Organizing and Managing Personal Electronic Files: A Mechanical Engineer's Perspective

B. J. HICKS
University of Bath
A. DONG
University of Sydney
and
R. PALMER and H. C. MCALPINE
University of Bath

This article deals with the organization and management of the computer files handled by mechanical engineers on their personal computers. In engineering organizations, a wide variety of electronic files (documents) are necessary to support both business processes and the activities of design and manufacture. Whilst a large number of files and hence information is formally archived, a significant amount of additional information and knowledge resides in electronic files on personal computers. The widespread use of these personal information stores means that all information is retained. However, its reuse is problematic for all but the individual as a result of the naming and organization of the files. To begin to address this issue, a study of the use and current practices for managing personal electronic files is described. The study considers the fundamental classes of files handled by engineers and analyses the organization of these files across the personal computers of 40 participants. The study involves a questionnaire and an electronic audit. The results of these qualitative and quantitative elements are used to elicit an understanding of the practices and requirements of engineers for managing personal electronic files. A potential scheme for naming and organizing personal electronic files is discussed as one possible way to satisfy these requirements. The aim of the scheme is to balance the personal nature of data storage with the need for personal records to be shared with others to support knowledge reuse in engineering organizations.

This research was supported in part by the EPSRC Innovative Design and Manufacturing Research Centre at the University of Bath, grant number GR/R67507/0.

Authors' addresses: B. J. Hicks, Innovative Design and Manufacturing Research Centre, Department of Mechanical Engineering, University of Bath, Bath, BA2 7AY, United Kingdom; email: b.j.hicks@bath.ac.uk; A. Dong, Design Lab, Faculty of Architecture, Design and Planning, University of Sydney, Sydney, 2006, Australia; email: andy.dong@usyd.edu.au; R. Palmer, Department of Mechanical Engineering, University of Bath, Bath, BA2 7AY, United Kingdom; HC McAlpine, Innovative Design and Manufacturing Research Centre, Department of Mechanical Engineering, University of Bath, Bath, BA2 7AY, United Kingdom.

© 2008 ACM 1046-8188/2008/09-ART23 $5.00 DOI 10.1145/1402256.1402262 http://doi.acm.org/10.1145/1402256.1402262

Although this article is concerned with mechanical engineers, the issues dealt with are relevant to knowledge-based industries and, in particular, teams of knowledge workers.

Categories and Subject Descriptors: E.5 [**Data**]: Files—*Organization / Structure, sorting / searching*; H.3.2 [**Information Systems**]: Information Storage—*File organization*; J.6 [**Computer Applications**]: Computer-Aided Engineering; K.7.m [**The Computing Profession**]: Codes of Good Practice

General Terms: Documentation, Human Factors, Management, Standardization

Additional Key Words and Phrases: Engineers, information management, directory and file naming conventions, file sharing and file recognition and recall

ACM Reference Format:

Hicks, B. J., Dong, A., Palmer, R., and McAlpine, H. C. 2008. Organizing and managing personal electronic files: A mechanical engineer's perspective. ACM Trans. Inform. Syst. 26, 4, Article 23 (September 2008), 40 pages. DOI = 10.1145/1402256.1402262 http://doi.acm.org/10.1145/1402256.1402262

1. INTRODUCTION

The importance of information and knowledge provision for organizations in today's highly regulated, fast-moving, dynamic markets is widely acknowledged [Stewart 1997]. Because of this, considerable work has been undertaken in the areas of design, implementation, and integration of information management systems to support business processes [Curtis & Cobham 2005; Jessup 2008]. In addition to these formal enterprise-wide systems, a wealth of information is also created and stored as electronic files. Some of these files may be stored and accessed by virtue of specialized business systems, such as product data management systems and document management systems. However, a vast number will reside within shared and personal file spaces of employees and as a consequence their access and reuse is problematic for all but the individual that created them.

For the purpose of this paper, the term *electronic file* is used interchangeably with computer-based file and encompasses the continuum of files—or what Jones [2007] defines as an *information item*—used to represent data generated by users of software applications. For example, in a typical medium-sized engineering organization, there may be hundreds of users working in different departments, on various projects and in different roles. Like their counterparts in the architecture firms studied by Schmidt and Wagner [2004], engineers use these documents to "order" and facilitate their cooperative and individual work. All of these individuals will use a variety of different software applications and information systems to support their activities and ultimately acquire, generate and manipulate the information necessary to support the core competencies of the organization, and its operational and strategic management [Chaffey and Wood 2004; Dietel 2000].

