

The Folders We Live In: What We Need and What We Can Get

Hong Zhang

GSLIS, University of Illinois at Urbana-Champaign
501 E Daniel St
Champaign, IL61820
h Zhang1@illinois.edu

Michael Twidale

GSLIS, University of Illinois at Urbana-Champaign
501 E Daniel St
Champaign, IL61820
twidale@illinois.edu

ABSTRACT

This paper reports one part of results of a qualitative study conducted in an information institute. By investigating how people creatively organize their information items in current hierarchical folder systems on computers, we try to identify what people need from current hierarchical folder structure and what they can get from the current structure. Specifically, people need two types of grouping in addition to ordering and highlighting, and especially better support on derivative relationships between items or groups of items. Current organization systems can provide overview and implicit contextual and workflow information. The impact of derivative relationship on multiple classification mechanisms is noted, and the connections between folders, tags and their possible use are discussed. The study provides implications for information organization system design.

Categories and Subject Descriptors

H.1.2 [User/machine system]: Human information processing

General Terms

Management, Documentation, Human Factors.

Keywords

Document Organization, Personal Information Management.

1. INTRODUCTION

We have long recognized the limitations of the current hierarchical folder system which leads to difficulty in filing and accessing. For example, documents can only be filed at one place in the hierarchy and thus can not be accessed according to a different criterion [1]; filing to a folder can hide information which reduces the visual cues for memory and recognition [2], among others.

But on the other hand, empirical studies show people prefer browsing to retrieve items and use search only as a last resort (e.g. [3]). Recent studies show that technology advances have little changed how individuals organize information [4][5]. People seem still to quite rely on hierarchical folder systems and there is no sign that we can live without them [2].

At the same time, new and sometimes radically different prototypes have been proposed as alternative systems including topical, temporal, and spatial metaphors (e.g. [6][7][8][9]). Recently, semantic tags have been used in variety of experimental systems to provide alternative or complementary ways of

organizing and accessing information items (e.g., [10][11][12]).

It has been noted that more understanding is needed in how people use the hierarchical folder systems and “where and how it is inadequate” [13] “before we discard folders as an outdated relic” [2]. As an information habitat for us for decades, it is notable that the current hierarchical folder system is an all-in-one place for storage and organization, visualization and retrieval, and for workspace. Our reliance on the tool as well as the problems with it can come from different layers. It deserves further investigation in terms of what we really need from current folder systems.

This paper reports some of the results from a qualitative study conducted in an information institute which tries to take a deep look at how people organize information items in current hierarchical folder systems and the extent to which this makes it possible or easy to re-access information. The study only focused on the users’ own files (not system files), and mainly with respect to a single computer, even though participants may have more than one computer. Some other results will be reported elsewhere in order to narrow down the scope of this paper.

2. RELATED WORK

In [1], Dourish et al. list several limitations of current hierarchical folder systems which include: (1) It does not allow storing an information item in more than one folder, which can cause filing difficulty. (2) It further restricts the way to retrieve it and can only be accessed in the same way to categorize it. (3) Since items are confined to the “physical” locations, information management activities such as backup are constrained to the structure instead of according to user needs. Additionally, for document organization, folder and file names are very limited in representing metadata about the files [10]. It also has been recognized that “folders can obscure” which can result in out of sight, out of mind [2]. Further on, some types of information e.g. recipes are better represented in a faceted object classification scheme [2][1].

But sometimes a seeming drawback can be an advantage. For example, it is observed that sometimes information hiding is exactly what people want, tidying some files ‘out of the way’ to enable the remainder to be more visible. Even the drawback of a document needing to be in only one place is sometimes an advantage because that means “we know exactly where to look for it later” [14]. We are reminded that current folder systems “do provide some (albeit limited) abilities” [13]. Creating hierarchical structure is a natural way to deal with information. When

reflecting upon the experience of using a personal electronic notebook Proteus, Erickson noticed the “gradual addition of more and more layers of structure” in which one notebook was broken into sections, and then subsections were added, basically a hierarchical structure, although he “neither wanted nor needed the structure” at the beginning [15].

In recognizing the importance of representing rich relationships - including hierarchical relationships, between information items, a few recent systems try to use semantic tags to label documents, and provide better support for multiple classification (e.g., [10] [11] [1]), although there are studies showing that people are not likely to spend time to assign metadata [16][17].

