Document Duplication: How Users (Struggle to) Manage File Copies and Versions

Sarah Henderson
The University of Auckland
Private Bag 92019, Auckland 1142, New Zealand
s.henderson@auckland.ac.nz

ABSTRACT
In personal document management, a common problem for users is handling file and folder duplication. Duplicates can be created deliberately (e.g. creating different versions of a document to preserve a history) or inadvertently (e.g. copying a file to a USB drive and then back to a different location). Users must spend time and effort to consciously and manually manage this duplication, or they run the risk of losing or overwriting data. This study of 73 knowledge workers combines a snapshot of their file system with a questionnaire about their document management practices in order to understand their document management structures, strategies and struggles. We find that current personal document management systems (i.e. the file systems built into modern Operating Systems) do not provide adequate support for managing file duplication. We explore the systems that users have developed to work around this deficiency and suggest some guidelines for the design of more effective document management systems.

Keywords
Personal information management, personal document management, document duplication, versioning.

INTRODUCTION
Most knowledge workers will spend a large part of their time working with documents: creating, editing, sending, receiving and reading information encoded within documents. Although the activities involved in managing these documents (creating, naming and renaming, filing, finding, deleting) are not particularly time consuming, they are undertaken by so many people so many times a day that they add up to a large expenditure of time. Understanding the challenges that are faced in document management is an important step towards designing systems that can better support these tasks.

ASIST 2011, October 9–13, 2011, New Orleans, LA, USA.

Documents and files are not synonymous. A document is a single conceptual entity, with an integrated form and purpose and a life history from creation, through editing to finally deletion or dormancy. A file is a logically connected chunk of data that exists on a storage medium. Files may contain documents, but they are also used to store data that is not a document, such as application files and configuration files. It is rare for a single file to contain more than one document, but reasonably common for a document to reside in more than one file. There are four different ways this can happen:

1. Splitting a document into subsections. A large document (such as a book or long report) might be split into one file per chapter. This is always deliberate.
2. Multiple file formats. A document might be saved in both Microsoft Word and PDF file formats. While there is still only a single conceptual document; it resides in two files. This is always deliberate.
3. Duplicate files. With duplicates, two or more files contain the exact same document content. A common personal document management activity is copying files from one device to another - from a work computer to a USB drive to a home computer for instance. These activities can result in multiple copies of a document existing in different places. Duplicates may be created deliberately, or accidentally.
4. File Versions. Versions are created as a document is edited throughout its lifecycle. As a document passes through several revisions, many users wish to keep track of the content they are editing or removing in case they need to go back to earlier content. Each time a change is made, a separate file is created with a snapshot of the content of the document at a particular point in time. Versioning is always deliberate.

The latter two issues are the focus of this paper. It is now very common for people to have multiple devices on which they access documents. Having documents in multiple places increases the risk that the documents may get out of sync, and this makes it ever more likely that they will face issues caused by file duplication. Versioning is also a

common activity, and is an issue that has been acknowledged and addressed in groupware and enterprise document management systems, but has not been studied at a more personal level. This paper aims to identify the issues knowledge workers may have in managing duplication and versioning and uncover the techniques they use to manage these issues. This will allow us to suggest improvements in document management systems which will allow people to be more productive.

BACKGROUND

Document management can be studied at three distinct levels: personal, group and enterprise. Personal document management is the process of an individual managing their own documents. It is personal in the sense that the documents are owned by or under the control of the individual doing the managing, not that the documents necessarily contain personal information. Group document management expands the individual situation to encompass a group of people who either work on documents collaboratively, or who interchange documents amongst each other. Enterprise document management considers the challenges that arise from managing documents across an entire organization. Appropriate systems must exist at all three levels in order to support fully effective document management.

Group and Enterprise Document Management. Version management is a more pronounced (and more widely acknowledged) issue in group and enterprise document management. With multiple people being able to edit documents (perhaps even concurrently), version control is included as one of the fundamental features needed in document management systems (Asprey & Middleton, 2008; Sprague, 1995). Other researchers have explored ways of providing support for managing versions in a group situation, such as the DocMan system (Backer & Busbach, 1996), which includes revision management, change histories and user notifications. A review of commercial groupware applications in 2006 showed that the majority provide some support for version management (Rama & Bishop, 2006).

Duplicate files is something that document management systems are designed to prevent, with an fundamental tenet of many systems being that each document has a single unique identifier and a single location. However, it is still possible that duplicates and near-duplicates could be added to a repository. While there is nothing in the literature about this problem, there is at least one commercial product that exists to solve the problem by detecting duplicates and near duplicates in enterprise document management systems (Equivio, 2006).

Personal Document Management. Versioning and duplication issues are almost completely absent from the personal document management literature. Personal document management is a subset of personal information management (PIM). This field aims to understand how people interact with their own information: their emails, web bookmarks, notes, contacts, calendar items, music, pictures, and of course, their documents.

The earliest studies of personal document management were conducted by Barreau (Barreau, 1995) & Nardi (Barreau & Nardi, 1995) in 1995. One study involved very computer-literate Macintosh users, while the other was managers with a wide range of computer experience who predominantly used DOS. The studies were predominantly concerned with understanding how users decide what documents to store and how they retrieve them again. Neither study mentioned any concerns with managing document versions or copies.

