Genre, Task, Topic and Time: Facets of Personal Digital Document Management

Sarah Henderson
Department of Information Systems and Operations Management
University of Auckland
Private Bag 92019 Auckland
s.henderson@auckland.ac.nz

ABSTRACT

Most operating systems provide the ability to create folders to contain documents, and to nest these folders to create a hierarchical organization. However, little is known about the kinds of folders people create using this type of organizing scheme, or how they structure those folders.

Exploratory research was conducted, analyzing the folder structures of six knowledge workers and it was found that most folder names represent the genre, task, topic or time dimension of the documents they contained. While these four dimensions were consistent across all participants, the order in which these dimensions are combined into a hierarchical structure varies substantially, even among people with the same job.

A number of interesting areas of investigation are highlighted for future research, including a proposal that these dimensions be treated as facets of document metadata and that exploring faceted navigation interfaces for personal digital document management would be a fruitful area for further research.

Categories and Subject Descriptors

H.3.2 Information Storage and Retrieval: Information Storage. H5.2 Information interfaces and presentation: User Interfaces.

General Terms

Human Factors, Design.

Keywords

Personal digital document management, personal information management, document organization

1. INTRODUCTION

Personal digital document management is the process of acquiring, storing, managing, retrieving and using digital documents. It is personal in the sense that the documents are owned by the user and are under their direct control, not that they necessarily contain information about the user [1]. Information overload is making document management increasingly difficult. Farhoomand and Drury found that the two most common definitions of information overload were "an excessive volume of information" (reported by 79% of respondents) and "difficulty or impossibility of managing it" (reported by 62%) [2].

One large part of managing documents involves organizing them so that they can later be easily retrieved. The software that many people currently use to manage their documents is the Windows Explorer file management utility that is included with Microsoft Windows (or a similar product on other platforms). These tools allow people to create folders and place their documents within folders. Using this simple containment mechanism, people can build up a large hierarchical structure of folders. This concept of folders evolved from early file systems developed over 40 years ago.

There is anecdotal evidence to suggest that people have trouble locating their files after having stored them in these hierarchical structures. Although much research has started with the premise that existing systems are flawed, very little research has been conducted on how people actually use the hierarchical structure provided by Windows Explorer and similar systems. Uncovering the actual problems with the existing systems, and understanding the organizing patterns people use could help to identify requirements that Windows Explorer doesn't meet, identify how the software could be improved, or even suggest the design of new software to take its place.

Understanding how the current hierarchical model supports users in organizing documents, and more crucially, where it doesn't, is important to being able to develop more usable systems that better support personal document management.

2. PREVIOUS RESEARCH

Many researchers have created experimental prototypes to explore alternative systems of organizing information. Most of these are based around a particular dimension of the information that is assumed to be primary. For instance, Lifestreams [3] is based on the premise that the most important dimension on which to organize things is time. TimeScape [4] also includes time as a primary dimension, but includes a spatial layout as well. Bellotti et al's Taskmaster system [5] is based on studies of email users that found that task or project is a common organizing principle. The Windows XP Start menu groups open documents according to the application used to edit them (essentially grouping by the file format of the document).

CHINZ '05, July 6-8, 2005 Auckland, New Zealand

At the other end of the spectrum are systems like the Google Desktop (desktop.google.com), and Copernic Desktop Search (www.copernic.com), which theoretically make organizing dimensions unnecessary, since they can locate documents by means of full text searching. Along a similar vein, the Placeless Documents project [6] doesn't impose any type of structured organizing scheme at all, but allows the user to give attributes and attribute values to documents, which can then be used to search and group documents for viewing.

Although the full text searching is attractive, it is not a full replacement for organizing document, since browsing through an organized collection of documents gives you an overview of what is available, as well as the ability to see how different items are related to each other. This doesn't happen with a full text search system, which can only retrieve documents matching a query you formulate explicitly. Sometimes users don't need to retrieve a specific document, but just to 'see what information I've got related to X' [7].

Another issue with many of these proposed systems (such as Placeless Documents), is that they require the user to supply metadata about the document. As encountered in many knowledge management initiatives it is difficult to get users to enter metadata about their documents [8]. Users are busy getting on with their work, and aren't really concerned about managing their documents beyond the minimum required to ensure the document doesn't disappear into oblivion.