For example, Dourish et al. proposed Placeless Document system based on document properties which avoids many problems of traditional hierarchical organization systems. It augments it with active properties which enables the provision of document -based services. It refines document properties into uniform properties and user-specific properties and provide “collection” mechanism instead of traditional “folder” concept [1]. In [18], Quan et al. proposed a user interface with Web browsing in which users can do multiple categorization with attribute-value pairs. In a prototype called Newdocms, Arriaga removes the “file name” concept, and replaces it with attribute-value pairs in describing files. Similar to [1], “collection” is used for sharing metadata within hierarchical relationships. Oleksik et al. created a tagging system that works with Windows desktop metaphor that can be used as a layer on top of hierarchical file system structure [12].

Inspired by such prior work, we wanted to know more about how people organize their files in current hierarchical folders and specifically, what we need and what we can get from current organization systems.

3. METHOD

The study is designed to try to obtain rich data through multiple channels. Two rounds of semi-structured interviews were conducted with six Ph.D. students and six administrative staff in an academic environment in front of their computers within three months. During the interview, the participants gave the investigator a “guided tour” of their main information organization systems. The interviews are guided with a set of broad, open questions which include background information, currently working project folders, completed project folders, and possibly miscellaneous folders. Re-access difficulties were discussed with the corresponding folders. The participants were asked to talk about their practices of organizing and re-organizing specific document and email folders and files, and to try to re-access several randomly selected files or emails. The interviews were audio recorded, and screenshots taken of selected folders. Participants were also asked to report via email any experiences of information re-access difficulty during the three months. The two rounds of interviews looked at similar issues although the second one focused on the new and changed part. Two rounds of interviews allowed evolving issues to be captured and explained to provide better understanding than a one-time snapshot. A second chance to look at the same issues for both investigator and subjects provided an opportunity to clarify and complement information collected in the first interview. It is especially important in studies on personal information organization behaviors because often people’s information organization

behavior is subconsciously conducted without much thinking. A second interview can sometimes reveal deeper information and perception the participant did not realize at first. The 3 months allows time for interviewees to pay attention to and report information re-access difficulty experiences via email and the second interview provides a chance to discuss them. Data reported here are mainly from the interviews.

Data analysis consisted of careful re-reading of the audio transcripts looking for patterns of interest. With such a small sample size, we are focusing on depth of understanding rather than making claims for breadth of coverage.

We use representative quotations to illustrate the findings. Student participants are referred to as PP and administrative staff participants as AP. Other identifying information is replaced by a <generic description>.

4. OVERVIEW OF STRUCTURE AND ORGANIZATION BEHAVIORS

When asking about their overall comfort level with using their computers to deal with files, given scale of 1 to 5, with 5 meaning very comfortable, the average comfort level was 4.3. Half of the participants (six) gave 5; one of each group gave 3, and the other 4 gave 4. When they were asked about their attitude to current hierarchical folder systems, almost all the participants made claims such as: “comfortable with the folder”, “it works pretty well for me”, or “I’m pretty happy with the trees”. There are though several features they hoped the system could support, e.g., version support in a broader sense than a linear versioning relationship; easy and flexible backup and digital preservation mechanism (e.g. dealing with data currently only readable with software on an old computer); the possibility of multiple classification in trees; a better drag-and-drop mechanism; and a file activity tracking mechanism.

4.1 What We Can Get From Folders

4.1.1 Organization as a view

We found that participants’ hierarchical folder systems, in addition to serving as organization and access tools, represent how they understand and see their information space and work. It is more apparent amongst the PPs, most of whom talked about the “big chunks” in their information work represented in the top level folders. The general terms used in top level folders such as “research”, “academic”, “school”, “job” or similar terms represent how the participants see their information world. As P2 said when she was asked what she thinks her information space looks like:

... based on my activities. I have my own study, my dissertation, my teaching assistant. In work, I have these two big chunks. I’m looking for a job, I have all the job related information. Then I have my personal life, my photos, music I like ... (P2)

P3 said “I mostly think about it just in terms of file folders”. P5 had 30 top level folders, but he used several quick links (on Mac) for current main parts. Compared to PPs, APs have more top level folders. This may be because their jobs involve more separate activities, although a subset of these may consume major parts of their work at any particular time. Participants generally have a global view of their files which guides most filing and re-

accessing behavior; although there are always new items coming where there is “no systematic place for it”. When an AP created a folder under her home directory for a course she was teaching, she created a subfolder “teaching” and then a subsubfolder “<coursename>” for this course, even though the “teaching” folder only included the single subfolder “coursename>”, and there was no substantial expectation of teaching several new courses in the near future.

