

    1. None

    2. Some

    3. All

f) Would you apply any of the recommended solutions to manage your documents?

    1. None

    2. Some

    3. All

g) The system is easy to use.

    1. Strongly disagree

    2. Disagree

    3. Undecided

    4. Agree

    5. Strongly agree

h) Please give any general comments you may have about the system:

_____

_____

_____

_____

_____

_____

Thank you for taking time to evaluate the system!

References

1. Barreau D. K. & Nardi B. (1995). *"Finding and Reminding: File Organization From the desktop"*. SIGCHI Bulletin. July 1995. Vol. 27. No. 3.

2. Babbie E. & Mouton J. (1998). *The practice of Social Research.* Oxford: Oxford University Press.

3. Boardman R., Sasse M.A. (2004). *"Stuff Goes into the Computer and Doesn't Come Out" A Cross-tool Study of Personal Information Management*. CHI 2004, April 24–29, 2004, Vienna, Austria.

4. Castillo, E. & Alvarez E. (1991). *Expert Systems: Uncertainty and Learning*. Southampton: Computational Mechanics Publications.

5. Celentano A., Pozzi S. & Toppeta D. (1992). *A Multiple Presentation Document Management System.* Italy: Politecnico di Milano. In Proceedings of the 10th Annual Conference on Systems Documentation, 1992. ACM SIGDOC.
Available: http://delivery.acm.org/10.1145/150000/147009/p63-celentano.pdf [15 March, 2006].

6. *Conversion Masters in Information Technology - Course Material.* [CD-ROM] February 2006. Department of Computer Science, University of Cape Town.

7. Davis R. & Buchanan B. G. (1977). *Meta level knowledge. Overview and applications*. In proceedings of the 5th International Joint Conference on Artificial Intelligence. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence.

8. Davis R., Buchanan B. & Shortliffe E. (1977). *Production rules as a representation for knowledge-based consultation systems.* Artificial Intelligence Journal, Vol 8 (1).

9. Dey I. (1993). *Qualitative Analysis.* London: Routledge.

10. Dourish P., Edwards K., LaMarca A., Lamping J., Petersen K., Salisbury M, Terry D.B. & Thornton J. (2000). *Extending Document Management Systems with User-Specific Active Properties.* ACM Transactions on Information Systems (TOIS). Volume 18, Issue 2 (April 2000).
Available: http://portal.acm.org/citation.cfm?id=348758 [15 March, 2006].

11. Engelmore R.S. & Feigenbaum E. (1993). *Expert Systems and Artificial Intelligence*. Available: http://www.wtec.org/loyola/kb/c1_sl.htm [16 March, 2006].

12. Expert System – Wikipedia. Available: http://en.wikipedia/wiki/Expert_System [11 April, 2006]

13. Feigenbaum E.A. (1982). *Knowledge Engineering in the 1980s*. Stanford: Department of Computer Science, Stanford University, Stanford CA.

14. Freeman E. & Gelernter D. (1996). *Lifestream: A storage Model for Personal Data.* New Haven: Yale University. ACM SIGMOD Record. Volume 25, Issue 1 (March 1996). Available:
http://delivery.acm.org/10.1145/390000/381893/P080.pdf [15 March, 2006].

15. Giarrantano J. & Riley G. (1989). *Expert Systems: Principles and Programming*. Boston: PWS-KENT Publishing Company.

16. Harmon P. & King D. (1985). *Representing knowledge.* I.N.C.A. Willey Press Book.

17. Hayes-Roth F., Waterman A.D. & Lenat D.B. (Editors). (1983). *Building Expert Systems.* Reading, Massachusetts: Addison-Wesley Publishing Company, Inc.

18. Henderson S. (2004). *How Do People Organize Their Desktops?*
CHI 2004, April 24–29, 2004, Vienna, Austria. ACM Press.

19. Jones W. (2005). *Personal Information Management.* Washington: University of Washington.

20. Lansdale, M. *The psychology of personal information management*. Applied Ergonomics, 19 (1).

21. Malone T. W. (1983). *How do People Organize Their Desks? Implications for the Design of Office Information Systems*. ACM Transactions on Office Information Systems, 1(1).

22. Mander R., Salomon G. & Wong Y.Y. (1992) *A 'Pile' Metaphor for Supporting Casual Organization of Information.* Human Interface Group, Advanced Technology Apple Computer, Inc. In Proceedings of CHI 1992 (May 3-7). ACM Press.

23. Marchand D.A., Davenport H., & Dickson T. (Editors). (2000). *Mastering Information Management*. London: Prentice Hall.

24. Marsden G. & Cairns D (2003). *Improving the Usability of the Hierarchical File System.* In Proceedings of SAICSIT 2003, ACM Press
http://delivery.acm.org/10.1145/960000/954027/p122-marsden.pdf   [11   July, 2006].

25. Meier J., & Spraque R. (1996). *Towards a Better Understanding of Electronic Document Management.* Proceeding of the 29th Annual Hawaii International Conference on System Sciences – 1996. Available: http://csdl12.computer.org/comp/proceedings/hicss/1996/7426/00/74260053.pdf   [16 March, 2007]

26. Nebendhal D. (1988). *Expert Systems: Introduction to the Technology and Applications.* Siemens Aktiengesellschaft: John Wiley & Sons Limited.

27. Payne E.C. & McArthur R.C. (1990). *Developing Expert Systems: A Knowledge Engineers Handbook for Rules and Objects.* New York: John Wiley & Sons Limited.

28. Sprague R.H. & McNurlin B.C. (Editors). (1993). *Information System Management in Practice.* Englewood Cliffs, New Jersey: Prentice Hall.

29. Ramike L., Mukunyandela K. & Suleman H. (2004). *Integration of Publishing and Document Management into Windows.* Available: http://pubs.cs.uct.ac.za/archive/00000178/01/research_document2.pdf [15 March, 2006]

30. Raviso P. (2004). *Personal Information Organisation: Studies on User-Appropriate Classification and Retrieval Strategies and their Implications for Information Management Systems Design.* Zurich: Swiss Federal Institute of Technology.

31. Riley G. (2004). *What is CLIPS?* Available: http://www.ghg.net/clips/WhatIsCLIPS.html [4 February, 2007]

32. Roberstson G., Czerwinski M., Larson K., Robbins D.C., Thiel D., & van Dantzich M. (1998). *Data Mountain: Using Spatial Memory for Document Management.* Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA.
Available: http://www.microsoft.com/usability/UEPostings/p153-robertson.pdf [16 March, 2006]

33. Woods W.A. (1975). *What's in a link: Foundations for semantic networks in representation and understanding.* Bobrow and A. Collins ed. London: Academic Press.