Whether the user is required to enter metadata about the documents, or the system tries to generate or intuit that information itself, it is still necessary to know exactly what information is required about a document. What are the attributes that are important about documents, and what are the important dimensions that people use to classify them?

Kwasnik conducted a study investigating the dimensions people use when they talk about their physical documents in their offices. She found 35 dimensions, which could be grouped into seven broad groups [9 p.208]:

> "Situation Attributes, such as source, use, circumstance, and access; Document Attributes, such as author, topic, and form; Disposition, such as discard, keep, postpone; Order/Scheme, such as group, separate, and arrange; Time, such as continuation, duration, and currency; Value, such as importance, interest, and confidentiality; and Cognitive State, such as "don't know," and "want to remember."

A similar study on digital documents found that the dimensions most commonly used in narratives about documents were: Time, Place, Co-Author, Purpose, Subject, Other Documents, Format, Exchanges, Tasks, Storage and Contents [10 p.248].

Setting up a hierarchy of folders is essentially equivalent to defining a set of attributes or keywords that can be applied to a document. For example, consider what it means to place documents into the Lectures folder in the structure shown in Figure 1 below:

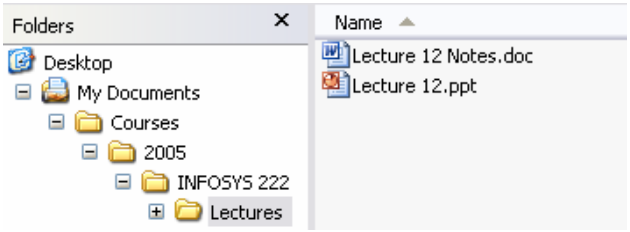

Figure 1. Example Folder Structure

By simply placing the document in this folder, the user is saying that this document is related to their Courses, that it is applicable for 2005, that it pertains to the course INFOSYS 222, and that the document is to do with lectures. These four pieces of information can be assigned to any document with the single action of placing it inside that folder. The folder names provide the context, and the file names and file formats serve to distinguish among documents within that context.

Thus, the folder names that are used, and depth of the folder structure tell us something about the type and quantity of metadata that must be used in order to differentiate documents for use by a particular person.

No research has yet been done that has looked at the type of folders people create and how they organize them into hierarchies. This research intends to fill that gap by examining folder naming practices and exploring how the folders are composed into hierarchies.

3. METHOD

This research was exploratory work conducted as part of a larger study aimed at understanding how knowledge workers manage their personal digital documents. The aim of this part of the study is to understand how people name their folders, and how people structure their folders into a hierarchy, in order to see what type of software support for folder management could be provided.

If there are commonalities in the types of folders people create, and the way they structure them, then there is scope for software to understand and support these processes. However, if folder naming is truly idiosyncratic, then there is less scope for automated support of folder creation and management.

In this part of the study, six knowledge workers were interviewed in depth about their document management practices, and a snapshot of their file system was taken (using custom-written software). These knowledge workers were all employed in a University setting, and held positions including Software Developer, Course Manager and Lecturer.

Table 1 shows the position help by each participant, and the total number of folders, and number of unique folder names that each participant had in their file system.

The list of the folder names each participant has used was extracted from the file system snapshot information. The number of unique folder names was lower than the total number of folders, sometimes substantially. This was largely due to repetition of the same folder names in different places in the hierarchy, and in some cases due to wholesale duplication of folders structures.

Table 1. Participant Summary

Participant	Position	Total Folders	Unique Names
A	Software Developer	196	172
B	Course Manager	4123	376
C	Course Manager	745	403
D	Course Manager	211	131
E	Lecturer	419	279
F	Lecturer	575	167

Many of the folders were system generated, with the names assigned by the Operating System, or the software that created them. These include the default folders Desktop, My Documents, My Pictures and My Music etc. Many others were created when HTML files are saved, since a folder is automatically created to hold images and other resources with the same name as the file plus '_files'. Still others were created by installed applications. All of these that could be identified were eliminated from the list of folders, so the analysis was restricted to only the folders that the participants named themselves.