More recent studies have been done in 2003-2004. Gonçalves & Jorge analyzed the file systems of 11 users in 2003, with the aim of understanding how people organize their documents across multiple machines or devices. They found that 30% of their users had only a single device on which they stored their documents, with 60% having two machines, and the remainder having 3 or more distinct locations. Their analysis did not look at file duplication or versioning.

Boardman & Sasse (Boardman & Sasse, 2004) report on a 2004 study aimed to understand how people manage multiple sets of hierarchically organized personal information. Rather than look at a single type of information across machines, they looked across information types and investigated the category overlaps people have between their documents, email and web bookmarks. Their focus was on understanding the folder names and structures, rather than the files themselves. They noted that one user reported creating document folders to separate document versions; an action that isn't applicable to email or web bookmarks. More information is provided in (Boardman, 2004), in which they observed that 4% of the folder names used by their 25 participants showed some evidence of being used for version control (such as folders called 'version1' or 'old').

More recent studies such as (Bergman, Whittaker, Sanderson, Nachmias, & Ramamoorthy, 2010) have examined how the structure of folders impacts on the user's ability to successfully retrieve files, but have not looked at any of these other issues that users might face in document management. Others have noted that versioning is an issue, finding that one of the common information management strategies abandoned is a versioning strategy (Bruce, Wenning, Jones, Vinson, & Jones, 2010).

Current user interface support. The document management user interface used by the majority of people is the file system interface in their operating system. Making copies in these systems is a very easy task, accomplished from either a context menu or with a keyboard shortcut. The copy of the file has the same content, and all the same attributes as the original with two exceptions. It will have a different date of creation, and if

the copy was made within the same folder, it will have a different name, due to the requirement that each file have a unique name within a given folder. However, there is no link or association between a file and any copies it has generated. No systems provide an explicit means of locating copies, although most systems have a search function that can be used to find other files with the same or a similar name. Once two files with the same name are located, the operating system doesn't provide any means of comparing them to assess their similarity. The file details can be inspected to see whether they are the same size and the same date modified, but the user will need to open both files and compare the content to manually assess the similarity. Some applications (such as Microsoft Word) have a feature that will compare two documents and highlight the content differences between them. To show differences, it is necessary to have an intimate knowledge of the specific file formats involved, and therefore that isn't something that would normally be expected of a file system. However, to be able to simply identify whether or not two files are identical in content is something the file system could easily do. There are a number of third party applications that are designed to compare files (e.g. UltraCompare (IDM Computer Solutions, 2011) or ExamDiff (PrestoSoft, 2010)), as well as applications that are designed to detect duplicates (e.g. Duplicate Cleaner (DigitalVolcano, 2011) or Easy Duplicate Finder (EasyDuplicateFinder.com, 2011))

The most commonly used Desktop file systems are currently Windows XP (56.7%), Windows 7 (20.9%), Windows Vista (12.1%) and Mac OS X (4.5%) (NetMarketShare.com, 2011). None of these commonly used file systems provide any support for versioning. Users are free to make copies and develop their own versioning schemes based on either folder or file names. To provide support for this task, some commonly-used applications do have their own versioning support built-in. For instance, Microsoft Word and Google Docs Word Processor both have the ability to track changes, which can allow a user to return to earlier versions of a file.

There are a number of third-party version control systems which were developed for software development source code control (Ruparelia, 2010), and there are a number of articles and blog posts on the web that show people how to use these systems for personal document management (e.g. (TechRepublic, 2007)). However, version control systems such CVS, Subversion or Git are not the most user friendly systems, with many having a command line as the primary user interfaces. Their use as document management repositories is probably limited to very technically competent people (primarily software developers).

There are a number of systems that have been proposed by researchers to support versioning in a file system. These have primarily been for Linux systems, and include: Elephant File System (Santry, Feeley, Hutchinson, & Veitch, 1999), CVFS (Soules, Goodson, Strunk, & Ganger, 2003) and Wayback (Cornell, Dinda, Fabi\, \#225, & Bustamante, 2004). However, none of these suggested features have so far been integrated into current operating systems.

This review of the state of research and the state of the art in version and copy management has shown that this area has received little attention from researchers. The following section describes a study that (among other aims), explores the issues that individual knowledge workers experience surrounding versioning and duplication and the practices they employ to cope with this lack of explicit tool support.

METHOD

A two-phase study was conducted with the aim of investigating personal document management practices among knowledge workers. The first phase was in-depth interviews with 10 participants, followed by a survey with 113 respondents.

The interviews took place in the participant's office and they were encouraged to show their file system as well as talk about it. Interviews were transcribed and analysed to extract common themes. All participants were employees of a University, and were a mixture of academic and non-academic staff, of varying ages and seniority.

Drawing on the issues raised in the interviews, a questionnaire was developed and used to conduct a survey to gather information about document management practices from a larger group of people. An invitation to a web based survey was sent to 428 people and 113 responses were received. The questions were based on practices and behaviours noted by participants during the interviews.

In addition, 73 participants additionally provided a snapshot of their file system was taken in order to quantitatively analyse the structure of the files and folders. This snapshot was obtained using a tool custom-written for the purpose. The tool prompted participants to select the locations where they stored their documents (the Desktop and My Documents folders were selected by default but could easily be removed). The tool analysed the folder structure and stored the names and structure of folders, and the name, extension, size and modified date of the files.

The file system snapshots were checked to ensure that no system folders were included, and where some had been, these were removed. Additionally, system-created files such as backup logs were removed from the analysis.

