

Folders as Workplaces and the Impact on Relationships Between Files

Hong Zhang

Michael Twidale

Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
{hzhang1, twidale}@illinois.edu

ABSTRACT

The current folder system on personal computers is not only a storage place for information organization and retrieval; it also serves as a set of potential workplaces for projects or tasks. We performed a qualitative study on two different participant groups in an academic institution. The study examined creation and use of folder content and structure. We found a continuum of folder types from homogeneous well-structured ones with minimal inter-file relationships to messier project folders with complicated relationships between files. Various derivative relationships between files and/or groups of files caused participants the most difficulties in finding and identifying files at re-access time. The ways that people cope and work around the limitations of the file system's organization can be useful in understanding needs, but also maybe why there is some inertia to change.

Author Keywords

Folders, personal information management

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI).

INTRODUCTION

The growing literature on Personal Information Management (PIM) reveals the complex and conflicting relationship that people have with the current widely used hierarchical file and folder system. We know that many people have difficult organizing and re-finding their files at least some of the time, and that there are various reasons why. Numerous innovative design solutions have been developed to address these identified problems and yet the hierarchical filesystem persists. People seem very attached to it, despite acknowledging its problems [18]. Why? As PIM research has shown there are many reasons, but in this paper we want to note one constellation of reasons around the use of the current folder system on personal computers not only as a storage place for information organization and retrieval, but also as potential workplaces for particular projects or tasks. The filesystem is not just about managing a large number of isolated files, but also about managing relationships between files. People innovate with the filesystem to create their own idiosyncratic organizations, to allow for different levels of messiness, and to cope with

changing circumstances. We don't want to say that is ideal – people still struggle – but there seems to be evidence of 'mundane innovation' to get some of the features that we see in the more radical proposed PIM design solutions. Again, we don't want to say that the status quo is 'good enough' but its robustness to many needs, practices and a variety of minor adaptive uses may help us understand its persistence. This in turn may help us think more carefully about incremental adaptations both to the filesystem and indeed to other applications with well-acknowledged flaws that remain bizarrely persistent.

RELATED WORK

The current hierarchical folder system on personal computers has long been criticized for its overly constraining structure and many limitations in filing and retrieving files [1, 9, 29]. For example, it does not allow storing an information item in more than one folder, which can cause filing and retrieval difficulty – information management activities such as backup are constrained to the structure instead of according to user needs [10]. Additionally, the mechanism to describe a file is very limited [1, 29, 32]. It also has been recognized that folders can obscure which can result in out of sight, out of mind [17,18] especially since a digital file folder structure can go much deeper than for a physical folder [6].

In a study drawing on paper use to inform digital system design, Bondarenko and Janssen noted that the hierarchical structure of files and folders is not a natural way for organizing information that people are still working on [6]. Visual cues of paper documents on a desk as well as textual cues in an email about the document are gone, and so people have to explicitly and effortfully create a context for a document by naming it and choosing or naming a proper folder, which in many cases is insufficient, since the names of files or folders are usually not descriptive enough to aid their recognition. The granularity of organization in the current file system is also a problem. Henderson found that many people have the problem of managing different versions of documents [14].

Despite the limitations of current folder systems, studies also found that users prefer to browse-and-locate information items in hierarchical file systems over direct keyword search [2]. Quan et al. found that folders are useful

for “file and forget” archiving and for organization schemes with fixed structures [29]. Boardman and Sasse pointed out that the stable folder systems provide a useful and usable familiar environment [5]. A seeming drawback can be an advantage. For example, folders can provide an “effective way to manage workflow” and clutter [8] - 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 that a document needs to be in only one place is sometimes an advantage because that means “we know exactly where to look for it later” [8].

Empirical studies show the important role of task or project in folder structure. For example, Henderson found four main categories of folder names which include genre, task, topic, and time based on categorization of folder names captured by file system snapshots [15], while Khoo et al. found that the labels for first level folders tended to be task-based or project-based [24]. Bondarenko and Janssen reaffirmed that document management is strongly related to task management for all information workers. They identified two types of activities they perform: research and administrative, in a spectrum with PhD students at the end of “research” and HR managers at the “administrative” end. They argue that all information workers have both unstructured research and structured administrative activities in varying degrees in their work. They further noted that the current digital document management system mainly supports pre-structured procedural tasks while leaving the research type unsupported, and suggest that “a document management system should place documents into their task-related context” [6].

A better task- or project-based organization method has long been recognized as an important need in information organization [11, 23, 30], since tasks or projects are facilities driving and organizing the daily work. Kaptelinin identified problems related to temporary file organization, personal activity planning, and recreating context of a file [22], and addressed these drawbacks with a task-centered system [23]. Ducheneaut and Bellotti designed Taskmaster which recasts email as task management facility using embedded task-centric resources [4, 11].

Different job content, personal strategies, and tools can lead to different document management and so the role of project or task in digital document management can vary. Kidd [25] made a distinction between knowledge workers, communication workers and clerical workers. Instead of fitting people into these categories, Bondarenko and Janssen looked at the two types of activities they perform: research and administrative. They argue that *all* information workers have both unstructured research and structured administrative activities in varying degrees in their work. They further reaffirmed in their study that document management is strongly related to task management for all information workers. They noted that the current digital document management system mainly supports pre-structured procedural tasks while leaving the research type

unsupported. They further suggest that “a document management system should place documents into their task-related context” [6].