A list of unique folder names was created. Because the duplicate folder names were eliminated, the information about containment of folders was not displayed, and the researcher coded from a simple list of unique folder names. These unique folder names were then inductively coded using a thematic open coding process (similar to that used in grounded theory). Every folder name used by every participant was assigned one more codes. The names assigned to the codes were continually examined to ensure that they accurately represented the material they were coding, and if necessary were changed to better represent their contents.

During the early part of the coding, new codes were added as needed, and sometimes codes were merged when it was recognized that they were actually coding the same concept. Eight codes were generated from Participant A (all except Security and Source). Source was added for Participant B, and Security was added for Participant A. Both of these codes were also used by other participants so were retained.

There were many folder names that could not be determined by the research, and these were classified as 'Unknown'. The percentage of folder names that could not be coded ranged from 3% to 36%, and averaged 18%. This is to be expected, as folder and file naming is a personal and idiosyncratic affair, and some names may only have meaning to their creator in a particular context.

The researcher who performed the coding was the same person who conducted the in depth interviews, and thus the knowledge gained in the interviews helped the researcher interpret some folder names that otherwise would be unclassifiable, and helped to ensure that the folders were coded according to what they meant to their creator as much as possible. However, this is necessarily an approximate process.

In order to check the extent to which the coding depended on the knowledge gained during the interviews, a second researcher coded Participant C's folder structure using the coding scheme established by the primary researcher. The percentage agreement between the coders was found to be 73%. After collaboration between the two researchers, this rose to 86%. This is surprisingly high given the subjective nature of the subject matter.

To assess patterns in the overall structure of the folders, the researcher made an assessment based on the interview and from inspection of the file system snapshot. This is a necessarily imprecise estimate, as the folder structures of these participants had hundreds or thousands of folders, and it is common for different parts of the structure to be organized differently and to a different degree. To allow for this, the researcher also assigned a confidence assessment, of either low, medium or high, which indicated how pervasive the identified primary scheme seemed to be across the entire folder structure. If there seems to be two equally pervasive schemes, both were identified.

The coding scheme inductively developed from the folder names is at a fairly coarse granularity. For instance, of the references to Time, some show a sequence (Week 1, Week 2), some indicate a relative age (Old, History, Archive), some indicate a particular year, and some indicate an exact date. Further analysis will be done to analyze the folder names with greater discrimination, once data is available from more participants. Being based on six participants, this study is necessarily limited in its generalizability, so the intention of this paper is not to make generalizations, but rather to identify areas where further study is needed, and to highlight interesting findings that should be investigated further.

4. RESULTS AND DISCUSSION

This section presents both the results found and a discussion of those results. First we report the results of inductively generating the folder name codes, followed by a discussion of the code combinations that occurred in the data. Next we report the results of the card sorting exercise, and finally the results of analyzing the overall folder structures.

4.1 Folder Names

The folder names were coded to describe what type of information the folder name conveys about the contents (files and subfolders) of the folder. For instance, a folder named "Data Communications" tells you about the subject matter (topic) that can be expected in the folder, whereas a folder named "Lectures" tells you about the form and purpose (genre) of the contents.

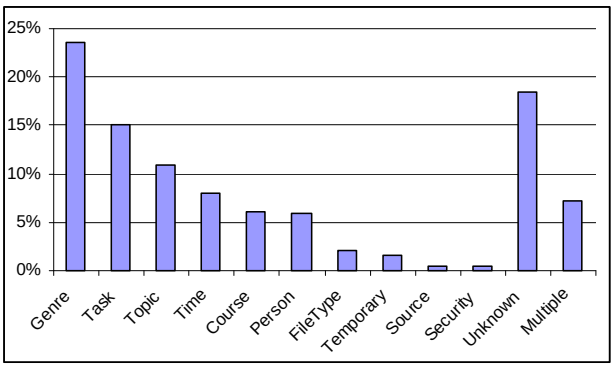

Figure 2. Average Proportion of Folders with each Code

Figure 2 shows the average proportion of folder names that was assigned to each code. The proportion of folders that were Unknown, and the proportion that was coded by multiple codes are also shown. Overall, the most frequently found types of folder name were Genre, Task, Topic and Time, followed closely by Course and Person.

Table 2 lists the codes that were derived from analysis of the folder names, with a definition and some examples of each code.