RESULTS

This section reports on the results obtained in this study. We first report on the file and folder duplication results from both the questionnaire and the file system snapshot. Next, we report on the versioning observations from both the questionnaire and snapshot, and finally, we present some of the qualitative results obtained from the interviews and the free-form answers from the questionnaires.

File and Folder Duplication – Questionnaire Results

Participants were asked two questions about the issue of proliferating copies of files on their computers. First, they were asked whether they felt this was an issue for them at all. 47% of respondents report sometimes accidentally having more than one copy of the same document. These people who do duplicate documents are overall significantly less satisfied with their document management practices overall (f=16.66, sig=.000).

They participants who answer affirmatively were asked why they thought this happens. They were given four choices based on the most frequent descriptions from the interview participants, plus an 'other' option where they could supply their own reason. Fig. 1 shows the proportion of responses in each category.

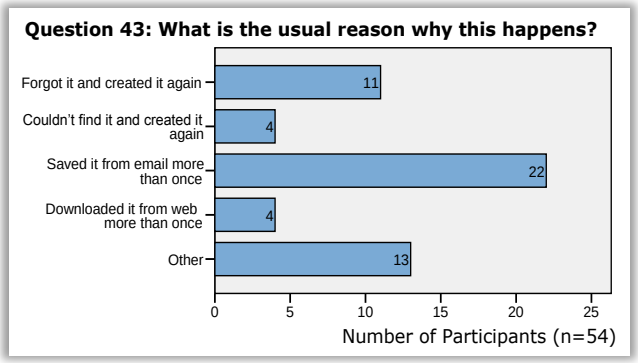

Figure 1. Bar graph showing common reasons given for accidental file duplication

The most common reason people think they have multiple documents is due to saving email attachments multiple times. However, forgetting a document existed and recreating it was considered the most common reason by 20% of respondents.

Of the 13 people who answered 'Other', four said they have duplicated files deliberately. Two say they have this situation because of having backup copies of a file. Two mentioned copying files to a different location to make them easier to upload into the university's learning management system. One person said that all the options apply, three said it arose because of transferring files between computers, and one said they thought they accidentally save the same file with a different name.

File and Folder Duplication – File System Snapshot Results

Duplication was measured by calculating the proportion of non-unique file names in the file system. A file or folder is considered to be a duplicate if another file or folder exists with the same name anywhere else in the file system. As the file system snapshot tool only records file names and not file contents, it is not possible to be sure that the two files are genuine duplicates or near-duplicates. There are legitimate reasons why two files or folders that are not the same may have the same name and these would be erroneously counted as being duplicates when in fact they are not. Likewise, an even more insidious form of duplication occurs when two files have the same content but have different names. However, this analysis does at least provide some indication of the level of duplication in file systems.

The mean level of file duplication was 21.8%. This means that on average, 21.8% of the documents in the file system have the same name as another file. The variation in file duplication across the sample was quite striking, ranging from 0.4% to 60.4% (shown in Fig. 2). The level of folder name duplication was slightly higher, with a mean of 23.5%, and ranging from 0 to 73.4% (shown in Fig. 3).

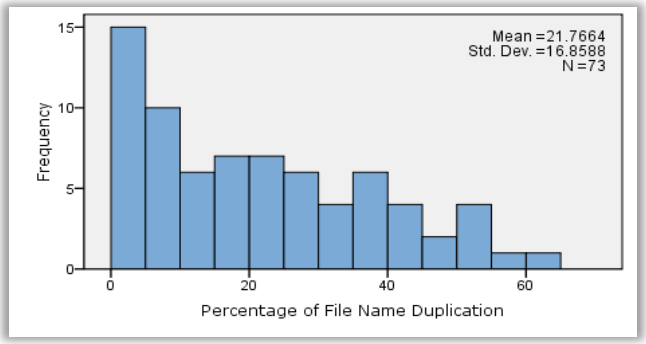

Figure 2. Histogram showing distribution of proportion of file name duplication.

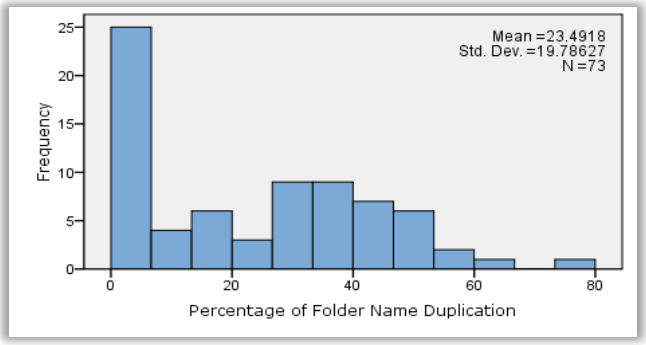

Figure 3. Histogram showing distribution of proportion of folder name duplication

There is a significant correlation between the level of folder and file name duplication (r = 0.79). One likely explanation for this is that entire folders and their contents are frequently being duplicated together.

The participants are fairly evenly split in whether the duplication is higher in files or folders, with 48% having higher folder duplication than file duplication. On average, the amount of folder duplication is about 2% higher than the amount of file duplication. However, the range is quite wide, with the participant at one extreme having 34% more file duplication than folder duplication, and the participant at the other end of the spectrum having 33% more folder duplication than file duplication.