Another body of research looks at how to connect information items and represent relationships between them. It is believed that the value of information “depended little on individual pieces of information, but rather on how the pieces were connected and thereby complemented one another” [30]. Some systems try to use semantic tags to label documents, and provide better support for multiple classification [1, 10, 29]. 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 enable the provision of document –based services. It refines document properties into uniform properties and user-specific properties and provides a “collection” mechanism instead of the traditional “folder” concept [10]. Quan et al. [29] 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 filename concept, and replaces it with attribute-value pairs in describing files [1]. Similar to [10], “collection” is used for sharing metadata within hierarchical relationships. Oleksik et al. created a tagging system that works with the Windows desktop metaphor and can be used as a layer on top of hierarchical file system structure [27]. However there are also studies showing that people are not likely to spend time to assign enough metadata that the effective use of many of these innovative systems would require [26, 31].

Some researchers have looked at the wider integrative complexity of PIM, of trying to integrate separate informational structures such as files, emails and bookmarks, handling multiple personal computers, laptops and mobile devices, or working with shared filesystems [19]. We certainly must acknowledge these as additional complexities, but have chosen to examine issues around the ‘simple’ case of just document files on a single PC.

The use of folders, the project nature of work with documents, and the relationships between files have been discussed separately in various settings. Our study was aimed at understanding peoples organizing and re-finding processes on their PCs. In this paper we look at one aspect of the findings that emerged in how people used folders as workplaces and struggled to manage the relationships between documents in these virtual places.

METHOD

Two rounds of in-depth semi-structured interviews were conducted with six PhD students and six administrative staff in an academic environment in front of their computers with a three-month interval between interviews. During each interview, the participants gave the investigator a guided tour of their main information organization systems. Although a set of broad, open-ended

questions were used to help guide the conversations, the actual interviews were directed by what was observed and what the participants talked about in their primary information organization systems which included mostly file folder systems and some email folder systems. The participants were asked to talk about their behavior in creating folders and organizing particular folders and files, as well as difficulties they might have in re-accessing them.

At the end of each interview, 3-4 folders were selected to run disk scan commands to capture the file folder structures. Based on the first interview disk scan data, at the time of the second interview several files or emails were randomly selected on each participant's computer to ask them to re-find them. During the three months between interviews, participants were also asked to report via email any experiences of information re-access difficulty.

These multiple instruments (interview, re-finding experiments, email, disk scan, screenshot) were designed to obtain rich data, although from an admittedly small sample. Nevertheless we believe that other small scale PIM studies have proved valuable, and the body of PIM findings enables some triangulation. The second interview followed a similar procedure to the first interview, although it included the re-access experiment and focused more on the new and changed parts of the folder structures. Two rounds of interviews allowed evolving issues to be captured and explained to complement the data collected in the first interview. Multiple interviews are especially valuable in studies of personal information organization behaviors because information organization behavior is often conducted without much explicit thought. The three months allowed time for interviewees to pay attention to and report information re-access difficulty experiences via email which were discussed during the second interviews.

	Operating System	Time in the Institution
Adm. N=6	Windows XP	3 months – 29 years
PhD N=6	3 Mac OS; 2 Windows XP; 1 Unix with XFS	1 – 6 years

Table 1

Each interview lasted 1 to 2 hours and were audio recorded and transcribed. The total audio length of 24 interviews is just over 34 hours. The transcripts were analyzed and coded in QSR NVivo 8. Table 1 shows the operating systems participants were using and how long they had been at the institution at the time of the first interview.

Using the two groups of participants provided rich and varied data. Since “activity type has a large impact on document management” [6], it is important to note that the two groups are not exclusively doing “research” and “administrative” activities. Rather, members of each group do both kinds (as defined by [6]), but in varying degrees,

giving an opportunity for a richer understanding of a continuum of multiple practices.

OVERVIEW OF THE FINDINGS

Many of our findings echo those of others in the substantial PIM literature. We include citations in the rest of the paper to indicate these recurrences. We use the extensive survey paper by Jones [19] as a convenient way to aggregate over a larger set of references than we have space for. These resonances are somewhat double edged. They increase the challenge of making clear the contribution of this paper – that various mundane innovations, coping behaviors and struggles around the relationship between documents-in-use are important to understand why people have problems with the existing file structure, that these workarounds can indicate requirements for systems innovation, but also may indicate why people could be reluctant to switch from a system that is problematic but perhaps almost-good-enough. The flipside of numerous resonances in the PIM literature is the opportunity for triangulation. If others have found very similar activities by different people doing different tasks in different settings, even with different technologies, then we can be more confident about the generalization of our findings from a small study. In HCI it has been relatively rare to be able to do such validation-by-lit-review because there have not been that many studies of the same issue. The pressure is always to do something completely new. In the case of PIM there is enough to begin to draw on the findings as an aggregation, and not just to seek to show how despite appearances your study is completely different to that of other studies of essentially the same topic. Of course this paper is in no way a meta-analysis of PIM; we just want to advocate for more citing-as-supporting-data as an approach and in contrast to just citing-to-make-a-distinction.