4.1.2 Context in folders and files

We found, as other studies (e.g. [10][14]) have, that folder/subfolder and file names are used to describe or even ‘tag’ a document’s attributes (e.g. author or genre) as well as provide contextual information (e.g. “what it is for”, “who asked me to do it”). In addition we found that the files or folders around a given file/folder can convey contextual or workflow information to help recognize and understand what a particular item is. As a participant said:

...It’s all about context. If you just give me a file name, I may not know what it is. But even like I’m looking at the folder now, oh what that .pdf, oh wait, a .doc file next to it, oh that’s the organization track.

On one case a file’s existence under a folder informed an AP of the status of an information item in the workflow because the specific file process procedure and convention is implicit in the folders. We found that several administrative participants claimed they actually intentionally save more files in order to keep a trail of activities or procedures for next time’s use.

However, it is notable that not all the effort of building context into names paid off. Several participants noted names that were so obvious when they created them, but may not be helpful when trying to find them later:

...I tried to name things systematically, like with the date I took I wrote it down and whatever, but it doesn’t always work. Or like a few months later, you look at the name, it doesn’t mean anything. It was so obvious when you did it, but it’s really not.

As another PP said, naming a file or a folder is “just what I think of it at the time. It seems very clear when I do it.” She has a single field exam file under a folder: “PhD evaluation”, and she said “I don’t know what it’s about. I mean I know that’s a field reading list, but I don’t know why I put it in this folder.”

While context may be impossible to be fully represented in any system, these examples indicate the limited functionality in current systems to support the representation of context and lack of help for the naming process.

4.2 Organization Structures

Unlike general hierarchical systems such as library subject classification, folders/subfolders in personal information organization are not always in a logical hierarchy and accrete over time. People may aspire to be systematic, but it can be too difficult or too effortful to be consistent. For example, a PP participant collected many articles from a KM journal at one time and so created a subfolder “J of KM” for those articles under her “dissertation references” folder. Later, when she found another article published in the same journal, she filed it to the parent folder “dissertation references” instead of the corresponding

subfolder, because: “I didn’t realize I had this folder, I didn’t even bother.”

This organization structure is not especially good for re-access, as the participant admitted, since she often did not remember which journal a desired article was from. But the main reason why the participant didn’t pay much attention on this folder’s organization was that: “If I know this is very important, I will print it out.” She relied mainly on printed out paper copies for her work.

This example also shows that organization structure on personal computers is affected by use of paper and web resources [19].

4.2.1 The current folder structure decays

Recalling the note from Ravasio et al. that “classification and document storage were considered on-going processes” and “hence no structure was thought of as permanent” [20], We saw this temporality in some situations when a folder structure that initially is useful does not make sense later. For example, two folders became overlapped and not separable over time; or the perspective with which a user wanted to access a group of files changed; or the purpose for using a group of files changed. As a PP said:

...originally, under “research”, I just have stuff I worked for my research assistantships, or also my own work. And then the “academic” folder I think I originally made because it was stuff that’s school related but not for classes. And the “school” folder was class work. But that’s all become kind of mixed up. So I know what my rationale was for making all three of them. It just doesn’t really make any sense any more.

Even when folder structures decay in their usefulness, participants may not always re-organize them, except for small refinements and adjustments. Re-organization is low priority and conducted only when it is deemed a really big issue. Participants chose not to re-organize for reasons such as: being familiar with the existing structure; or not bothering since it is not important; not bothering since re-access is infrequent; or it does not affect important work.

4.2.2 Incomplete re-organization

We found across many participants that when they do re-organize, it is usually incomplete and not systemic. It is typical that a participant created a folder for a new file at some time, but did not collect other related files scattered in other folders into that folder. This can cause re-access difficulties. One main reason for incomplete re-organizing is the understandable perception of “too much effort”.