The amount of duplication is related to the overall size of the file system. The level of file name duplication is correlated to the total number of files (r = 0.61) and the level of folder name duplication is correlated with the total number of folders (r = 0.65). Therefore, the more folders and files a person has, the more they are likely to have duplicates.

There is no significant correlation between the levels of file and folder duplication and the width of the file system. However, there is a significant correlation between the average depth of the file system and the level of file name duplication (r = 0.59) and folder name duplication (r = 0.71). One possible explanation is that people with deep file systems are more likely to have repeating groups of folders and files (which perhaps are differentiated with a high level folder name).

The level of duplication can also be examined within specific locations as well as across the file system as a whole. On the Desktop (and its subfolders), the average level of folder duplication is only 2%, and the level of file duplication is 9.7%. Within the My Documents folder, the folder duplication is 14% but the file name duplication is only 2.2%. Within other locations (network drives, flash memory, other C: drive folders), the folder name duplication was 10.5%, and file name duplication is 17.6%.

All of these within-location levels of duplication are lower than the average level of duplication across the whole file system, indicating that files and folders are being duplicated across locations. This can be done deliberately for a number of reasons, including portability and backup purposes, as well as accidentally, when transferring files from one location to another.

File Versioning – Questionnaire Results

A separate section of the questionnaire asked about issues with versions. Respondents were asked whether they sometimes use separate files for different document versions and 83% responded affirmatively. Those 97 respondents who version files were then asked how they distinguish between the different versions. The results are shown in Figure 4.

Of those 97 participants who have multiple files for version of documents, the majority (57%) say they distinguish between them using version numbers in the file name. The next most common option is to use dates to differentiate the version, with descriptions or folders being used by only 5 respondents each.

The 6 respondents who chose 'Other' all indicated that they would use a combination of these methods. 2 people said they use all of these methods, a further 2 says they use all of the first three, 1 person uses the first two, and 1 uses the first two plus also the name of the person who changed the file. In addition, one respondent who chose dates added the extra information that they used either dates or semesters. And one respondent who said they put the files in another folder added "*I actually combine the above and put the version name in the file name AND put them in a separate folder (usually labelled 'Old Whatever')*"

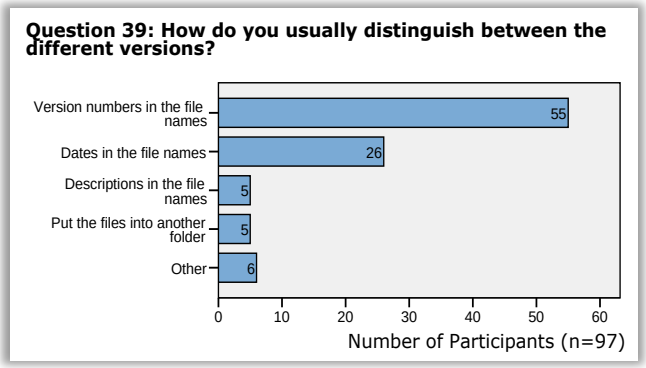

Figure 4. Bar graph showing how participants distinguish between file versions.

The respondents were also asked whether they sometimes lose track of which document is the most recent version. 42% answered affirmatively. The 42% of respondents who report they sometimes lose track of which document is the most recent version were significantly less satisfied with their document management overall (f=22.11, sig=.000).

These people who sometimes lose track of which file is the current version were then asked why they think this might happen. No options were given to the participants, just a free text response field.

The most common reason, reported by nine people was that they have no systematic way of identifying versions, or at least, no consistent way of doing it. One mentioned not being able to store dates in the name of the file, which is due to the slashes in the most common short date format being prohibited in file names. Five people didn't specifically mention that it was a systematic problem, but just said that their files weren't named well enough for them to be able to tell the version from the file name.

The next most common reason was that they forget to add the version identifiers, with a couple noting that it is particularly a problem when you come back to a document after a few days. On the same theme, some mentioned that they were too busy or too lazy or just didn't take the time to assign proper identifiers, and two simply said it was carelessness

Four people identified collaboration as a problem, saying that other people gave documents different names, used different version identifiers or forgot to change version identifiers.

Three people fingered multiple locations as the problem, saying for instance that they "*transfer documents between computers and forget where I last worked on the file*." Three report running into problems because different versions are stored in different folders on the same

computer, which can cause them to overlook the most recent one.

One person didn't know, one said they mistakenly put the wrong version numbers on a file, one person mentioned problems with email attachments being saved in temporary locations, and another said they had accidentally overwritten a version of a file.

Academics are more likely to have multiple files for different versions of a document (Chi-square=4.15, sig=0.042). 90.3% of academics have multiple files for versions, versus 76.2% of non-academic staff. Academics are also more likely to accidentally have multiple copies of a document (Chi-square=5.25, sig=0.022), with 55.6% of academics experiencing this compared to only 33.3% of non-academic staff.

File Versioning - File System Snapshot Results

The 72 participants who provided a file system snapshot have a total of 584,162 files in their file systems. Two participants were found to have several versions of a very large web help library amongst their files, which accounted for over 50,000 files. These were removed from the analysis. For the purposes of this analysis, it was decided to focus on the most common file types: Microsoft Word documents, PDF Files, Microsoft Excel Files, and Microsoft PowerPoint Files. Together these amounted to 191,863 files. The most common files that were left out were image files, since these were primarily photos and are rarely versioned.