Although a small study, a substantial variety of different organizing practices emerged [33]. We had two distinct groups of participants, but there was considerable variation within each group [19]. Furthermore, the same person may organize different parts of her file system very differently, typically with a good reason why [3, 19] (“Many participants employed multiple PIM strategies within specific collections” [5]). Although we do not dispute the findings that some people are more organized than others, we also found that sometimes the same person will have some subfiles that they consider “neat” and others that they find “messy” and that they may even feel guilty about [19]. Organization methods seemed to be driven by the nature of the task, including how familiar and/or systematized the task was. Although the administrative staff had more systematized tasks and the PhD students more open ended tasks, all members of each group had both kinds, just in varying proportions.

TWO EXTREME TYPES OF FOLDERS

We found that folders could be described as belonging in a continuum between two extremes that we name “genre

folder” and “project folder”. A genre folder is one that includes one kind of file, e.g. “employee reimbursement forms, “invoice vouchers”, “meetings”, and “readings”. Figure 1 shows a part of an example genre folder of employees’ reimbursement forms, from an administrative participant’s computer. Genre folders had certain characteristics that recurred, but did not define them. These included often having the same file format (spreadsheet, pdf, etc.), and the files being of the same form (employee reimbursement, research papers, old drafts, etc.) [3].

At the other end of the continuum, a project folder was one created for a more open ended or less structured project, perhaps something the user was doing for the first time (such as planning a wedding [18]), and was often named with the project name or initials, organization names, or the corresponding person’s names. It often had a more heterogeneous collection of forms and formats and a less systematic naming structure. Figure 2 is a part of a research project folder from a PhD student’s computer. The project and person names in both figures have been changed to preserve privacy.

A key difference between these two folders is that in the genre folder, when the participant was pointed to a randomly selected file and asked about what it was, he or she was able to say exactly what it was from the file name, without the need to look at the files or folders around it. For example, in Figure 1, the file “November 2006 Austin Tx.xls” under “Jen” is the reimbursement spreadsheet for Jen’s trip to Austin in Nov. 2006. However, in a project folder, when the participant was asked about a randomly picked file, he or she often had to look at the other files or folders around it before being able to recognize or remember exactly what it was, although the file was named seemingly well. As a PhD student explained:

...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’s that .pdf, oh wait, a .doc file next to it. That’s the organization track.

This ‘look-around-and-recognize’ seems in part caused by the relationships between files in a project folder being more complicated than those in a genre folder. In the genre folder “employee reimbursement”, two files are simply two separate payment statements. To understand one does not require looking at any other files. The file names were sufficient to describe and understand what the file is. This type of folder is more like a traditional paper folder - the source metaphor of the current folder system on personal computers, so it is not surprising that the folder system works well for this type of folders.

```
H:\ZF07 EMPLOYEE REIMBURSEMENT\Alice\June
2007 St Louis.xls
H:\ZF07 EMPLOYEE REIMBURSEMENT\Ben\Ann
Arbor MI October 2006.xls
H:\ZF07 EMPLOYEE
REIMBURSEMENT\Jen\November 2006 Austin
TX.xls
```

Figure 1. Part of a genre folder.

```
ProjectA/4.txt
ProjectA/AmunChronicle11-17-07onecolumn.pdf
ProjectA/ResMacro_2007_win.sxw
ProjectA/new_members.txt
ProjectA/2008forcompart.txt
ProjectA/formatNewRes.txt
ProjectA/flashdrive_bkup.tar.gz
ProjectA/ResMacro_2007.sxw
ProjectA/macro_errors.pdf
ProjectA/macro_errors.ps
ProjectA/pwd-hg.gpg
ProjectA/drop-reasons.txt
ProjectA/amun_res_2007_upd.sql
ProjectA/amun_res_2007_upd.sqp
ProjectA/XrayTool52_en.sxw
ProjectA/XrayTool52_en.zip
ProjectA/Current-DB.mysql
ProjectA/res_update.mysql
ProjectA/tech-stuff.txt
```

Figure 2: Part of a project folder

On the other hand, under the project folder in Figure 2, the relationships between files and/or subfolders are more complicated. There might be data files, code files, reference articles, notes, and some files may have different versions. This type of folder is not as closely analogous to the traditional paper folder. It is more like the assembly of materials of work-in-progress on a physical desktop, or an agglomeration of materials hastily stacked up together in order to make room or tidy up a desktop. But there is no mechanism in the current folder system to capture and represent the relationships between these agglomerated materials. With the limited representative capability of file/folder naming mechanism, the user has to bear the various relationships in mind in order to recognize and understand a file later at the re-access time.

Another way to look at these two extreme cases is from the perspective of physical libraries and archives. A physical library shelves books together according to a subject classification scheme, but this is to optimize access to the particular one you want. The books nearby on the shelf will likely be about the same or a similar subject, but you don’t need to look at all of them to understand what the book in question is all about. This is very similar to the genre folder. By contrast, a physical archive often contains lots of boxes. Within a box can be many different items of

different types – reports, leaflets, drafts, receipts, letters, photographs etc. Historians spend a lot of time understanding the meaning of these individual artifacts by their relationship to each other. Each may not be so easily interpretable on its own. This is very similar to the project folder. The big difference is that we believe project folders are a result of using the folder as a workplace.