4.3 What We Need From Organization Systems

It was widely observed that participants used numbers or letters as a prefix of file or folder names to order them in a particular way when viewed in alphabetical order.

A related behavior is highlighting. For example, a participant highlighted several folders with the function provided by MacOS, and an AP use different colors to highlight emails. These two functions can help better organize items and shorten navigation time in access.

In addition to these, there are two basic needs observed in this study: grouping with hiding, and grouping without hiding, which

will be described in section 5. Another need is to represent versioning and more complicated derivative relationships, which will be covered in section 6. Multiple classification will be discussed in section 7.

Under current systems, these needs can be more or less supported in current hierarchical folder-based file system or by users' work-around strategies. For example, "grouping and hiding" can be well supported by creating a folder or subfolder; "grouping without hiding" and "ordering" functions are not supported and currently rely on naming mechanisms as a workaround; highlighting is supported by coloring in the Mac operating system and some email systems.

5. GROUPINGS

Two groupings are needed in both the organization stage and later on provide context in retrieval stage or in workspace: grouping with, and grouping without hiding.

5.1 Folders: Grouping With Hiding

Grouping and at the same time hiding files can be achieved by creating a folder. It is identified as a basic need in information organization in this study, similar to the observations in [2 & 10]. When the participants were asked why they created a particular folder or subfolder, a majority of the answers are to "separate ... from ..." or "didn't want to mix it up with the rest of my stuff". Hiding is also sometimes a need, especially when the number of items grows to a level that it imposes visual complexity:

Because if I have all of my files out at a once, I wouldn't be able to see, or I might not remember just the way I titled the documents. But once I have them into certain folder, it eliminates some of the other choices.

The folder metaphor in current systems is a very good tool for grouping and hiding. The ease of moving files and folders is an important part of the usability of a system, because personal information structure can be provisional.

Although participants have some implicit or explicit general criteria, sometimes the decision of creating a folder is a best-guess judgment at the moment and could be found wrong later. For example, participants may create a folder when they were "desperate to save something". Several participants have folders including only one file because "when you create this you never know how many are going to come," or "you thought you'd have more stuff, but later on you stopped doing it."

In fact, this study found that instead of well guided information management behaviors, many decisions in personal information organization such as to keep or not, to delete or not, how to name a file or folder, in addition to create a folder or not, sometimes are more of a "judgment call at the moment" in "a very brief interaction", and "it's in the moment, whatever seems right in the moment", which can be partly explained by the "cognitive obstacle between knowing or anticipating how information is to be used" and encoding these factors in the folder names [10].

5.2 Grouping Without Hiding

Files and subfolders in a particular folder are sometimes in more than one sub-group even though they are not put in subfolders. For example, one participant commented on a folder:

...it just confuses everything in here. Because these are archive or reference directories, these are actual content directories, these are actual content directories, and this is a specific kind content directory that I used a lot.

One AP had 83 files under her home directory. They include "very general things", files that "don't have matching folders", files saved by software by default to this location, and files that "once used heavily", "at least temporarily they were important", and files "for easy access". Among them, some files are old and not used, those that until the interview she had even forgot the existence of; files described as right now useless although may have been used heavily; used recently, used periodically or used sometimes; and files that have not been but "should be foldered".

Although sometimes this is a case of a 'miscellaneous' folder that is just allowed to grow until it gets so unwieldy as to need and justify the effort of tidying, sometimes participants did not want to hide a group or have it too deep by filing them 'away' to a subfolder.

Since current systems do not explicitly support grouping without hiding, one workaround was quite commonly used: renaming files so that in an alphabetic list view of the folder they would be adjacent and so could be seen as a subgroup. For example, a participant had a group of files with the same name except the ending number (NS_Update_1, NS_Update_2, through NS_Update_8). Other participants did nothing and got used to ignoring the unrelated ones in their daily work, and later on rely on memory to feature it out when they have to.

As noted above, information items grouped together can provide implicit context for each other. In the study, there were a few times that participants could not remember what an individual file was or was for because it was not together with other related items. For example:

...I don't remember what this is. ...oh oh oh, I know what this is now. ... this used to be filed under the class I took with him. But since I no longer on the class, but I still attended the group meeting sometimes, it's here now. And I don't think I moved the class stuff over which is why I didn't recognize it.