Table 2. Inductively Generated Codes

Code	Description & Examples
Genre	Indicates that the contents of the folder are a particular class or type of document, with a commonly recognized form and structure. Examples: Lecture Notes, Presentations, Timesheets, Budgets, Letters.
Task	Indicates that the contents of the folder are related to a task, project, event or some other type of activity. Examples: Assignment 5, Lec01, PhD, recruitment, evaluation, For DSS Presentation.
Course	Indicates that the contents of the folder are related to a specific course. (This is a special case of Task above) Examples: Database Systems, 222, INFOSYS 222
Topic	Indicates that the contents of the folder are all about a particular subject matter. Examples: Web development, Database Architectures, JavaScript
Time	Indicates that the contents of the folder are related to a particular time period, or have a time related aspect. Examples: 2005, 2003 SC, Old, History, Week12, Archive
Person	Indicates that the contents of the folder are related to a particular person, group or organization. Examples: Matthew, Audit Committee
File Type	Indicates that the contents of the folder are all a particular file format. Examples: zips, PowerPoints, Excel docs
Temp	Indicates that the name of the folder appears to have no intrinsic meaning and that little thought was given to assigning the name. Examples: foo, bar, fffff, asdfasdf, New Folder
Source	Indicates where the contents of the folder originated, either a location or person. Examples: From Brenda, From J Drive, Copy of R Drive
Security	Indicates that the contents of the folder are subject to particular security constraints or permission level. Examples: Personal, Confidential, Private

Overall, the most common type was **Genre**, and it is the one that deserves the most explanation. The genre of a document tells you what kind of document it is, something about its purpose and form. Orlikowski and Yates define it as a "distinctive type of communicative action, characterized by a socially recognized communicative purpose and common aspects of form" [11 p.543]. For instance, knowing that a document is a newsletter gives us a different set of expectations as to what we can do with it than if we knew it was a journal article, a budget or a map. These distinctions are more to do with the purpose and form of the document than with the subject matter (topic) it is about.

Genre is something that is easily understandable to people but is rather difficult for a computer to understand. Assessment of document genre is not a completely objective classification, since different people can have different assessments of the genre of a document [12]. In addition, people may deal with a vastly differing set of genres, depending on their job. Work is currently being done by Roussinov et al on automatic genre classification and using that to facilitate web searching [12], a line of investigation that these results suggest should be pursued further.

In this study, **Task** was defined quite broadly to include activities, projects and events, as well as more traditional tasks. Task often appeared in two quite distinct places in the hierarchy, either at the top or at the bottom. At the top, the tasks were very broadly defined, like Teaching and Research, and might more properly be thought of as roles. At the bottom, the tasks were more discrete, detailed activities, such as Tutorial 4. The concept of Tasks would be a good place to further investigate and refine the coding scheme with more participants.

The **Course** code is really just a specialized form of the Task code. Because all participants work at a University, courses figure prominently in their lives. It was decided to retain Course as a separate code, so that there would be flexibility to either separate it out or include it with Task as required.

Topic is something that can sometimes be determined by keywords in a file. A lot of research has been done on automatically categorizing documents into pre-established topic structures based on content analysis of the document itself (e.g. [13]) This is possibly the area where current automated assistance could be most useful.

Time was quite heavily used by most of the participants, although this is probably reflective of the fact that the study was conducted within a university. In an academic situation, the same course runs again and again and each instance of the same course needs to be distinguished from the others through some kind of time designation. It is entirely possible that in other situations that don't have regularly repeating activities, time would not be nearly so important. However, it is interesting to notice that the Software Developer also had moderate use of the time cue, and the two lecturers had almost no use of it at all. The use and importance of time in a wider setting is something that requires further work.

Most operating systems have time-stamping mechanisms that record when a file was created, accessed and modified, however, these are unlikely to be a substitute for the time dimension as observed here. As an example, consider a lab exercise that is handed out to the students. It is copied from a previous year's lab exercise, so the creation date is a couple of years ago. The accessed date is frequently changed by the backup process and other automated processes, but the modified date may give a reasonable indication of the age of the file. However, sometime software processes can interfere with this, such as an auto-save process saving the document while it is opened for printing. This could alter the modified date even if the file was not actually

modified. Events like this mean that the modified date (as it is implemented by current file systems) cannot truly be trusted in order to locate the file in time, although it can provide a good starting point for assisting with automatic determination of date.