These two types of folders represent two ends of a continuum from simple to complex inter-file relationships. Most of the folders we studied fitted somewhere between these two extremes. For example, one student’s course folder had more complicated relationships than the genre folder but more structured relationships than the project folder, since the course had certain typical structural elements such as a syllabus, sets of readings, assignments, etc. Most of the students’ CV or resume folders were somewhere between the two ends, although some of them were close to the genre folder.

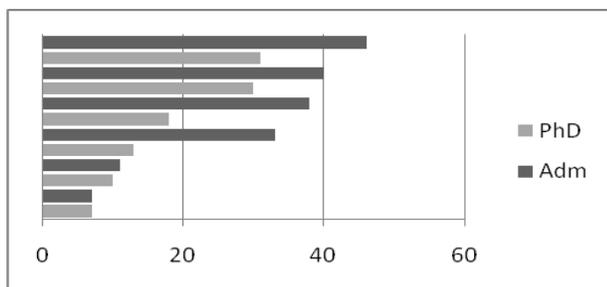


Figure 3. Numbers of the top level folders (1st interview)

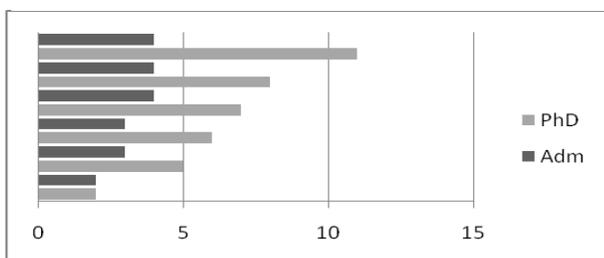


Figure 4: Maximum depth of the folders (1st interview)

PhD students had more folders at or close to the project folder end than administrative participants, and vice versa for folders at or near the genre folder end. This might be related to the finding that the administrative participants had broader but shallower folder structure than the PhD students, as shown in Figures 3 and 4. For the convenience of illustrating differences between and within groups, the values of the participants in the two groups are ordered by size and interleaved. Figure 3 shows that the administrative participants generally had more top level folders than the PhD student participants at the first interview time. Figure 4 shows that the maximum depth of folders among the disk scanned folders at the first interview time. The results at the second interview had a similar pattern.

Note the substantial individual differences within any group in the two figures. Similarly, although there was a

difference between the two groups of participants, a participant usually had both types of folders on his or her personal computers. For example, an administrative participant had a project folder “annual funds” while a PhD student had a top level genre folder “readings.”

FOLDERS AS WORKPLACES RATHER THAN STORAGE

We attribute the genre-project continuum in folders in part to the way that folders are used as well as to what is in them. This is related to two of Barreau and Nardi’s types: working and archived [2]. Some folders are used as workplaces in addition to serving as storage for organization and retrieval. In a workplace, people need to assemble needed and possibly useful resources so that they are easily to hand, both for efficient use and as a reminder that they may need to be used. In this workplace, people need to keep work progress and status traces. Thus many working files are kept and accumulated under a project folder to help convey a status and track of work progress. A simple example is the old version files many participants kept under some folders, which were useful for tracking when the user was working on the document. After the document was completed, the old versions may (see later) become useless. Given that “information management and task/ project management are two sides of the same coin” [19] it is not too surprising that people choose to do the latter within folders created for the former.

Desktops all the way down?

Many researchers have looked at the ways in which people use, organize (and fail to organize) files on their computer’s desktop, and the way it does or does not parallel activities on physical desktops [19]. Files are placed on the computer desktop for a number of reasons including: for quick access to frequently used files, for quick (and temporary) access to work in progress, to serve as a reminder to do a task, and as temporary storage for incoming files to be read, processed and then subsequently filed or deleted.

These actions and reasons make a distinction between the desktop and the rest of the hierarchy of folders within which the user organizes her files. Clearly from the perspective of the operating system, there is no great difference – the desktop is just a particular folder within the larger hierarchical file organization. Equally clearly, from the perspective of most end-users, the desktop looks and feels radically different from ‘normal’ folders. Similarly, the way that files are placed in the desktop is very different from the way they are placed in other folders. And yet we noticed surprising parallels between the way people talk about why they put some files on the desktop and why they put some files in certain folders.

Several researchers have advocated for multiple desktops, virtually increasing the screen real-estate, allowing for places or indeed rooms [13] to work on different projects, just as certain fortunate people have multiple or large physical desktop surfaces allocated to different ongoing

projects. We noticed that our participants sometimes used folders as workspaces in a ‘desktop-like’ way, using various workarounds to achieve some of the advantages (but not all and not as elegant) of their one actual virtual desktop. A workplace folder assembles the resources to do the task to hand. This can involve a heterogeneous assembly of resources that may ‘belong’ elsewhere.