In a recent study by Berkeley's School of Information Management and Systems, it was revealed that, in 2002, around 1.986 million terabytes of information was created and stored on hard disks worldwide, more than twice that stored in 1999 (0.926 million terabytes). The study also revealed that about 800 megabytes (Mb) of recorded information is produced per individual each year [Lyman and Varian 2003] or just over 2Mb per day. Similarly, recent data on the growth of personal electronic files shows that users scale their capacity needs for storing electronic files as storage capacity increases [Agrawal et al. 2007]. Furthermore, since Barreau and Nardi [1995] published their study indicating that users archive relatively little information, factors such as the decreasing cost of digital storage, dramatic increases in storage capacity, and improvements in search technologies, make users of today more inclined to archive information. As a consequence, the amount of information both produced and stored increases cumulatively with respect to time, negatively impacting on the users' ability to locate their own personal electronic files, let alone those of others. This is supported by recently published studies which highlight that users still "consider it difficult to find a distinct piece of information within their own personal information space" [Ravasio et al. 2004, pp. 158].

This ever-increasing volume of information is perpetuated further by the wide variety, diversity and number of sources, data formats and tools for generating and accessing information. This is particularly the case in the engineering sector, where there are a vast array of both commercial-off-the-shelf and bespoke software tools. These include simulation environments, data management tools, systems for managing customer and supplier relationships and electronic data exchange (EDI) [Culley and McMahon 2005; Hicks et al. 2002; Ward 2001]. The consequences of the this diversity of information-producing tools impacts not only on the amount of information that needs to be managed—which in itself contributes to information overload [Edmunds and Morris, 2000; Stewart, 1994]—but also on the provision for the storage and organization of the files generated by such software systems. Whilst these systems generally create similar types of file formats such as text files, word processing files, drawing files, spreadsheets, software code, and video and audio files, they all impose different requirements for their organization and access, which contributes to the problem of "information fragmentation" [Karger and Jones 2006]. In the case of engineering, these assumptions are generally based on whether the data should be organized according to the design process or according to the design artifact (i.e., the product) [Regli et al. 2000]. These different assumptions are built into commercial product data management systems (PDM) for computer-aided design (CAD) files and associated documents and product-life cycle management (PLM) systems. In addition, PLM systems manage engineering data according to a business strategy framework rather than the process- or product-centric view of PDM systems [Rachuri et al. 2007].

Despite the commercial availability of these systems, a large number of electronic files remain within an engineer's personal file space and directory structure. It is the nature of engineering work that causes engineers to retain a significant amount of project information and knowledge in personal electronic files. This characteristic is not uncommon and can be attributed to a range of reasons. For example, Whittaker and Hirschberg [2001] find that when it comes to paper documents, people want to keep personal copies for cognitive and emotional reasons; this characteristic is likely to extend into the digital domain.

The issues concerning the organization of personal file spaces becomes particularly germane in engineering organizations, where much of the information stored within electronic files is fundamental for the activities of design and manufacture. A study by Anderson et al. [2001] reported that 60% of the aerospace scientists and engineers they studied referred to their personal store of information first before consulting others in their department. Increasingly, the personal collections exist in digital form. Paper-based copies no longer hold much relevance for the 3D CAD models and computational simulations that typify engineering design work. In contrast to the findings by Barreau and Nardi [1997], the nature of engineering work demands that engineers keep their notes of past projects as they could contain useful knowledge that can be deployed in current projects. For example, drawing files, analysis models, meeting minutes, service reports, and technical reports are all critical for effective product design. Engineers seek out previous engineered works in the context of the current design problem to provide knowledge for describing and analyzing the current design work and to explain and predict the nature of the new design work. The integration of information from electronic files drawn from various projects constitutes knowledge in context which allows engineers to make use of personal memory or past experiences in the current engineering project. [Coughlan and Johnson 2008; Demian and Fruchter 2006] The data in these electronic files constitute cases which assist in engineering problem-solving by analogy through case-based reasoning [Maher and Gómez de Silva Garza 1997]. In engineering, the electronic files can be thought of as parts of design cases, which are indexed, recalled, and then adapted to suit the current engineering design problem. The benefits of referring to knowledge from different projects to support their highly technical work produce an imperative for engineers to record their work as reference for future projects. For these reasons it follows that the identification and recall of relevant personal electronic files is an important task for engineers. Furthermore, in the modern engineering environment, where projects involve large distributed teams, there is an increasing need to be able to share not only formal information such as CAD models and component data but also informal information [McMahon et al. 1993] such as personal notes and logbooks [McAlpine et al. 2006].

The difficulties engineers face with information access and reuse are widely reported [Szykman et al. 2000; Ward 2001; Hicks et al. 2002; Clarkson and Eckert, 2005]. Engineers also report that information gathering is their most frustrating task [Crabtree et al. 1997] and around 20–40% of their time is spent searching for and accessing information [Culley et al. 1992]. Despite this considerable effort, engineers frequently experience delays associated with information acquisition and access ranging from a day to as much as a year [Crabtree et al. 1997]. As a consequence, engineers, like any other knowledge worker, create and use personal information stores. However, although potentially valuable, these stores also present a variety of challenges for their effective reuse [Boardmann and Sasse 2004] and sharing [Erickson 2006].