Person is a dimension of document that could be readily supplied by document management software. Already Microsoft Office documents include an Author attribute that is automatically filled in based on the login name of the user when the document was created. Mechanisms similar to this could be used to supply this attribute value.

File Type is an interesting code, since some researchers have suggested that the segmentation by file type is an artificial distinction that has limited relevance for document management and retrieval [14, 15]. Not only did file type appear in the way the folders are named, but during the interviews, all subjects reported sorting their documents by file type or searching on file type in order to quickly locate documents of a certain type. This would tend to suggest that file type is a necessary cue to allow people to distinguish and retrieve documents. However, it is possible that file type could be being used as a proxy for genre, since genre information is not available in current file systems. For instance, sorting by file type would allow you to easily distinguish between a presentation (typically a PowerPoint file), a journal article (commonly stored as PDF), and a budget (likely to be an Excel file). Although this doesn't have very much power to discriminate between documents, it nonetheless might be useful for that purpose in the absence of genre information. These findings would suggest that more investigation needs to be done on the usefulness of file type before it is dismissed in the design of future document management systems.

Source is another code that is very interesting, although not commonly used. The folders coded as Source also included two that were actually specifying a destination, but since this only occurred twice and for one participant only, it was not coded separately. However, source was also sometimes implied in folders named for people, such as 'Annie's lectures', and 'Jim's'. Whether a document arrives as an email attachment, is downloaded from somewhere or copied from another location, the document management software should be able to detect where it came from and automatically store that information.

Since email is now the primary form of document exchange [16], most documents that were not created by the owner probably arrived as email attachments. This gives potential for even more information to be stored, such as exactly who they came from and when. Ideally, this information should be accessible when managing documents, and not solely in the email system.

Security was the least frequently used of the codes, but was encountered in three of the six participants. This designation of certain things as private, confidential or shared is something that could be easily supported by document management software. Since it seems that security designations are relatively seldom used, it would probably be appropriate for it to default to a 'Normal' setting, and which could be changed by the user when required.

The **Temp** folders were an interesting group, although they only appeared in three of the participants. More investigation needs to be done on why these folders were created, and what breakdown in the process is causing them to appear. It would also be interesting to know how long they last, and whether they are eventually given a more meaningful name, or deleted altogether.

Table 3 shows the percentage of each participant's folders that were classified using each Code.

Table 3. Proportion of Folders Coded with each Code

Code	A	B	C	D	E	F	Avg
Genre	12%	20%	29%	32%	32%	13%	**24%**
Task	11%	17%	4%	3%	1%	55%	**15%**
Topic	7%	5%	7%	7%	23%	15%	**11%**
Time	8%	6%	22%	9%	1%		**8%**
Course	2%	2%	10%	20%	1%	3%	**6%**
Person	8%	1%	9%	16%	1%		**6%**
Temp	8%	1%	<1%				**2%**
File Type	4%	1%	1%		3%	3%	**2%**
Source		<1%	1%		2%		**1%**
Security			<1%	2%	1%		**0%**
Multiple	2%	25%	3%	7%	1%	8%	**8%**
Unknown	36%	20%	13%	3%	35%	4%	**18%**

This table shows the percentages after system-created folders have been excluded from the analysis. Due to rounding for display purposes, the totals in each column may not add up to exactly 100%.

For all participants except Participant F, Genre was the most common type of file name encountered. For Participant F, it was outweighed only by Task. Folders of type Temp, Source, and Security were only employed by three of the six participants in this study. Participant F only exhibited use of five of the 10 codes, while participant C showed use of all of them.

Figure 3 shows radial graphs of how each of the six participants different with respect to the proportions on their documents that were classified according to the top four codes. For the purposes of this graph, Course has been included with Task. This gives a graphical view of the profile of the top four codes for each participant.