Furthermore, we observed several cases when grouping-without-hiding seem more suitable for files currently being worked with, and filing them to a folder (grouping-and-hiding) occurs at a later stage. For example, a PP had physical folders for hard copy papers according to the subject of an article. But she only filed them after she is done with the class. When she's taking the class, "they are all in a big pile".

Another participant worked on files under her home directory ("My Documents") and filed them after the work was done. She explained that she wanted to pay more attention to doing the work at that time rather than on where to put it and how to organize it.

6. THE DIFFICULTY OF DERIVATIVE RELATIONSHIPS

Current file systems have no good mechanism to represent many relationships other than generic-specific, whole-part, or template-instance relationships. For example, different configuration files to be included in a program for different purpose; website registration receipt and the purchased article. But similar to [4], we found that managing different versions was a troublesome problem for the participants. We also found that participants were even more troubled by complicated derivative relationships between files.

6.1 Versioning: Current and Final Version

Participants used file/folder names to differentiate versions. For example, a participant used “final_” and “final_final_” as a prefix trying to indicate the final version of her dissertation. Another participant used “stage 1” and then “stage 1 new” to differentiate between folders of two versions.

Participants had a common need to make the current version and especially the final version very distinct from others. A PP said “I only want to see what is the most current, but also with this idea that I might still want to know what was in an old version”. An AP wanted the final version to “popup” form others so she “would know what I need to look at first, instead of looking through 20-25 folders or files.”

In some cases where the final version had to be put under a different folder for particular use, the connection to the version chain became disconnected. For example, several APs worked in their own directory and then put the final version on a shared drive/folder to share with colleagues. A PP put all final papers together in a folder for a particular use. But later on, the version under the original folder could be changed without updating the final folder, which caused confusion and hesitance when they tried to find the real final version some time later.

It is important to note that versioning can be more complex than a simple linear temporal sequence. There are other types of versions that could cause access difficulties. For example, during the interview, a PP spent a while to remember the connection between two related files mixed among other files, finally recalling that one was a “lengthy notes about the whole thing” written first, and the other was a synopsis produced from the previous one for submission.

6.2 Complex Derivative Relationships

It is even more difficult in current hierarchical structures to represent and display complex derivative relationships between files. As a PP said:

...part of the problem is that there is so much overlap between, that's why I have so many multiple files because everything is connected.

This same participant reported a re-access difficulty in trying to find a particular excel spreadsheet among a few which have a complex derivative relationship between them:

...I was never clear when I was doing it. It was very confusing because we change a few things at different times. ... (Interviewer: but when you modified and produced a new file, you were clear what it's about) yes. I was for a little while. But then I forgot.

Although they are under the same folder, the grouping can not convey the exact relationships between the files. Sometimes the relationships could be between information items, and could be between projects since projects evolve. A PP had over 38 top level folders, and most of them were projects or course folders that served for a large study with complicated derivative relationships. She felt her files were not well organized and had frequent file re-access difficulties.

Since there is no way to represent these relationships, people have to rely on their memory to reconstruct the specific derivative relationships between folders as much as they can in order to make sense of them.

7. MULTIPLE CLASSIFICATION PROBLEMS

When participants were asked what they would do when one file is related to multiple folders, most of the APs said they would make a choice, often based on which is the most important. They would often create folders for people they worked with and then if a file related to more than one person, they could make a decision that they found simple: “whoever has seniority”, “who decided it, who initiated it”, etc. Multiple classification dilemmas seemed not to be a major problem for them. They seemed to have clear criteria in deciding where such an item will go, and later on where to find when there is overlap between two folders. Consequently, even when there are multiple folders a file is related to, they have different priorities that can be relatively unambiguously ranked.

In the PhD group, this multiple classification issue is usually manifested in a downloaded article that covers multiple issues, such as being related to two theories. One participant's solution was:

...if I need it for both theories, I saved it in both, which is redundant, but it lets me file it. I figure space is cheap enough that the only thing that really affects (is) the amount of time takes of backup of my files.(P1)

This file duplication solution was observed in another PP's folders, while yet another PP made the choice to “put it in a place that's more frequently used.”