It is this need to improve information and knowledge reuse with engineering organizations and the important role of personal information in the organizational knowledge base [Groth and Eklundh 2006] that are addressed in this study. In particular, the use and the current practices of mechanical engineers for managing electronic files are examined. The study considers the fundamental classes of file handled by engineers and analyses the organization of these files across their personal computers. The article first describes the research method and summarizes the development of a computer based audit tool. The results of the qualitative and quantitative elements are discussed in detail and used to develop a strategy or convention for naming and organizing directories and files, and archiving and backing up file spaces within the context of engineering. The proposed strategy is presented as a potential remedy to the issues of navigation around the file space of other users and ultimately to support more effective reuse across the organization.

2. RESEARCH METHODOLOGY

To investigate the personal electronic file management practices of engineers, forty participants were studied in detail. In order to evaluate their current practices, it is first necessary to define the various aspects of electronic file management and the dimensions considered in this study. In particular, it is necessary to collect both qualitative data about the participants and quantitative data concerning their personal computers and file space. The definition of terms and the methodology are discussed in the next sections, following which the backgrounds and roles of the participants are summarized.

2.1 Definition of Terms

For the purpose of this study, the management of personal electronic files includes the naming of directories and files, and their relative organization within the hierarchical disk (file) management structure. A hierarchical disk management structure is implemented across most common operating systems, including Windows, Unix and Macintosh. These structures are implemented so that files can be referenced and accessed in a consistent and logical manner that is also independent of the physical location and formatting of the disk. A typical structure is illustrated in Figure 1 and highlights five elements: the file space, directories and directory names, and files and filenames.

1. *File space.* A file space is an area where directories and files are stored. A file space can be a stand-alone device or an allocated area of a physical disk defined by either partitions or access rights. For the purpose of this study the term personal file space is used to represent all ‘allocated’ or ‘personal’ computer disks (storage) used by an individual.
2. *Directory.* A directory is an organizational unit or container that is used to reference files and or subdirectories. Directories can be arranged in a hierarchical structure. The uppermost directory is called the root directory and directories that are below another directory are called subdirectories. The term directory is often used interchangeably with ‘folder’ which reflects the graphical representation used in Windows File Explorer. For the purpose of this study, the term ‘root directory’ denotes the top level directory of a user’s file system. It is not necessarily the root directory of the drive, such as c:\ or d:\.

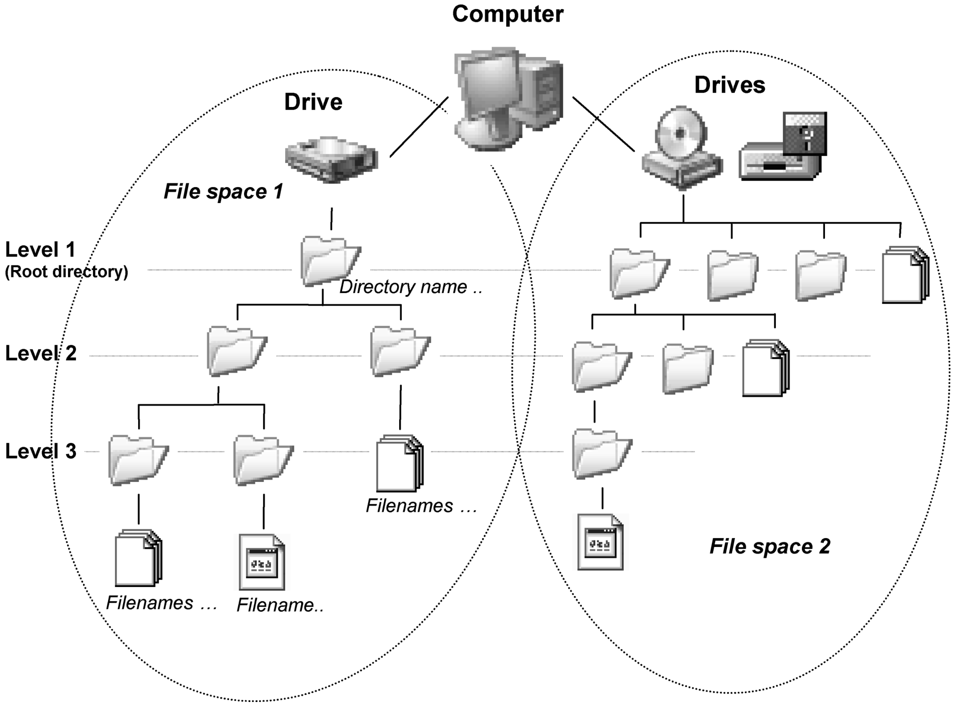


Fig. 1. The hierarchical structure of file management on personal computers.