The file names were scanned to look for particular character sequences that may indicate versioning has occurred. We tried to be as conservative as possible in our matches, erring on the side of overlooking versioned documents rather than including too many. However, the process is necessarily approximate.

The following list describes the different kinds of versioning were searched for, and the numbers found for each technique are listed in Table 1 below.

Versioning using the word 'version'. This can either be preceded or followed by a description, date or numeric version indicator (e.g. "final version", "submitted version", "version 5"). In order to avoid counting file names that contain words like diversion or conversion, the string searched for was 'version' (with either a preceding space, hyphen, period or underscore). Examples of matched files were: 'AMA Version 2.doc', 'OM Part A version 8.ppt' and 'PhDThesis-version196.doc'.

Versioning using the letter 'v'. Another common technique for versioning is to use the letter v followed by a numeric identifier: e.g. v1, v5, v23. The pattern searched for was a "v" followed by any digit. Optionally, there could also be a space, period, underscore or hyphen between the v and the digit. This eliminates matches for files with names like nov02 or cv2002. Examples of matched files were: 'proposal v.3.doc', 'Product Development Strategy v 1.4.pdf' and 'Chapter One V4.doc'.

Versioning using years. Many participants reported using dates in their version identifiers. There are many different ways of matching dates, so years were checked first. The pattern searched for was any 4 digit sequence that began with either '19' or '20'. There was no match if there were additional digits either side of the sequence, to eliminate files named with ID numbers that just happened to contain a 19 or 20. This identifies 4 digit years, but misses any years that are coded with only two digits. Examples of matched files were: 'Retreat Agenda 1999.doc', 'CV2003.pdf' and 'Exam Details 2005.doc'.

Versioning using months. The pattern searched for was either the full month name, or the 3 letter month abbreviation, provided it wasn't followed by any other letters. This is to eliminate matches for things like 'Mark' or 'Julie'.. Examples of matched files were: 'Notes from Meeting 11 June.doc', 'Expenses March.xls' and 'ReportDec09.doc'. 37% of the files versioned with a month name also included a 4-digit year. A further 59% of month-versioned files also included a two digit identifier. The majority of these are likely to be two digit year identifiers, but some could be day identifiers.

Versioning using dates. The pattern searched for was a three part numeric date, for instance 15.06.2004, 6-4-98 or 2008_06_6. Valid separators included spaces, hyphens, periods or underscores. Either one or two digits were acceptable for day and month, either two or four digits for the year, and the year could go either at the beginning or the end. Examples of matched files were: 'Annual Report_20_1_04.doc', 'Seminar 28.2.2005.ppt' and 'Timesheet 17 3 2002.xls'.

Versioning using 'final' or 'old' labels. Labels such as 'old' and 'final' can be used to indicate versions. The pattern searched for was 'old' anywhere in the file name, and final only if it appeared at the end of the file. This is to eliminate the many matches for 'Final Exam' that appear in academic file systems. Final was by far the most common identifier, accounting for over 90% of this category. Examples of matched files were: 'Questionnaire_final.doc', 'Project report FINAL FINAL.pdf' and 'Old version of template.doc'.

Versioning using numbers. Many files included numbers in the name, however it is impossible to determine automatically whether these numbers represent a version identifier. For instance, there are many files with names like 'Chapter1.doc', 'Chapter2.doc', which are likely to be different files. It is very difficult to distinguish these sorts of files from files which might be versioned such as 'FinalReport-1.doc', 'FinalReport-2.doc'. There were 49,179 files (25.6%) that don't fit into any of the above categories and that have numbers at the end of the name. However, these are not included as versioned files in the

table below since without human judgment, it is impossible to know how many are actually versioned.

Table 1 below summarizes these results. Overall, 71 out of the 73 participants used some form of versioning scheme, and a total of 24.3% of documents in the entire collection have some form of versioning. The majority (55%) of participants used all 6 of the different versioning schemes at least once in their document collection ('version', 'v', year, month, date and labels). A further 16% used 5 of the 6, and only one participant did not use any of these schemes.

Versioning Scheme	Partici-pants	Partici-pants %	Files	Files %	Avg Files	Std Dev
'version'	48	66%	3162	1.65%	65	342
'v'	56	78%	2270	1.18%	40	82
Year	72	99%	29410	15.30%	408	651
Month	70	97%	13403	6.99%	157	299
Numeric Date	55	75%	2602	1.36%	47	124
Final or Old	60	83%	2087	1.09%	34	86
Year and Month	65	89%	4158	2.16%	63	103

Table 1. Number of participants and number of files using different file versioning schemes. There are 73 participants and 191,863 files in total.

Folder name versioning. The above analysis can be repeated using folder names. There are a total of 85,311 folders in the document collections of the 72 participants. System-named folders (including My Document, My Pictures, _vti_cnf, etc) were excluded, as were all the folders containing the web help library identified earlier. This left 45,604 folders in total. The patterns searched for were the same as detailed above. The results are shown in Table 2.

Versioning Scheme	Partici-pants	Partici-pants %	Folders	Folders %
'version'	20	27%	65	0.14%
'v'	6	8%	40	0.09%
Year	64	88%	3044	6.67%
Month	40	55%	750	1.64%
Numeric Date	17	23%	214	0.47%
Final or Old	37	51%	179	0.39%
Year and Month	31	42%	239	0.52%

Table 2. Number of participants and number of folders using different folder versioning schemes. There are 73 participants and 45,604 folders in total.