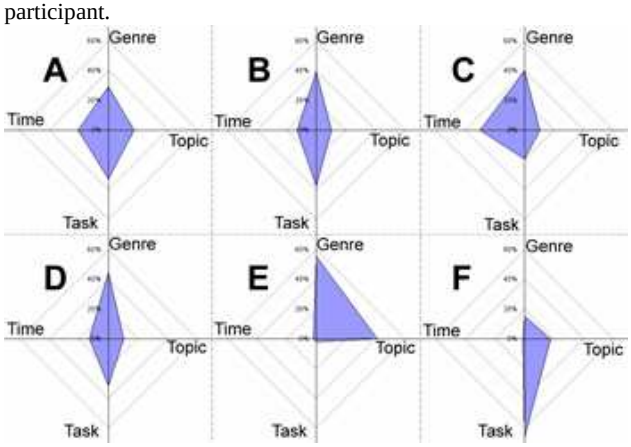

Figure 3. Radial Graphs showing profile of the top four codes

Participants A, B, C and D all have reasonably similar profiles. Participant A has an approximately equal proportion of each of the four codes, with a slight tendency to more Task-based folder names. B and D have a very similar distribution, with a tendency towards more Task-based folders and fewer Time-based folders, whereas C tends to the opposite, with more Time based and fewer task based. Participants B, C & D are all Course Managers, performing essentially the same duties, so it is perhaps expected that their folder name profiles are similar to each other.

However, Participant E and F both have the same position (both are Lecturers, with similar teaching and research responsibilities), but their folder name profiles show quite different tendencies. E tends to have mainly Genre and Topic oriented folder names, with very little use of the Task or Time dimensions, whereas F tends to have overwhelmingly Task based folder names, with some Genre and Topic but no Time-based folders at all.

4.2 Multiple Coded Folder Names

In addition to the folders that were coded with a single type, some folder names were assigned multiple codes. This happened when the folder name was made up of multiple parts, each which could be classified differently. Some examples of common multi-code folder names are given here. If a folder could be classified as more than one name, it was given a primary code and a secondary code.

Not all possible combinations of codes were actually present in the folder names of the participants. Table 4 shows the combinations that were actually used.

Table 4. Examples of Multiple-Coded Folder Names

Folder Name	Primary Code	Secondary Code
2004 Excel	Time	Topic
INFOSYS 222 2005AC	Course	Time
Recruiting 2003	Task	Time
INFOSYS 222 Exams	Course	Genre
Jim's Timesheets	Person	Genre

Table 5. Combinations of Multiple Coded Folder Names

Second-ary → / Primary ↓	Course	Genre	Person	Task	Time	Topic	File Type	Security
Course		6%		3%	**14%**	2%	1%	
Genre	1%		1%		3%		1%	1%
Person		4%			1%			
Task		6%			7%			
Time		1%				**38%**		
Topic		4%	1%		1%			

These are relative percentages of multiple-coded documents, not percentages of all documents. These percentages add up to more than 100% due to rounding for display purposes.

The most common combination of Time and Topic was largely due to Participant B, who had 25% of her folders multiple coded. Participant B has an ongoing project that she works on every day. Every time she finished working on it, she makes a new copy of the entire folder containing the project material, and names it with the current date, as well as the topic that she most recently addressed in the project. These folders accounted for almost all of the Time+Topic coded folders, and also accounted for the high level of duplication in Participant B's folders, since she creates a duplicate copy of the entire folder structure on a daily basis.

The next most common set of codes was the combination of Course and Time. This is unsurprising, since courses run over multiple semesters, and it is common at this University to refer to one particular instance of a course using the course code, year and semester designation.

Other combinations include the combination of Task and Time, commonly seen in names like Lab Week 4, Recruiting 2003; the combination of Course and Genre, as in INFOSYS 222 Tutorials; and Task and Genre, as in Lecture 5 Demos.

It is very interesting that only 8% of the folder names were multiple coded, and even that figure was very much inflated by Participant B's naming practice. This doesn't necessarily mean that the folder names are single word, just that the folder names tend to only represent one of these dimensions. Thus, it seems that these multiple dimensions are applied to documents through combining several single-dimension folder names into a hierarchy, rather than by constructing multiple dimension folder names and using those.

One plausible reason for this is reuse. Especially in the university environment where courses repeat year after year, it is possible to copy a folder and all its subfolder and documents, and by changing one of the folder names higher up the tree, change the context of all the documents below it. Consider the example in Figure 1 at the beginning of this paper. By simply copying the 2005 folder and naming the copy 2006, the user can change the context of all the documents and folders below it. It would be interesting to see whether the same findings occur in a situation where there was not such strong time based replication as in the university.

In addition to these folders that are dual coded, Table 6 shows the combinations of triple coding that were found.