3. *Directory name.* All directories possess an identifier, or name. Different operating systems impose different restrictions on the length and allowable characters. Furthermore, no two subdirectories within a directory can share the same name.
4. *File.* The terms “file” and “electronic file” are used interchangeably with “computer based file” and encompasses the continuum of file types that contain data necessary for, and generated by, software applications. Almost all information on a computer is stored in a file in order that its content resides over consecutive memory addresses and can hence be accessed and referenced by the operating system and software applications. There are many different file types, including standard formats and numerous proprietary formats. Examples include text files (*.txt), executable files (*.exe) and image files such as Graphical Interchange Format (*.gif) and Joint Photographic Experts Group (*.jpg) files.
5. *Filename.* All files have a uniform resource identifier (URI) commonly known as the filename. Different operating systems impose different restrictions on filenames—for example, some prohibit the use of certain characters and others limit the length of the filename. Within a single directory filenames must be unique, but two files in different directories may have the same filename. Some operating systems, such as Unix and Macintosh, allow a file to have more than one name through a symbolic link or an alias, respectively.

Table I. Qualitative and Quantitative Data Requirements

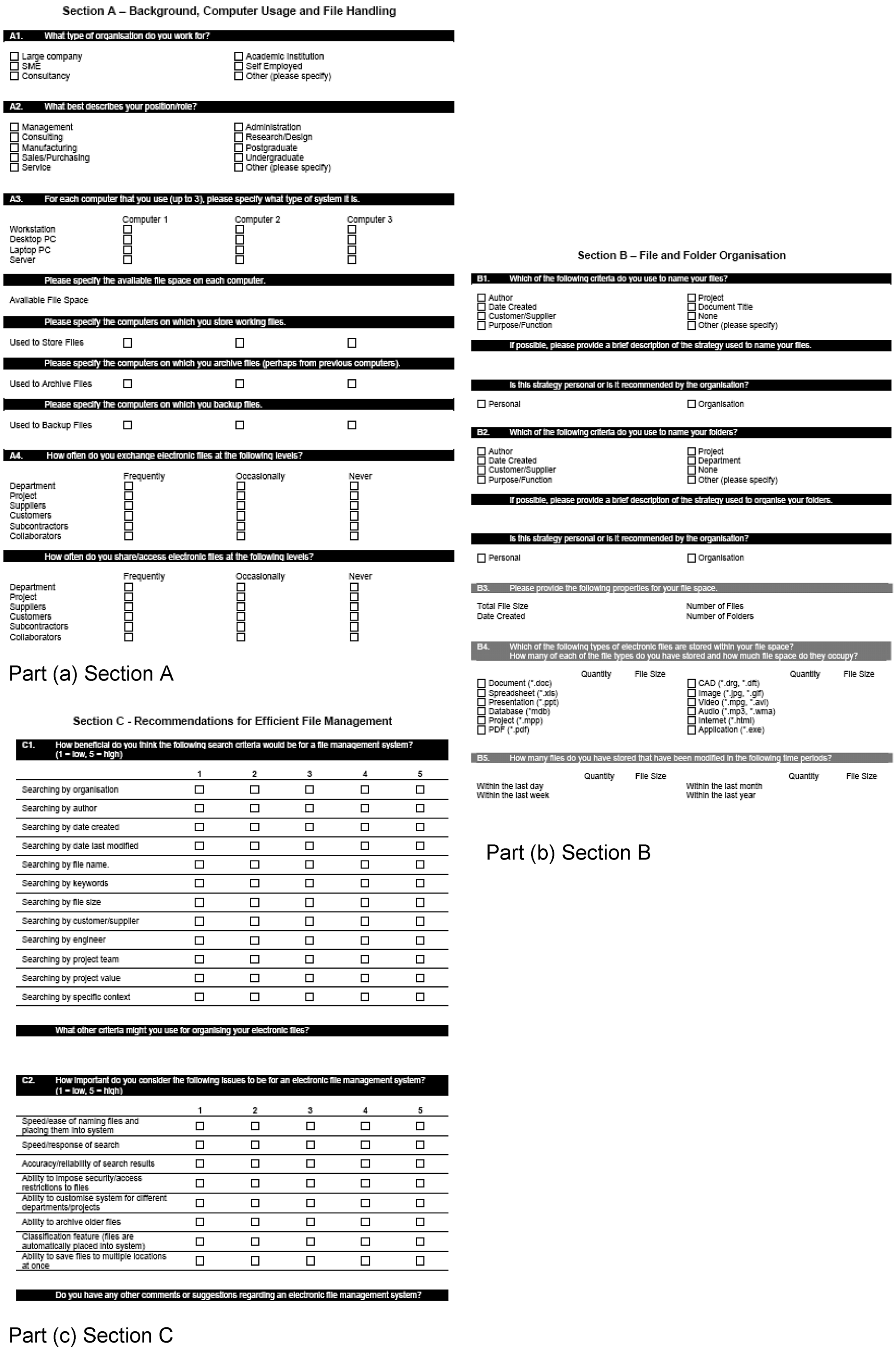
	Area	Data Required	Type
Background	Participants role	Type of organization and role of engineer	Qualitative
	Computer setup	Number of computers, their type and use	Qualitative
	File sharing and exchange	Level of file exchange, sharing and access – internally and externally	Qualitative
Personal strategies / conventions	File naming	Criteria or conventions used to construct filenames	Qualitative
	Directory naming	Criteria or conventions used to construct directory names	Qualitative
Improving access and management	Improving access	Important features / functions for improved access and retrieval	Qualitative
	Improving management	Important features / functions for improved organization and management	Qualitative
File space properties / characteristics	File space properties	Size and structure	Quantitative
	File space organization	Number of directories and files, relative organization and file properties	Quantitative
	File types	Details of the occurrence and distribution of types of files	Quantitative
	File access / use	Details of the files accessed / modified over recent periods of time	Quantitative
	Duplication	List of duplicate directory names and filenames	Quantitative