Overall, 65 of the 73 participants used some kind of version identifier in their folder names, and version identifiers featured in 8.9% of folder names. Most of the participants used either 3 different types of version identifiers (31.5%), just one type (17.8%) or 2 types (13.7%). Three participants used all 6 different types.

Qualitative Results

Almost all the interview participants kept multiple versions of documents in separate files and added identifiers to the file name in order to distinguish between versions of the document, and in particular, so they would know which file represented the most recent version of the document.

Version Management. The most common method of version identifiers that the interview participants demonstrated was to append a **number** to the file name, as Participant B[1] explains: "*Normally, they'll share the same name followed by an underscore and a sequential number, 01, 02 and so on*." Participant B has many versions of some of his documents: "*something like the thesis, that went up into like a couple of hundred versions of that. Coursebooks generally have a half dozen versions. Exams normally have a half, maybe a dozen versions*." Because of his sequential numbering, he normally doesn't have any difficulty in figuring out which version is the most recent. Participants D, E, G and J all reported using the same system. Participant D also noted that sometimes the final version in the numeric sequence would be given a different version identifier to indicate it was the final product: "*I'll have a final version which is the production one. And I call that final version as well so that I don't get mixed up*."

Participant E also noted that he applied version numbers after the fact if he discovered two copies of the same file in different locations and wasn't sure whether they were the same or not: "*Normally I just check the date if they are the same or not, or the file size, they are the same or not. But very frequent they are different, by very minor changes of the date or the file size. And sometimes I keep them all there, and I name it 1, yes for example myfile1 and myfile0 depending on the date. So the zero one is the earliest date one, because at that moment, when I do the file reorganisation, I don't want to take too much time to look into detail, but I don't want to destroy the other one too. So I don't want to use different name, so I just use a little bit different name, but I still put them there, just in case I do need them at the same time.*"

Adding a **date** to the version file names was also very common, with Participants B, C and I all reporting doing this with some of their documents. Participant I describes her process: "*So this is a paper that I've just been working on recently. It's a revision so I've got them all dated. [Topic] for this journal and a certain date. So we've got February, 14, 16, 21 March and then this is the response letter that I've been working on. I will get rid of all the old stuff once the paper is in print. But until it's in print, it all stays*."

Participants C and I noted that they often have both people's **names and dates** in documents that they are collaboratively working on. Participant I describes this: "*We get into a rhythm, if I use [Alice] for example, she has her own way of labelling, as I have my own way of labelling, so I might keep portions of what she has and then*

[1] Participant's identities are kept confidential. The 10 participants are identified with letters A through J.

I'll have 'cb[me]' meaning changes by [me] and then give the date. So she'll recognise her file but it's been changed by me." When working with her colleagues, they all usually add their initials to the file name "*so we all know who we're talking about when we're passing stuff around*." Different conventions apply to different collaborators: "*I have a colleague in Scotland who we don't use our names anymore, we're ping and pong. Because this paper we were working on was just going ping pong ping pong and we now, she's ping and I'm pong and that's the way we do everything*."

Participants A, F and H said they didn't keep versions. Participant F says "*it's something I've toyed with, but life is complicated enough*". Participant A uses the track changes feature in Word for versioning, while H periodically backs up all her files and prints important documents, so she maintains a version history that way while only keeping one file.

File duplication. Four participants reported having problems resulting from multiple copies of the same document (excluding backups).

Participant A reports sometimes ending up with multiple copies of the same file in different locations: "*what happens is I generate a document, it ends up in one of my cleanup folders, and I can't find it, or I forget that I had it there. I generate it again by whatever means, and that ends up in another cleanup folder, so I've got two copies of the same document lying around*." He adds "*it is also likely that a file on my Desktop is in My Documents as well, but they may be out of sync, so I may have created it in My Documents, copied it over to my Desktop so I can work on it and not synched back to My Documents*." He says he rarely detects that sort of thing, and if he notices in a search that he has multiple copies, he will compare dates and use the most recent document.

Participant D ran into trouble after trying to integrate files on his USB flash drive with his Desktop: "*I keep my main copy on here now [USB drive], but then I copy across, but then I changed stuff and then I forgot to delete this one first [the copy on the hard drive], so it's double copied things because I changed the file names, which is bloody annoying*."

Participant E has fairly often ended up with multiple copies of the same file. He says "*just in case that I lose some files in somewhere, most of time I keep three copies, or even four copies. For example I have a memory sticker [USB drive], I keep some file there and also here [Desktop], and also my laptop and my home computer. Sometimes it's a problem because there are too many different kind of versions, so different kind of copies*." He tries to avoid problems by trying to immediately synchronise any files he changes, but he doesn't always remember to do so.

Participant F is very aware of this problem: *"There's always a problem when you have multiple copies of a file, and you inadvertently do an edit on what in fact is an earlier version, and somewhere else on your many disk drives there's a later version that is really the one you should have done the latest edit on. It's not a good idea to have multiple copies of a file."* To avoid this issue, he has a special tool to keep his document collections synchronised, which he runs periodically to make sure all his locations are updated with the latest version. It notifies him if there are any problems synchronising and gives him choices about which files to keep and which to overwrite. He does very occasionally run into problems after having independently edited more than one copy of the same file: "*when the situation arises, more likely with a current document where I have inadvertently edited the wrong disk's version of it. And then a day later, I can't remember quite which of the versions I was editing. I may have to open up all of the versions to find out where they are and open them all up to make sure which is the one that I most recently changed*."