Table 6. Triple Coded Folder Names

Code	Example	Incidence
Course+Time+Genre	INFOSYS 222 2005 AC Tutorials	1%
Person+Topic+Genre	Yin's Modeling Tutorials	1%
Time+Genre+Source	OldExamsFromEmma	2%

This triple coding was relatively uncommon, accounting for less than 0.5% of the folder names overall, and less than 5% of the multiple-coded names, and only appearing in two of the participants' file systems. This is probably due to the fact that more flexibility is obtained from nesting single folder names than from encoding the same information into a single folder name.

For example, consider the case of Course+Time+Genre, and a folder called "INFOSYS 222 2005F Tutorials 222" representing the tutorials for a database course taught in the first semester of 2005. These three pieces of information can all be encoded into a single folder name, or the same information can be discerned by nesting three separate folder names. The following three hierarchies would all give the same information:

- INFOSYS 222 > 2005F > Tutorials
- Tutorials > INFOSYS 222 > 2005F
- 2005F > INFOSYS 222 > Tutorials

Having the separate folders allows other documents and folders to be placed into the intermediate levels, providing context to those without having to create additional folders.

4.3 Folder Hierarchy

Table 7 shows the primary organization scheme employed by each participant. The confidence column indicates how much confidence the researcher had in how rigorously this scheme was followed throughout the file system.

Table 7. Primary Organizational Scheme

Participant	Scheme	Confidence
A	Time > [various]	Low
B	Time > Course > Task	Medium
C	Genre > Time	Medium
D	Task > Course > Time > Genre	High
E	Task > Time > Course > Genre or Task > Course > Time > Genre	High
F	Genre/ Task > [various]	Low

Of interest is the different primary organizing schemes used by B, C and D. These three participants are all Course Managers whose profiles of folder types were all very similar to each other; however, their dominant organizing schemes are all quite different. None of these people expressed significant dissatisfaction with their organizing scheme, and all seems able to effectively use their structure to perform their jobs.

This would suggest that perhaps the order in which these folder types are combined is not particularly important. As noted before, if you place a document in the bottom level, the combination of all the folders in the hierarchy above supply the required meta data to be able to distinguish the file from others, regardless of the order in which they were encountered.

During the interview with Participant E, he noticed that what he thought were two identical folder structures that he maintained in two different places, were actually different. In one place, the order was Task > Time > Course > Genre, in the other it was Task > Course > Time > Genre. Despite priding himself on keeping these two structures perfectly synchronized with each other, he'd never noticed that in fact the order of two of the primary dimensions was different. This would seem to confirm, that for this participant at least, it makes little difference which order these dimensions appear in. He commented that there was "nothing to choose between them".

Figure 4. Three equivalent folder structures

Each of the three folder structures shown in Figure 4 encodes the same information about the files and folders in the bottom level. All three of these structures are showing ways to assign three dimensions to a documents: a course dimension (2 options), a time dimension (2 options), and a genre dimension (3 options). Since there is no natural subordination between these dimensions, any of these three structures will do the job. However, precisely because there is no natural subordination, all of them involve considerable duplication of information.

One possible way of overcoming this problem would be to consider these dimensions as separate facets of the document, and to allow the user to provide values for these facets separately. The document management system could then allow any of the views shown in Figure 4 to be dynamically created, and to be changed on the fly by reordering the facets as needed. In addition, the facets could be used to filter the information displayed if required. Hearst et al [17] have had success in using a faceted search system called Flamenco, which operated on a collection of landscape and architecture photographs. Although personal document collections have quite different characteristics to an image library, this technique seems a promising avenue of exploration for document management.

5. CONCLUSION

The document dimensions identified provide a starting point for understanding how people structure their documents. More data needs to be collected in order to further test and refine these dimensions and to see if there are other important dimensions that may be required.

Some of these seem to provide some opportunity for automatic software support of document management (Person, Source, Topic, Time, File Type), relieving the user of having to enter this metadata manually. For the other dimensions (Genre, Course/Task, Security), some research is under way investigating how software can automatically detect this, but for now, users still have to add this information manually. More research of this sort is needed.

More research is also required to understand how widely applicable these identified dimensions are. If these dimensions are widely applicable, it is necessary to investigate the potential significance of the different profiles of these dimensions in a person's folder system, and how these profiles correlate with other document management behavior, job types, or personality types. A study is currently underway to investigate this with over 50 participants.