2.2 Data Collection

For the purpose of understanding the current practices of engineers for organizing electronic files, both qualitative and quantitative data are required. The qualitative data concerns the background of the participants, their personal strategies or conventions for organizing files and their views on mechanisms for improving file access and management. The quantitative data concerns the properties and characteristics of their personal file space(s). The scope of this primary research data is depicted in Table I, which also summarizes the data required in more detail.

The qualitative and quantitative data was collected by virtue of two mechanisms: a questionnaire and a computer-based audit tool respectively.

2.2.1 *Qualitative Data.* Qualitative data was collected by a questionnaire which comprised of three sections: background, personal strategies/conventions, and improving access and management. Where possible, the questions present a predefined list of responses to allow the data to be aggregated. The sections of the questionnaire are shown in parts (a) to (c) of Figure 2



Section A – Background, Computer Usage and File Handling

A1. What type of organisation do you work for?

□ Large company
□ SME
□ Consultancy
□ Academic Institution
□ Self Employed
□ Other (please specify)

A2. What best describes your position/role?

□ Management
□ Consulting
□ Manufacturing
□ Sales/Purchasing
□ Service
□ Administration
□ Research/Design
□ Postgraduate
□ Undergraduate
□ Other (please specify)

A3. For each computer that you use (up to 3), please specify what type of system it is.

	Computer 1	Computer 2	Computer 3
Workstation	□	□	□
Desktop PC	□	□	□
Laptop PC	□	□	□
Server	□	□	□

Please specify the available file space on each computer.

Available File Space

Please specify the computers on which you store working files.

Used to Store Files	□	□	□

Please specify the computers on which you archive files (perhaps from previous computers).

Used to Archive Files	□	□	□

Please specify the computers on which you backup files.

Used to Backup Files	□	□	□

A4. How often do you exchange electronic files at the following levels?

	Frequently	Occasionally	Never
Department	□	□	□
Project	□	□	□
Suppliers	□	□	□
Customers	□	□	□
Subcontractors	□	□	□
Collaborators	□	□	□

How often do you share/access electronic files at the following levels?

	Frequently	Occasionally	Never
Department	□	□	□
Project	□	□	□
Suppliers	□	□	□
Customers	□	□	□
Subcontractors	□	□	□
Collaborators	□	□	□

Part (a) Section A

Section B – File and Folder Organisation

B1. Which of the following criteria do you use to name your files?

□ Author
□ Date Created
□ Customer/Supplier
□ Purpose/Function
□ Project
□ Document Title
□ None
□ Other (please specify)

If possible, please provide a brief description of the strategy used to name your files.

Is this strategy personal or is it recommended by the organisation?

□ Personal
□ Organisation

B2. Which of the following criteria do you use to name your folders?

□ Author
□ Date Created
□ Customer/Supplier
□ Purpose/Function
□ Project
□ Department
□ None
□ Other (please specify)

If possible, please provide a brief description of the strategy used to organise your folders.

Is this strategy personal or is it recommended by the organisation?

□ Personal
□ Organisation

B3. Please provide the following properties for your file space.