A workplace folder often contained both files and subfolders. At least some users seem to perceive files mixed with folders as problematic. People talked about co-mingled files and folders being “messy”. The idealizations of file hierarchies in texts are almost always perfect if imbalanced trees – leaf files at the bottom (in a folder containing only files) and only folders of folders everywhere else – not lots of folders containing mixtures of files and folders. And yet commingled folders seem to pervade peoples’ filestructures [18]. Some talked about their effort to create “clean” or “well-organized” folders in which very few files co-mingle with sub-folders. However they also noted their failure to achieve this in other folders. Users may blame themselves more than the system for perceived shortcomings in their file organization [19].

The reasons users gave for why files were in a given folder alongside sub-folders closely parallel the reasons given above for placing files in the desktop. Not all of these files lurk there as unfiled, or as a to-be-filed-when-I-have-time. Some files had a logical home in a sub-folder and yet were ‘promoted’ to this higher level to avoid them being hidden from view (and consciousness): “I probably need to take these and put them in a folder. But the minute I put them in a folder, I will hide it, and then I’m gonna have to figure out, well, what folder did I put that in?”

Although no participant articulated it this way, the similarity of reasons for placing files higher in the file hierarchy than they might be seems to imply that participants use folders as pseudo-desktops to store some exceptional files slightly ‘above’ the overall folder hierarchical structure, but benefitting from the context of where they are stored and the files and folders stored alongside. It seems that moving between certain folders has some of the characteristics of moving from one pseudo-desktop to another.

The main computer desktop is used for many reasons, but one is as a workplace. It enables people to keep to hand the files they want to work on. It also allows for various kinds of grouping. People can quickly and with low effort convey meaning for themselves or ease access by exploiting the 2D desktop layout to move related files next to or close to each other. Our users could not do this in a regular subfolder acting as a pseudo-desktop – but they did create some adjacency workarounds by naming files so that when displayed in alphabetical order (by far the most common display option we saw) certain files will be adjacent, or pop to the top or the bottom of the list. [18]

After the work has been completed the folder-workplace may be tidied (or not). A typical tidying act is to create sub-folders to push down certain resources that are less likely to be referred to leaving the main products (or final versions of documents) as part of the now less cluttered main folder.

We choose to call these very simple workarounds ‘mundane innovations’. To a technical audience they can seem utterly banal. They don’t seem exactly like the end user innovations reported by von Hippel [16]. They typically don’t solve the problem, but they ameliorate it. They do allow people to get some of the features of advanced alternatives to the traditional hierarchical fine structure – in the case of this example to create a series of workplaces with (some) desktop-like features.

However, although this can be useful in adding a little flexibility, creating a space to do work (and have several projects on the go at once), it does not mean that people have solved their problem. Indeed their design fix can create its own set of extra problems.

Context in Folders

This use of folders as workplaces makes context information an important component in folders. The current hierarchical folder structure and naming mechanism provides a way for users to build in some metadata about files and relationships between files and/or groups of files. Participants organized and named files and folders to explicitly engrave an information item’s content (e.g. author or genre) and context (e.g. “what it is for” and “who asked me to do it”) into the folder system [15, 20, 30].

We also found implicit contextual or workflow information. For example, an administrative participant intentionally kept more files in order to keep a record of the way of doing a particular task:

I’m saving more things because I find that it’s helpful sometimes to reflect back on what I did previously. Keeping a trail of activities pertaining to a specific event or program is helpful as well because then I’m not relying totally on my memory. ...That has been extremely helpful.

This is a case of exploiting the falling cost of storage. The user *could* have created a document to record their process for future use, if they have to account for what they did or how they did it. Or the document could be consulted when this task or a similar one needs to be done a year hence. But documenting process is incredibly effortful. Not bothering to delete files is not only less work, it is even less work than deciding what can be deleted to tidy things up [3, 19, 35]. Deriving process and rationale from these related documents will indeed be more effort than reading a process document, but that is a cost to be discounted in an uncertain future. In another case, a file’s existence in a given folder informed an administrative participant of the status of the corresponding paper work in the workflow, because the specific file process procedure and convention is implicit in the folder structure.

Pseudo Tagging

Another mundane innovation is the way that the file and folder naming process can be appropriated to have some of the advantages (but not all and not in the most elegant way) of tagging – a feature that has often been advocated as a supplement or replacement for file organization.

Examples include adding “final” or “submitted” to the end of a document to indicate version status (considered more below). From Figure 1 the file “Alice\June 2007 St Louis.xls” has aspects of ‘tag-ness’ about it telling us different facets of information about the reimbursement and using tag-terms consistently within this structure. One participant created a folder under her home called *teaching* and within this a subfolder *coursename* for the course she was teaching, even though the teaching folder only contained the single subfolder and there was no substantial expectation of teaching several new courses in the near future. That seemingly superfluous extra level created a way to say more about what was at the bottom through two layers of folder names.

Relationships between files remain problematic

Despite these creative mundane innovations, the hierarchical structure and naming mechanism remains limited in conveying context, and people often have to largely rely on their memory. Several students spent a while to recognize or find a file. One explained “I just don’t remember what I did in the class.” Two students had difficulty in finding particular files because they forgot the way they worked on the files – using LaTeX, which had caused the files to be stored in a different location.