These dimensions can be combined in many different ways, since there is no "best" way to combine them into a hierarchy. Forcing them into a hierarchy results in duplication, so a more promising approach appears to be to treat them as independent facets and design a facet-based document management system which allows the facets to be dynamically combined into a hierarchy according to the preferences of the user or their needs at the time. Creating a sophisticated, easy-to-use system that incorporates these facets will not be an easy task, but identifying the most useful facets to support is a good starting point.

6. ACKNOWLEDGMENTS

Thanks to all the participants for their co-operation and to Stephen Witherden for his comments and feedback on the draft of this paper.

7. REFERENCES

[1] M. Lansdale, "The psychology of personal information management," *Applied Ergonomics*, vol. 19, pp. 55-66, 1988.

[2] A. F. Farhoomand and D. H. Drury, "Managerial Information Overload," *Communications of the ACM*, vol. 45, pp. 127-131, 2002.

[3] E. Freeman and D. Gelernter, "Lifestreams: A Storage Model for Personal Data," *SIGMOD Bulletin*, vol. 25, pp. 80-86, 1996.

[4] J. Rekimoto, "Time Machine Computing: A time-centric approach for the information environment," presented at UIST'99 Symposium on User Interface Software and Technology, Asheville, North Carolina, USA, 1999.

[5] V. Bellotti, B. Dalal, N. Good, P. Flynn, D. G. Bobrow, and N. Ducheneaut, "What a To-Do: Studies of Task Management Towards the Design of a Personal Task List Manager," presented at CHI'2004 Conference on Human Factors in Computing Systems, Vienna, Austria, 2004.

[6] P. Dourish, W. K. Edwards, A. LaMarca, and M. Salisbury, "Presto: An Experimental Architecture for Fluid Interactive Document Spaces," *ACM Transactions on Computer-Human Interaction*, vol. 6, pp. 133-161, 1999.

[7] E. Svenonius, *The Intellectual Foundation of Information Organization*. Cambridge, Massachusetts: MIT Press, 2000.

[8] A. Kao, L. Quach, S. Poteet, and S. Woods, "User assisted text classification and knowledge management," presented at Twelfth International Conference on Information and Knowledge Management, New Orleans, LA, USA, 2003.

[9] B. H. Kwasnik, "How a Personal Document's Intended Use or Purpose Affects its Classification in an Office," presented at SIGIR'89 Conference on Research and Development in Information Retrieval, Cambridge, Massachusetts, USA, 1989.

[10] D. Gonçalves and J. A. Jorge, "Describing Documents: What Can Users Tell Us?," presented at IUI'04 International Conference on Intelligent User Interfaces, Madeira, Funchal, Portugal, 2004.

[11] W. J. Orlikowski and J. Yates, "Genre repertoire: The structuring of communicative practices in organizations," *Administrative Science Quarterly*, vol. 39, pp. 541-574, 1994.

[12] D. Roussinov, K. Crowston, M. Nilan, B. H. Kwasnik, J. Cai, and X. Liu, "Genre-based navigation on the Web," presented at HICSS-34, Maui, Hawaii, USA, 2001.

[13] C. Apté, F. Damerau, and S. M. Weis, "Automated learning of decision rules for text categorization," *ACM Transactions on Information Systems*, vol. 12, pp. 233-251, 1994.

[14] O. Bergman, R. Beyth-Marom, and R. Nachmias, "The User-Subjective Approach to Personal Information Management Systems," *Journal of the American Society for Information Science and Technology*, vol. 54, pp. 872-878, 2003.

[15] R. Boardman and M. A. Sasse, "Too Many Hierarchies? The Daily Struggle for Control of the Workspace," presented at HCI International'03 International Conference on Human-Computer Interaction, Crete, Greece, 2003.

[16] N. Ducheneaut and V. Bellotti, "E-mail as Habitat: An Exploration of Embedded Personal Information Management," *Interactions*, vol. 8, pp. 30-38, 2001.

[17] M. Hearst, A. Elliott, J. English, R. Sinha, K. Swearingen, and K.-P. Yee, "Finding the Flow in Web Site Search," *Communications of the ACM*, vol. 45, pp. 42-49, 2002.