Total File Size
Date Created
Number of Files
Number of Folders

**B4. Which of the following types of electronic files are stored within your file space?
How many of each of the file types do you have stored and how much file space do they occupy?**

	Quantity	File Size		Quantity	File Size
□ Document (*.doc)			□ CAD (*.drg, *.dft)		
□ Spreadsheet (*.xls)			□ Image (*.jpg, *.gif)		
□ Presentation (*.ppt)			□ Video (*.mpg, *.avi)		
□ Database (*mdb)			□ Audio (*.mp3, *.wma)		
□ Project (*.mpp)			□ Internet (*.html)		
□ PDF (*.pdf)			□ Application (*.exe)		

B5. How many files do you have stored that have been modified in the following time periods?

	Quantity	File Size		Quantity	File Size
Within the last day			Within the last month		
Within the last week			Within the last year		

Part (b) Section B

Section C - Recommendations for Efficient File Management

**C1. How beneficial do you think the following search criteria would be for a file management system?
(1 – low, 5 – high)**

	1	2	3	4	5
Searching by organisation	□	□	□	□	□
Searching by author	□	□	□	□	□
Searching by date created	□	□	□	□	□
Searching by date last modified	□	□	□	□	□
Searching by file name.	□	□	□	□	□
Searching by keywords	□	□	□	□	□
Searching by file size	□	□	□	□	□
Searching by customer/supplier	□	□	□	□	□
Searching by engineer	□	□	□	□	□
Searching by project team	□	□	□	□	□
Searching by project value	□	□	□	□	□
Searching by specific context	□	□	□	□	□

What other criteria might you use for organising your electronic files?

**C2. How important do you consider the following issues to be for an electronic file management system?
(1 – low, 5 – high)**

	1	2	3	4	5
Speed/ease of naming files and placing them into system	□	□	□	□	□
Speed/response of search	□	□	□	□	□
Accuracy/reliability of search results	□	□	□	□	□
Ability to impose security/access restrictions to files	□	□	□	□	□
Ability to customise system for different departments/projects	□	□	□	□	□
Ability to archive older files	□	□	□	□	□
Classification feature (files are automatically placed into system)	□	□	□	□	□
Ability to save files to multiple locations at once	□	□	□	□	□

Do you have any other comments or suggestions regarding an electronic file management system?

Part (c) Section C

Fig. 2. The three sections of the questionnaire.

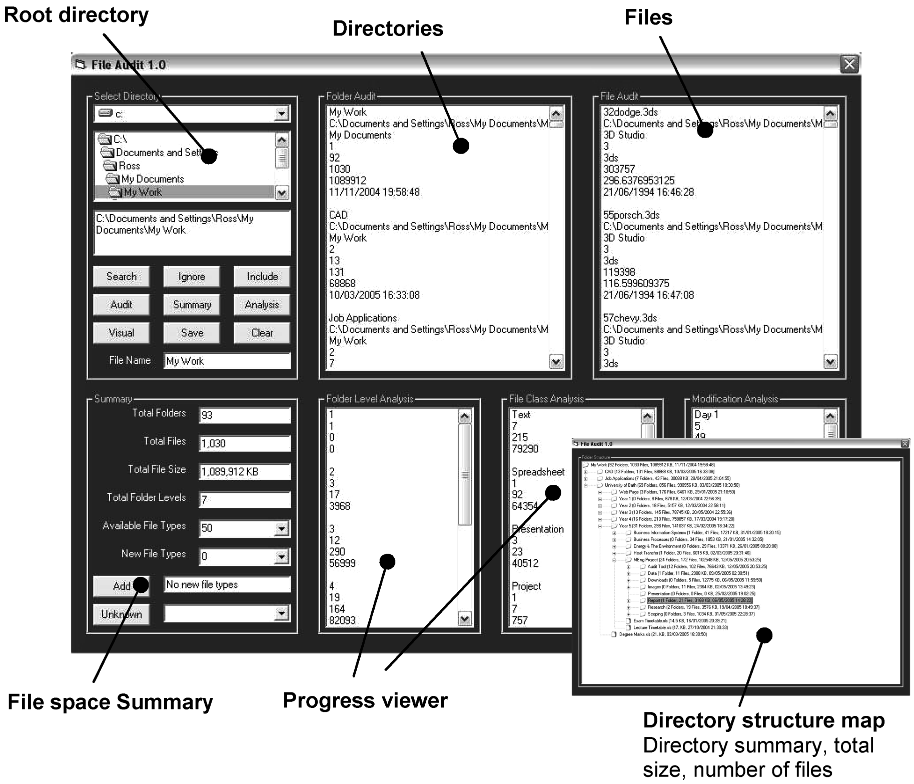
and include:

- —Section A. Background information related to the size of organization, role, computer setup, total file space available, backup and archiving policies, and frequency of file exchange/sharing with internal and external groups.
- —Section B. Personal strategies/conventions including the criteria used to name directories and files (author, date, purpose, project, document, whether naming conventions were imposed by their organization, etc.) Part B also includes the quantitative data concerning the file space(s). This is obtained through an audit tool discussed in more detail in Section 2.2.2.
- —Section C. Improving access and management. Participants were asked to rate the potential benefit of searching or indexing files according to a range of criteria. These criteria include a variety of elements which are generic to electronic files (organization, author, date created, modification, keywords, size) and also elements that are specific to engineering organizations such as customer/suppliers, projects, engineers and specific context. In addition, participants were asked to rate the importance of the performance and functionality of a file management system. These included speed and accuracy of file manipulation, search and access. The aim of these questions is to provide insight into the relative importance of the functionality and performance of approaches or tools for file management.

2.2.2 *Quantitative Data.* In order to collect the quantitative data and measure the properties and characteristics of the file space, a computer-based audit tool was created. This tool was a standalone software application that indexed the entire contents of a user's personal file space. An annotated image of the interface is shown in Figure 3. In particular, the tool was constructed to audit the following:

- —Search the root directory and list all subdirectories and files. Functionality was also included to allow multiple root directories to be analyzed and particular subdirectories to be ignored. The latter was important where participants were bound by confidentiality agreements or similar.
- —Audit all directories and record their relative level within the hierarchical file space, the number of subdirectories and files, total number of files and their sizes, the date the directory was created and last modified. These dimensions are summarized in the left-hand side of Table II.
- —Audit all files and measure the relative level within the file space, the file type (discussed in Section 4.0), the file size and the date last modified. Summarized in the right-hand side of Table II.
- —Generate a summary of the file space, including total files and directories, total file size, and the number of directory levels within the hierarchy.

In addition to auditing the file space, the tool also performed some analysis of the file space to establish the key characteristics. These are summarized in Table III and include:



Select Directory: c:

```
C:\
 Documents and Settings
  Ross
   My Documents
    My Work
```

C:\Documents and Settings\Ross\My Documents\My Work

File Name: My Work

Folder Audit

```
My Work
C:\Documents and Settings\Ross\My Documents\M
My Documents
1
92
1030
1089912
11/11/2004 19:58:48

CAD
C:\Documents and Settings\Ross\My Documents\M
My Work
2
13
131
68868
10/03/2005 16:33:08

Job Applications
C:\Documents and Settings\Ross\My Documents\M
My Work
2
7
```

File Audit

```
32dodge.3ds
C:\Documents and Settings\Ross\My Documents\M
3D Studio
3
3ds
303757
296.6376953125
21/06/1994 16:46:28

55porsch.3ds
C:\Documents and Settings\Ross\My Documents\M
3D Studio
3
3ds
119398
116.599609375
21/06/1994 16:47:08

57chevy.3ds
C:\Documents and Settings\Ross\My Documents\M
3D Studio
3
3ds
```

Summary

Total Folders: 93
Total Files: 1,030
Total File Size: 1,089,912 KB
Total Folder Levels: 7
Available File Types: 50
New File Types: 0
No new file types

Folder Level Analysis

```
1
1
0
0

2
3
17
3968

3
12
290
56999

4
19
164
82093
```

File Class Analysis

```
Text
7
215
79290

Spreadsheet
1
92
64354

Presentation
3
23
40512

Project
1
7
757
```

Modification Analysis

```
Day 1
5
49
```

File Audit 1.0

Folder Structure

```
My Work (92 Folders, 1030 Files, 1089912 KB, 11/11/2004 19:58:48)
  CAD (13 Folders, 131 Files, 68868 KB, 10/03/2005 16:33:08)
  Job Applications (7 Folders, 43 Files, 30088 KB, 28/04/2005 21:04:55)
  University of Bath (69 Folders, 856 Files, 990956 KB, 03/03/2005 18:30:50)
    Web Page (3 Folders, 176 Files, 6461 KB, 29/01/2005 21:18:50)
    Year 1 (0 Folders, 8 Files, 678 KB, 12/03/2004 22:56:39)
    Year 2 (0 Folders, 18 Files, 5157 KB, 12/03/2004 22:58:11)
    Year 3 (13 Folders, 145 Files, 78745 KB, 20/05/2004 22:55:36)
    Year 4 (16 Folders, 210 Files, 758857 KB, 17/03/2004 19:17:20)
    Year 5 (31 Folders, 298 Files, 141037 KB, 24/02/2005 18:34:22)
      Business Information Systems (1 Folder, 41 Files, 17217 KB, 31/01/2005 18:20:15)
      Business Processes (0 Folders, 34 Files, 1853 KB, 21/01/2005 14:32:05)
      Energy & The Environment (0 Folders, 29 Files, 13371 KB, 26/01/2005 00:20:08)
      Heat Transfer (1 Folder, 20 Files, 6015 KB, 02/03/2005 20:31:46)
      MEng Project (24 Folders, 172 Files, 102548 KB, 12/05/2005 20:53:25)
        Audit Tool (12 Folders, 102 Files, 76643 KB, 12/05/2005 20:53:25)
        Data (1 Folder, 11 Files, 2988 KB, 09/05/2005 02:38:51)
        Downloads (0 Folders, 5 Files, 12775 KB, 06/05/2005 11:59:50)
        Images (0 Folders, 11 Files, 2364 KB, 02/05/2005 13:49:23)
        Presentation (0 Folders, 0 Files, 0 KB, 25/02/2005 19:02:25)
        Report (1 Folder, 21 Files, 3168 KB, 06/05/2005 14:28:22)
        Research (2 Folders, 19 Files, 3576 KB, 19/04/2005 18:49:37)
        Scoping (0 Folders, 3 Files, 1034 KB, 01/05/2005 22:28:37)
      Exam Timetable.xls (14.5 KB, 16/01/2005 20:39:21)
      Lecture Timetable.xls (17. KB, 27/10/2004 21:30:33)
    Degree Marks.xls (21. KB, 03/03/2005 18:30:50)
```