There is another type of situation observed in 3 PPs in which participants had downloaded articles in different course folders for each course use. But later on at some stage, they wanted to re-collect those articles for research use and according classify them in a different way. But there is no easy way to re-collect them in current systems, and so the participants ended up re-downloading most of them and creating new folders.

Other than these cases, PPs did not emphasize the difficulty of dealing with multiple classification. One possible reason for this group is that 5 of the 6 PPs, (including two very technical people), declared that they relied more on hard copies of these readings in doing the field exam, course work or their dissertation. The reasons include: preferring to read on paper rather than on computer, doing mark-up and note taking more easily on paper, and several readings only having a paper format. Our sample size makes us especially hesitant to generalize from this finding, but we did find it surprising and wonder if it would show up in a larger study. With the current growth in ebooks we also wonder if this preference for paper will persist.

Version consistency is always a challenge for any multiple classification strategies. An AP expressed the concern of inconsistent versions and the confusion it may cause if she put one thing in two folders. Since the final or current version is a big concern for administrative participants, the benefit of more access points does not seem to outweigh the risk of using the wrong version. When an administrative participant talked about the risk of losing her final version file on shared folder, she said “I almost rather do that (lose it) than risk having the wrong letter.”

For PPs, “saving in both” may only apply for the downloaded pdf files which usually do not have a version consistency problem, since they are not normally modified. But for editable files, before the version consistency problem is well addressed, they may have to adopt the same solution of making a choice as the APs. This also partly explains why multiple classification is less problematic in email system than in file organization system – received or sent emails are not changed, although there might still exist minor version confusions in such email systems. For example, in Gmail many people see labels as folders (e.g., one participant called the labels “folders” in this study). If one message has two labels, deleting a message under a “folder” (instead of de-labeling) could make it disappeared under both “folders”.

8. DISCUSSION AND IMPLICATIONS

8.1 Why Both Grouping With and Without Hiding?

This study identifies two groupings as basic needs for information organization systems, with the understanding that current simple hierarchical folders on computers (files, emails) are a multiple-layer tool for storage and organization, retrieval and visualization, and workspace. Each layer imposes its own requirements and systems may have different problems on different layers. Each one plays a role in the required/provided functions. Identifying “grouping without hiding” is to recognize a need from the visualization layer. The problem of “hiding” information in current folder systems may be solved by a visualization design change. There are already some examples elsewhere to provide this “grouping-without-hiding” function. For example, in Firefox Bookmarks, a list of bookmarks can be arbitrarily ordered to group related links more closely together, and a horizontal line can be inserted to further support this separating and grouping function. Another example is the way current browsers display xml documents. A node can be expanded and collapsed to show and hide.

8.2 What Do We Need From Systems?

Overall, we need two grouping mechanisms to support the “group and separate” need, user-specified ordering and highlighting mechanisms and better support for derivative relationships.

A related issue is the observed struggle with the “where am I” problem within a file hierarchy. With a deep folder structure such as A/B/C/D/E/F, you first have to know where you want to be going, and then you have to get there. This is effortful, especially in command line mode, but even with a GUI, it leads to a lot of clicking. One participant having deep folders needed to use “scp” very frequently to upload files to servers. After typing an extremely long path in most of the semester, he created a quick link file including all the links he needed, and put the file in

multiple places where he would be doing files uploads. This example shows that to have a “bird’s eye view” instead of a sequential path view is not just for visualization. It’s also for accessing.

Another related problem is the linear view in current folder and file lists. There may be alternative views such as ordered by file name or modification date, but they remain linear. Although there are studies arguing that spatial information organization is not preferred [21], allowing moving files or folders around in a two dimensional space makes it possible to represent richer and subtle relationships between files and/or folders.

Some benefits we can get from current systems include:

- We can get some kind of overall view from top level folders;
- we have a stable structure that allows us get more and more familiar with it over our use, represents how we think about and look at our information space (although may not always represent the “current” view and thinking), and represents our particular priorities in our work and life;
- we can get implicit context and workflow information that can help us access, understand, and work with information items.

These functions and hidden benefits should be considered in alternative systems.