It is not feasible to try and consider all possibly useful relationships, but there are some that seem to recur enough to merit attention. In the PREMIS Data Dictionary for Preservation Metadata working group, three basic types of relationships between digital objects are identified: structural, derivation, and dependency [28]. We found examples of all three, as well as others. For example, a kind of structural relationship exists between a dissertation file and the files of the individual chapters; a configuration file and the main program file have a kind of dependency relationship. An example of another type of relationship is that between a website registration receipt and the purchased article. Among all these types of relationships, various derivative relationships seem to be the ones that the participants found the most difficult.

Derivative Relationships

PREMIS defines a derivation relationship as one between digital objects where one is the result of a transformation performed on the other. We use a looser definition where one can also be copied or adapted from the other. We noted two types of derivative relationships between files: versioning relationships, and the other more complicated derivative relationships.

Versioning: the Current and Final version

Managing different versions was troublesome for the participants [14, 19]. File and 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 the others. An administrative participant wanted the final version to “pop up” from the others so she “would know what I need to look at first, instead of looking through 20-25 folders or files.” A PhD student said:

I only want to see what is the most current, but also ... I might still want to know what was in an old version.

Folder naming strategy to note the versioning relationships did not work well when the final version had to be put in a different folder for particular purpose, since the connection to the version chain became disconnected. For example, several administrative participants worked in their own directory and then put the final version on a shared drive/folder to share with colleagues. When an administrative participant talked about the risk of losing her final version file on a shared folder, she said “I’d almost rather do that (lose it) than risk having the wrong letter.” A PhD student put all their final papers together in a folder for a particular use. But later on, the version under the original folder was changed without updating the final folder, which caused confusion when the participant tried to find the real final version some time later.

Complex Derivative Relationships

Even if we just consider single-authored documents, versioning can be more complex than a simple linear temporal sequence. There are other version types that can cause access difficulties. For example, during the interview, a PhD student spent a while to recall the connection between two related files mixed among others, finally recalling that one was “lengthy notes about the whole thing” written first, and the other was a synopsis produced from the previous one for submission. That relationship was essential to understanding what those two documents were (and why both had been saved).

It is difficult in current hierarchical structures to represent complex derivative relationships between files, other than by adjacency to mean ‘somehow connected’. A participant reported a re-access difficulty in trying to identify a particular excel spreadsheet among four spreadsheets 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 there are only four files in the folder, the relationships between them were so complicated that they were beyond the participant's capability to remember from their file names. In some other cases, the relationships could be between information items, and could be between projects since projects evolve. A PhD student 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:

...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.

Other examples are: the relationships between a paper, an abstract submitted based on it, and a poster which includes the abstract; various resume versions for different positions; a template file and three customized letters sent to three different audiences about the same topic.

Citation Relationships

Related to derivative relationships, citation relationships also caused some difficulty and participants noted they would like a mechanism to link the citing and cited items together. It can be a paper citing or including quotations from several other articles, or a reading note reviewing one or several articles, etc. A student reported that:

One thing that I've struggled with is finding pieces of information to include in my dissertation. ... I was looking for a complete citation for a particular book. But here's the thing, I could only remember the author and that I had cited it in another paper I wrote. It was not difficult for me to find the information once I remembered what paper I cited it in.

Another student reported a re-access difficulty case involving a citation mixed with a versioning issue:

*I was looking for a particular article citation that I knew I had used in a previous paper I hadn't used the file in quite some time, ... I was trying to provide the citation for a colleague. ... I knew I had come across the citation while working on a particular paper for a particular class. I use a familiar file naming scheme that includes the course number, which are also organized into directories by course. Since it was course related I also had some rough idea of dates. I looked at all the files from the course that I thought was related to the citation, but it wasn't in the bibliography of the document I thought I belonged in. I looked at the several earlier drafts located in the same directory, but still couldn't find the citation - although I did easily find the file I *thought* it should have been in.*

I did eventually find the citation I was looking for, but it wasn't in the file I thought it was in. Turns out I had not used the citation for the final paper, but I then recalled that I had been experimenting with a new citation software that used BibTeX files as a database. The BibTeX file wasn't in the directory with other Word documents, but stored in

another directory. I couldn't recall the file name, but knew it would have a .bib extension and the approximate date I last used it. I was able to perform a search using Spotlight that found all the .bib files, which I then sorted by date to find the right file.

During an interview, another student expressed the hope that "ideally I can specify which readings are used for which papers." Bidirectional links to connect them together reminds us of the association trails in Bush's Memex [7].

IMPLICATIONS AND DISCUSSION

Since most of the folders are not the well-organized systematic genre folders, extra features are necessary to accommodate extra needs. We can see many of these needs manifested through mundane innovations. Specifically, this study shows the need for multiple workplaces and suggests the importance of keeping the relationships between files and/or groups of files, especially the derivative relationships which are the most difficult for the participants in terms of understanding and re-accessing files. Although detailed recording and representation of the relationships (say through tagging or using other metadata) would be very effortful for users, the trail of derivative use relationships could be captured computationally. The challenge is to develop visualizations to help identify the corresponding derivative relationships.