5 DISCUSSION AND CONCLUSION

Document duplication

The field study indicated that duplicated documents are an issue for many people, with 47% of survey respondents reporting this as a potential problem. This was especially true with files shared between multiple computers or storage media. The survey confirmed this issue; with 47% of respondents have this problem. On average, around 20% of participant's files are duplicates of files that exist elsewhere in their file system, however, these tend to be most frequently duplicated between different locations. For instance, a user might have some of their local files duplicated onto a network file share or a USB drive.

Survey respondent indicated that some of the main reasons for this duplication were due to multiple saving of email attachments, or forgetting they had a file and creating it again. Other reasons included deliberately making copies, for backup or archive purposes, to put a file in a location where it was easier to upload into online systems, and because of files being transferred between computers.

Suggested document management features to prevent document duplication

In order to support the user to prevent and manage duplication, the file system needs to identify duplication when it occurs. For instance, whenever a file is copied onto the hard drive, it should be checked against the files that already exist to check whether it is a duplicate. Current Operating Systems will prevent files with the same name being copied into the same folder, but it needs to handle the situations where the file is copied into a different folder, and also situations where the file contents are the same but the name is different. Identification of identical content is possible using the mathematical technique of hashing, which allow a shorter 'signature' of a file to be computed. These cryptographic algorithms ensure that each content generates a unique signature that can reliably be used to identify unique documents (Mead, 2006).

Once the system has identified that a file being copied is a duplicate of a file elsewhere in the system, there are two user interface strategies that can be deployed: pre-emptively alerting the user, or subtly notifying the user. Using the first strategy, the user will be immediately alerted that they are copying a duplicate, informed of the location of the other copy and asked how they wish to resolve the situation. One option for resolution is by keeping the document in both locations in the file system. Another is to keep the document in one location in the file system. In this case, the user can select which location the file should remain in, and the duplicate will be removed from the other folder. A third option is to keep the document in one location and provide a shortcut in the other location.

A less intrusive method involves allowing the action to take place that creates a duplicate, but then visually tagging the duplicate so that the user can identify it as such. This can take the form of an icon overlay, similar to the shortcut icon overlay currently used in Windows operating system. The system can then provide information about the duplication (perhaps via a context menu, or properties panel), showing the location of any duplicates, and give the user the same options as above to resolve the duplication.

Document versioning

Managing multiple versions of documents in separate files is a very common practice and a source of many problems since there is no system support for this. People come up with a range of possible versioning schemes, and are not necessarily consistent in their use. The survey confirms that 80% of participants use multiple files for versions, and that 42% report sometimes losing track of the current version. Needless to say, this causes them to spend extra time opening files to check contents, or even having to redo work, and therefore they are less satisfied about the personal document management system because of it. Those that report losing track indicate that they think that either the fact that they have no systematic way of versioning is responsible, or simply that they forget or are too busy to do it.

By far the most common versioning scheme reported in the survey is to include a version number in the file name (a practice adopted by 57% of the respondents). Although we couldn't unambiguously identify these files in the file system snapshot results, we did note that all participants had files with names ending in numbers, and fully 25% of the documents analyzed fit this pattern. Only a few people reported using descriptions or moving old or current versions to another folder, which was confirmed by the file system snapshot analysis. While only a few people reported in the survey that they used a combination of these techniques, the file system analysis shows that the majority of people do use almost all techniques somewhere in their file system. However, it is likely that most people have a dominant scheme that they use most of the time.

The second most common scheme reported in the survey and observed in the file system snapshots was to use dates as version identifiers. Because there are so many different date formats, it is difficult to be sure exactly how widespread this is, but over 15% of files included a year in the title and almost 7% had a month name or abbreviation. Years were also very common in folder names. It is particularly Interesting that so many people put dates in their folder and file names, since all that is required for versioning is a sequence identifier, and all files are date stamped by the file system. The file system stamps each file with the date it was created, the date it was last modified and the date it was last accessed.

Does the prevalence of manually encoded dates in file names indicate a lack of trust in the file system's date stamping mechanism? In some instances, the answer is probably yes, since there are many actions which interfere with these dates. Copying a file from one drive to another or one device to another can reset the dates, which means they cannot be relied upon as indicators of when the file was actually created or last changed. Another issue is that not all the dates in the filenames necessarily indicate a version of a document that was last edited on that date. A document about a meeting held in May could be labelled with May in the name, and still be being edited a month or two later.

Suggested document management features to support document versioning

Users should not have to manually create and manage multiple files for multiple versions of a document - the file system should handle that automatically and invisibly. Every time a file is saved, it should automatically create a new version with the current information. The file should be visually tagged in some way to indicate that a version history is available, perhaps with an icon overlay similar to that suggested above for tagging duplicates. The document management interface should always show only the most recent version of a file, but the previous versions should be accessible on demand. It is not necessary that they be easily or quickly accessible, since going back to previous versions is not a common activity, but it should be possible to do so when necessary. This should eliminate one major source of version problems.

Keeping a potentially infinite number of saved versions possibly raises some storage issues, even though hard drive sizes are continually increasing. The system should provide a feature for automatically purging old versions when disk space makes it necessary, perhaps in a similar way to the Windows Disk Cleanup utility which prompts users to select and delete files that are no longer needed. This can be configurable to remove versions older than a certain date, or to keep only the most recent 2 or 3 versions of a document.