8.3 Folders, Tags, and Multiple Classification

Naming a folder and a file can be seen as an easy way to tag a file or a group of files, similar to the observation in many studies (e.g. [10]). As a participant declared, folder and subfolder names are like metadata tags which usually indicate information such as what the files are for, what they are about, what kind of genre they are, and what particular feature they have such as year or semester. It suggests that file retrieval tools within file organization system should make use of these “tags” in the search tools. For example, searching “paper on information theory” should be able to find the files under the folder research/paper/information theory/, even though the files do not include the term “paper” and “information theory”.

Indeed, folder/file names at many times are more like tags when we do not have a specific purpose or we do not know or care about the main tags. For example, in the “J of KM” example described in 4.2, the journal name folders are more like tags, because that’s all the user knew at that time and wanted to separate them up, but not necessarily the way the participant wants to organize them, which may emerge later when a clear purpose or idea come out. This analysis is consistent with the observation in [12] which proves that tags are especially useful in pre-organizational stage as an intermediary step “when the scope is still unclear” and “before the resources are included into the storage and organizational structure, e.g., by creating a folder in the file system hierarchy.”

Folder and subfolder names can be used as tags to help re-accessing from any folder/subfolder names. For example, if the participants described in section 7 can collect all the papers scattered in the course folders by clicking a tag “papers” – assuming they have a “papers” folder under each course folder, they can re-collect those articles and classify them according to

their new needs. Reciprocally, participants may be better oriented and pay more attention in folder/file naming.

Folder-subfolder chains can be seen as ordered tags, and the order can reflect the particular priority the user has. The sequence (structure) can represent or convey relationships between the tags. It limits the combination possibilities of tags, reducing a core functionality of real tags but this may help decrease ambiguity to some extent. These findings and design implications seem to indicate that hierarchical structure is still good at the organization stage, and that incremental improvements are more desirable than radical innovations in how people organize their information items..

As described in section 2, there are already several proposed systems that use tags for multiple classification, some of which work with current file folder systems. But folder/file names and tags might be better in a single system, otherwise people will need to do extra, duplicate tagging. Tag quality might be another issue, since it will directly affect later retrieval. For example, tags may not differentiate enough, since they are given in a flat way, and not under a specific context.

Finally, we should note that on PC, many files are editable working files, not just web pages, downloaded pdf files, or emails. The version issue is a challenge for all multiple classification mechanism, which is not addressed in currently proposed systems.

Even for non-editable items such as emails, subtle concept differences between folders and categories/labels (or looked at in another way, between items and links to them), can be confusing for users. This is especially so because the physical folder metaphor exists in our everyday lives. New concepts may need new metaphors.

8.4 Limitations of the Study

With just 12 participants, all from the same department, our findings are necessarily highly provisional. Nevertheless we think they are indicative. They show that even in such a small sample, there are many different things that people do to enable them to organize and re-access their files. There are various approaches and workarounds and any one person may have multiple needs that change over time. Our current hierarchical file systems, for all their limitations, do allow for considerable diversity of use. They allow for messiness, for guesses about likely use that prove wrong, for organizing “in the moment”, and for subsequent reorganizations and partial reorganizations. We see differences between doctoral students and administrators, but also many similarities. We have attempted to show that certain themes recur, even when manifested in different forms.

The academic context may lead to certain peculiarities of use. However it does not guarantee an effortless and elegant approach to personal information management. We see intelligent people with considerable familiarity with computers struggling to stay on top of their organizational structures. They mostly succeed, but it is hard work. We would not expect the problems identified to be substantially less in other settings. If anything we predict that they would be magnified and perhaps accompanied by issues and problems we did not see in our small study.

It is important to extend these studies with larger numbers of subjects and across different use contexts. One simple low cost

triangulation is for the reader to compare the findings with his/her own file organization practices and those of others that are visible in public file hierarchies, as well as findings already reported in the literature. There is a lot to be found in studying the activities of a small number of people in depth.

9. CONCLUSION

A good system needs to support various levels of organizational rigour, lapses, emergent structures, organizing on-the-fly and sometimes inherent messiness.

Our analysis of what people do and how they cope with thousands of existing files, floods of new files, constantly changing tasks, considerable uncertainty about future needs and uses, limited formal skills in information organization, and limited time to devote to organization, only serves to remind us of the creativity of people in doing actions that many (including possibly our participants themselves) would view as mundane – merely a boring adjunct to their main work. It also reminds us how robust the existing hierarchical file system is.