Workplaces, Placelessness, Tags, & File Relationships

Personal Project Planner [21] is one approach that supports workplaces and integrates with the filesystem, meeting many of the needs identified here. The complicated relationships result from using folders as workplaces. For the example genre folder "reimbursement," the administrative participant was using her physical desktop as the workplace with paper form receipts. Her folder structure was just the repository for the final document, not 'where' the work was done. But we can imagine that if she changes her way of working and more receipts, paper works, and even correspondences become digital files, the files under the reimbursement folder would have more complicated relationships than the current folder (currently she prints out digital receipts to have them in the same place as the physical receipts). Indeed, the continuum from pure genre folder to project folder indicates that any folder on a personal computer can be a workplace.

Version control systems are specifically designed to address some of the file versioning issues that people had. However, to be acceptable it would need to be part of the file system. That is where people work. Furthermore, the version tracking in current applications is achieved by strictly following the check-in/out sequence. What is needed is to capture the relationships automatically based on co-occurrence, copy-paste and drag-and-drop activities.

This use of folders as workplaces seems to validate the place metaphor used in several systems, even though it has been blamed as the source of many current problems. As

noted earlier, the most criticized problem is the inability to put a file into two related folders, and various placeless [10] metaphors such as using document attributes or tagging/labeling mechanism to replace the current “folder” metaphor have been proposed. We suggest that although a document attribute-based approach is good for describing and organizing to some extent, the use of a workplace, which is usually observed collocated with the corresponding information collection in the physical world, is an unavoidable part of organization on personal computers. Furthermore, people typically have multiple projects, so take steps to create multiple workplaces. At the same time people innovate to have some documents in multiple places – they want and act to get aspects of both place and placelessness.

The possibility that any folder can be a workplace is a challenge for any tagging/labeling mechanism for editable files. For example, although a file can be tagged with label (project or workplace) A and label (project or workplace) B, it is possible that later on the user wants to modify the version under label A but not for the one under B, which means that file is split into two items with a derivative relationship between them. In this sense, a mechanism for derivative relationship is also necessary for maintaining a good tagging/labeling mechanism for editable files, when files are organized by what they are *used for* rather than what they *are*.

Overloaded File Naming Mechanism

A related issue is that the folder and file naming mechanism in current folder system is overloaded, not least by mundane innovations. Currently, file and folder names have to describe “what it is,” “what it is for,” plus the relationships with other files or folders. For example, it was widely observed that the participants use numbers appended to file names to indicate versioning sequence. However, if some of the files were located under another folder for a different purpose, this connection could be difficult to recognize. One case seen several times is where the final version of a file was moved away from the workplace folder where it was created to a different folder (sometimes a workplace where it would be used for another project) and so caused confusion later when the participant tried to find the final version. The WikiFolders system shows one way of addressing this issue while keeping many benefits of the hierarchical filesystem [34].

To make the situation even worse, file naming sometimes is not controlled by the user and has to serve other purposes. For example, after a PhD student talked about the effort he spent in keeping a convention in file naming, he added that “but at the same time, I don’t always have a control over that,” since the coauthor named the file differently. Another student had a file name changed before she sent it to another person “in order to make it more communicative for her so she knows this is mine.”

CONCLUSION

Our study showed how the folders that people created on their personal computers could be usefully understood as fitting in a continuum from “genre folders” with no inter-file relationships to “project folders” with complicated relationships between files. This continuum corresponded with the way that some folders are used as workplaces. We found that various derivative relationships between files and/or groups of files are the ones that caused the participants most difficulties in finding and identifying files at re-access time.

People use mundane innovations to cope with problems of information use and access. They are often low effort and ad hoc. These innovations are not necessarily solutions. They may not even be ‘good enough’ But they may be ‘not bad enough to precipitate doing something else or trying another system’. We believe that studying mundane innovations can be useful as an indication of user requirements – but not a guarantee that people will use the new feature thoroughly (just because they innovate a bit of tagging does not mean they want or will do a lot of tagging). Consequently mundane innovations are also a challenge to designing systems that people will bother to adopt.

REFERENCES

1. Arriaga, M. Going beyond the hierarchical file system: a new approach to document storage and retrieval. (2002). <http://pages.stern.nyu.edu/~marriaga/papers/beyond-the-hfs.pdf>
2. Barreau, D. and Nardi, B. A. Finding and reminding: file organization from the desktop. In *SIGCHI Bulletin*, 27, 3 (1995), 39-43.
3. Barreau, D. The Persistence of Behavior and Form in the Organization of Personal Information. *JASIST*, 59, 2 (2008), 310-317.
4. Bellotti, V., Ducheneaut, N., Howard, M., and Smith, I. Taking email to task: the design and evaluation of a task management centered email tool. In *Proc. CHI 2003*, ACM Press (2003), 345-352.
5. Boardman, R. and Sasse, M. “Stuff goes into the computer and doesn't come out”: A cross-tool study of personal information management. In *Proc. CHI 2004*, ACM Press (2004)583-590.
6. Bondarenko, O. and Janssen, R. Documents at Hand: Learning from Paper to Improve Digital Technologies. In *Proc. CHI 2005*, ACM Press (2005), 121-130.
7. Bush, V. As we may think. *The Atlantic*, 1945.
8. Civan, A., Jones, W., Klasnja, P., and Bruce, H. Better to Organize Personal Information by Folders Or by Tags?: The Devil Is in the Details. *ASIS&T 2008 Annual Meeting*, 45 (2008), 1-13.
9. Dourish, P., Edwards, W. K., LaMarca, A., and Salisbury, M. Presto: an experimental architecture for