Another useful feature is that it should be possible to mark a document or folder as being locked or archived, which essentially makes the document read only. Views of the

document should visually represent this, with either an icon overlay or perhaps a colour change to show that the document is no longer active. This provides both versioning support and some support for task management, since people can essentially mark their document as being complete. This feature could possibly also be an opportunity to purge earlier versions of the document in order to recover disk space.

CONCLUSION

This paper has found that issues of document duplication and versioning have not been studied in the research literature to date. Current file system user interfaces do not provide any automated support to users trying to minimise duplicates and manage versions. We conducted a study involving 10 interviews, 113 questionnaires and 73 file system snapshots, which shows that these two issues potentially affect significant numbers of people, and therefore are issues that deserve to have some attention paid to them by researchers. We have provided some empirical evidence on the scale of the problem, as well as both quantitative and qualitative descriptions of the ways in which users currently address these problems. From this, we suggest some improvements to the usability of personal document management systems that will hopefully result in increased productivity for those who use them.

REFERENCES

Asprey, L., & Middleton, M. (2008). Integrated Document Management For Decision Support. In F. Burstein & C. W. Holsapple (Eds.), *Handbook on Decision Support Systems* (Vol. 1, pp. 191-206). Berlin: SpringerLink.

Backer, A., & Busbach, U. (1996, 3-6 Jan 1996). *DocMan: a document management system for cooperation support.* Paper presented at the Proceedings of the Twenty-Ninth Hawaii International Conference on System Sciences.

Barreau, D. K. (1995). Context as a Factor in Personal Information Management Systems. *Journal of the American Society for Information Science, 46*(5), 327-339.

Barreau, D. K., & Nardi, B. A. (1995). Finding and Reminding: File Organization from the Desktop. *SIGCHI Bulletin, 27*(3), 39-43.

Bergman, O., Whittaker, S., Sanderson, M., Nachmias, R., & Ramamoorthy, A. (2010). The effect of folder structure on personal file navigation. *Journal of the American Society for Information Science and Technology, 61*(12), 2426-2441. doi: 10.1002/asi.v61:12

Boardman, R. (2004). *Improving Tool Support for Personal Information Management.* Doctoral Dissertation, Imperial College, London, England.

Boardman, R., & Sasse, M. A. (2004, April 5-8, 2004). *"Stuff Goes into the Computer and Doesn't Come Out" A Cross-tool Study of Personal Information Management.* Paper presented at the CHI'2004 Conference on Human Factors in Computing Systems, Vienna, Austria.

Bruce, H., Wenning, A., Jones, E., Vinson, J., & Jones, W. (2010). *Seeking an ideal solution to the management of personal information collections*. Paper presented at the ISIC 2010 Information Seeking in Context Conference, Murcia, Spain.

Cornell, B., Dinda, P. A., Fabi\, \#225, & Bustamante, n. E. (2004). *Wayback: a user-level versioning file system for linux*. Paper presented at the Proceedings of the annual conference on USENIX Annual Technical Conference, Boston, MA.

DigitalVolcano. (2011). Duplicate Cleaner. Retrieved from http://www.digitalvolcano.co.uk/content/duplicate-cleaner

EasyDuplicateFinder.com. (2011). Easy Duplicate Finder. Retrieved from http://www.easyduplicatefinder.com/

Equivio. (2006). Equivio Announces Version 2.0 of Near-Duplicate Detection Software. *Press Release*. Retrieved from

IDM Computer Solutions. (2011). UltraCompare (Version 8). Retrieved from http://www.ultraedit.com/products/ultracompare.html

Mead, S. (2006). Unique file identification in the National Software Reference Library. *Digital Investigation, 3*(3), 138-150. doi: DOI: 10.1016/j.diin.2006.08.010

NetMarketShare.com. (2011). Operating System Market Share, from http://www.netmarketshare.com/operating-system-market-share.aspx?qprid=10

PrestoSoft. (2010). ExamDiff (Version 1.8). Retrieved from http://www.prestosoft.com/edp_examdiff.asp

Rama, J., & Bishop, J. (2006). *A survey and comparison of CSCW groupware applications*. Paper presented at the SAICSIT '06 - The 2006 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries, Somerset West, South Africa.

Ruparelia, N. B. (2010). The history of version control. *SIGSOFT Software Engineering Notes, 35*(1), 5-9. doi: 10.1145/1668862.1668876

Santry, D. J., Feeley, M. J., Hutchinson, N. C., & Veitch, A. C. (1999, 1999). *Elephant: the file system that never forgets.* Paper presented at the Proceedings of the Seventh Workshop on Hot Topics in Operating Systems, Rio Rico, AZ , USA.

Soules, C. A. N., Goodson, G. R., Strunk, J. D., & Ganger, G. R. (2003). *Metadata Efficiency in Versioning File Systems*. Paper presented at the Proceedings of the 2nd USENIX Conference on File and Storage Technologies, San Francisco, CA.

Sprague, R. H., Jr. (1995). Electronic Document Management: Challenges and Opportunities for Information Systems Managers. *MIS Quarterly, 19*(1), 29-49.

TechRepublic. (2007). Use open source Subversion for personal document management. Retrieved from http://www.techrepublic.com/article/use-open-source-subversion-for-personal-document-management/6167205