The functionality provided and the interface through which people use those features of course affects what they do.

It is tempting to devise exotic new filing mechanisms. But we believe it is important to understand what we need and what we can get from current folders systems, in order to ensure that alternatives do not fix one issue but worsen three others.

10. REFERENCES

- [1] P. Dourish, W.K. Edwards, A. LaMarca, J. Lamping, K. Petersen, M. Salisbury, D.B. Terry, and J. Thornton, “Extending document management systems with user-specific active properties,” *ACM Transactions on Information Systems*, vol. 18, 2000, pp. 140–170.
- [2] W. Jones, A.J. Phuwanartnurak, R. Gill, and H. Bruce, “Don’t Take My Folders Away! Organizing Personal Information to Get Things Done.”
- [3] D.K. Barreau, “Context as a factor in personal information management systems,” *Journal of the American Society for Information Science*, vol. 46, 1995, pp. 327–339.
- [4] D. Barreau, “The persistence of behavior and form in the organization of personal information,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 59, 2008, pp. 307-317.
- [5] O. Bergman, R. Beyth-Marom, R. Nachmias, N. Gradovitch, and S. Whittaker, “Improved search engines and navigation preference in personal information management,” *ACM Trans. Inf. Syst.*, vol. 26, 2008, pp. 1-24.
- [6] E. Adar, D. Karger, and L.A. Stein, “Haystack: per-user information environment,” *CIKM*, 1999.
- [7] E. Freeman and D. Gelernter, “Lifestreams: A storage model for personal data,” *SIGMOD Bulletin*, vol. 25, Mar. 1996, pp. 80–86.
- [8] J. Rekimoto, “Time-Machine Computing: a time-centric approach for the information environment,” *UIST99: Proceedings of the 12th Annual ACM Symposium on User Interface Software and Technology*, Asheville, NC: 1999.
- [9] G. Robertson, M. Czerwinski, and K. Larson, “Data Mountain: Using spatial memory for document management,” *UIST98: Proceedings of the 11th Annual ACM Symposium on User Interface Software and*

- Technology*, San Francisco, CA: 1998.
- [10] M. Arriaga, "Going beyond the hierarchical file system: a new approach to document storage and retrieval," 2002.
 - [11] D. Quan, K. Bakshi, D. Huynh, and D.R. Karger, "User Interfaces for supporting multiple categorization," *Proceedings of Interact*, 2003.
 - [12] Gerard Oleksik, Max L. Wilson, Craig Tashman, Eduarda Mendes Rodrigues, Gabriella Kazai, Gavin Smyth, Natasa Milic-Frayling, and Rachel Jones, "Lightweight Tagging Expands Information and Activity Management Practices - Microsoft Research," 2009, pp. 279-288.
 - [13] S. Henderson, "How do people organize their desktops," *CHI*, 2004.
 - [14] Andrea Civan, William Jones, Predrag Klasnja, and Harry Bruce, "Better to Organize Personal Information by Folders Or by Tags?: The Devil Is in the Details," Columbus, Ohio: 2008.
 - [15] T. Erickson, "A Personal Electronic Notebook, by Thomas Erickson," *CHI'96*, 1996.
 - [16] Gary Marsden and David E. Cairns, "Improving the usability of the hierarchical file system," *Proceedings of the 2003 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology*, 2003, pp. 122-129.
 - [17] K. Rodden and K. Wood, "How do people manage their digital photographs," *CHI*, 2003.
 - [18] Dennis Quan, Karun Bakshi, David Huynh, and David R. Karger, "User Interfaces for Supporting Multiple Categorization," 2003.
 - [19] Hong Zhang and Michael Twidale, "To keep, or not to keep: or options in between," Champaign, Illinois: 2010.
 - [20] P. Ravasio, S.G. Schär, and H. Krueger, "In pursuit of desktop evolution: User problems and practices with modern desktop systems," *ACM Trans. Comput.-Hum. Interact.*, vol. 11, 2004, pp. 156-180.
 - [21] W.P. Jones and S.T. Dumais, "The spatial metaphor for user interfaces: experimental tests of reference by location versus name," *ACM Transactions on Office Information Systems*, vol. 4, Jan. 1986, pp. 42-63.