- fluid interactive document spaces. *ACM Transactions on Computer Human Interaction*, 6, 2 (1999), 133-161.
10. Dourish, P., Edwards, W. K., LaMarca, A., Lamping, J., Petersen, K., Salisbury, M., Terry, D. B., et al. Extending document management systems with user-specific active properties. *ACM Transactions on Information Systems*, 18, 2 (2000), 140-170.
 11. Ducheneaut, N. and Bellotti, V. Talking about objects in e-mail. *Human-Computer Interaction*, 18, 1-2 (2003), 85-110.
 12. Henderson, S. How do people organize their desktops. In *Proc. CHI 2004*, ACM Press (2004), 99-112.
 13. Henderson, A. and Card, S. Rooms: The use of multiple virtual workspaces to reduce space contention in a Windows-based graphical user interface. *ACM Transactions on Graphics*, 5 (1986), 211-243.
 14. Henderson, S. Genre, task, topic and time: facets of personal digital document management. In *Proc. CHINZ '05*, ACM Press (2005), 75-82.
 15. Henderson, S. *How do people manage their documents? An empirical investigation into personal document management practices among knowledge workers*. (Dissertation). University of Auckland, (2009).
 16. von Hippel, E. *Democratizing Innovation*, Cambridge, MA: MIT Press, (2005).
 17. Jones, W., Bruce, H., and Dumais, S. Keeping Found Things Found on the Web. In *Proc. of the tenth International Conference on Information and Knowledge Management* (2001), 119-126.
 18. Jones, W., Phuwanturak, A. J., Gill, R., & Bruce, H. Don't Take My Folders Away! Organizing Personal Information to Get Things Done. *Ext. Abstracts CHI 2005*, ACM Press (2005), 1505-1508.
 19. Jones, W. Personal information management. *The Annual Review of Information Science and Technology*, 41 (2006), 453-504.
 20. Jones, W. and Teevan, J. *Personal information management*. University of Washington Press, 2007.
 21. Jones, W., P. Klasnja, A. Civan, & M. Adcock. The Personal Project Planner: Planning to organize personal information. In *Proc. CHI 2008*, ACM Press (2008), 681-684.
 22. Kaptelinin, V., Nardi, B. A., & Macaulay, C. The Activity Checklist: a tool for representing the "space" of context. In *Interactions*, 6, 4 (1999), 27-39.
 23. Kaptelinin, V. UMEA: Translating Interaction Histories into Project Contexts. In *Proc. CHI 2003*, ACM Press (2003), 353-360.
 24. Khoo, C. S., Luyt, B., Ee, C., Osman, J., Lim, H., and Yong, S. How users organize electronic files on their workstations in the office environment: a preliminary study of personal information organization behaviour. *Information Research*, 11, 2 (2007).
 25. Kidd, A. The marks are on the knowledge worker. In *Proc. CHI 1994*, ACM Press (1994), 186-191.
 26. Marsden, G., and Cairns, D. E. Improving the usability of the hierarchical file system. In *Proc. SAICSIT 2003*, (2003), 122-129.
 27. Oleksik, G., Wilson, M. L., Tashman, C., Rodrigues, E. M., Kazai, G., Smyth, G., Milic-Frayling, N. Lightweight Tagging Expands Information and Activity Management Practices. In *Proc. CHI 2009*, ACM Press (2009), 279-288.
 28. PREMIS editorial Committee. PREMIS: Preservation Metadata Maintenance Activity (version 2.0). (2008). <http://www.loc.gov/standards/premis/>
 29. Quan, D., Bakshi, K., Huynh, D., and Karger, D. R. User Interfaces for supporting multiple categorization. In *Proc. of Interact* (2003), 228-235.
 30. Ravasio, P., Schär, S. G., & Krueger, H. In pursuit of desktop evolution: User problems and practices with modern desktop systems. *ACM Trans. Computer-Human Interaction*, 11, 2 (2004), 156-180.
 31. Rodden, K. and Wood, K. How do people manage their digital photographs. In *Proc. CHI 2003*, ACM Press (2003), 409-416.
 32. Soules, C. A. N., and Ganger, G. R. Toward automatic context-based attribute assignment for semantic file systems. *Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-04-105* (2004).
 33. Volda, S. and Mynatt, E. D. It feels better than filing: everyday work experiences in an activity-based computing system. In *Proc. CHI 2008*, ACM Press (2008), 259-268.
 34. Volda, S. and Greenberg, S. WikiFolders: augmenting the display of folders to better convey the meaning of files. In *Proc. CHI 2009*, ACM Press (2009), 1679-1682.
 35. Whittaker, S., and Sidner, C. Email overload: Exploring personal information management of email. In *Proc. CHI 1996*, ACM Press (1996), 276-283.

The columns on the last page should be of approximately equal length.