Fig. 3. The graphical user interface of the computer based file audit tool.

- The proportion of directories and files stored in each level of the hierarchy.
- The occurrence and distribution of particular file types (classes).
- The proportion of files and directories that were modified within various time periods. These periods included each day for the last 7 days, each week (consecutive 7-day periods) for the last 5 weeks, monthly for the last 12 months and those modified or created more than 1 year ago.
- The level of duplication of filenames and directory names.

2.3 Participants

The study involved 40 engineers from a variety of industrial and academic backgrounds. The participants included: two departmental managers, three project managers, twelve engineers, fifteen research engineers, and eight trainee engineers. Of the fifteen research engineers, twelve belonged to a university research center involved in variety of collaborative projects with industry (The Innovative Design and Manufacturing Research Centre at the University of Bath). The study was conducted over a period of four days (Tuesday to Friday) during June 2005. Importantly, no auditing took place on a Monday because of the need to analyze files modified on the previous working day.

Table II. Quantitative Data for Directories and Files

Directories	Files
Directory name	Filename
Directory path	File path
Parent directory	Parent directory
Relative level in the hierarchy	Relative level in the hierarchy
Total subdirectories contained	File extension
Total files contained	File type
Total file size (in kilobytes)	File size (kilobytes)
Date last created / modified	Date last modified

Table III. Data Analysis

Directory Analysis	File Analysis	Modification Analysis
For each directory level 1 to n (where n is user dependent)	*For each file type (File types are characterized in Section 4.0)*	*For each time period (Last 7 days, 5 weeks, 12 months, 5 years)*
Number of directories	Number of files	Number of directories modified
Number of files	Number of files	Number of files modified
Total size of all files	Total file size	Total file size modified

3. A CLASSIFICATION OF FILE TYPES USED BY ENGINEERS

In order to understand how engineers manage personal electronic files it is necessary to be able to determine the fundamental classes of electronic file generated and used by engineers and their associated software applications. In today's computer-driven business environments there are a wide variety and diversity of software applications available, many of which require and generate files based on proprietary standards denoted by the file extension. For some common desktop applications and environments, such as the Internet, there are standards governing data representations and formats. However, there are still many hundreds of different formats for what can be considered to be similar documents or file types. For example, image files can be represented by GIF, BMP, JPEG, or TIFF formats, to name but a few. This variety and diversity of formats and files makes characterizing the practices of engineers very complicated and arguably divergent. To overcome this, and provide more insight into file types and their management, it is necessary to implement a content or context-based classification of similar files types. In order to generate this classification, a preliminary study of ten engineers was undertaken to establish and characterize the file types. This preliminary study revealed sixteen classes of electronic file: Text, Spreadsheet, Presentation, Database, Project, Graphics, CAD, Code, Simulation, Application, Image, Audio, Video, Internet, Data, and Compressed files. These file classes are described with examples in Table IV.

For the purpose of auditing the file spaces, the sixteen fundamental classes were incorporated into the audit tool (2.2.2). More specifically, all of the files and file extensions associated with each category were incorporated into the tool to allow the automated classification of file types. The typical file types and extensions classified are given in Table V. During the study, additional file

Table IV. The Sixteen Fundamental Classes of File Used by Engineers

	File Class	Description	Typical File Types
1	Text	Many files may contain text in some form, however, the text file classification refers to files (documents) that have been created with the intention of communicating predominantly textual information to a reader. Text files can contain non-text objects such as images, provided the overall focus of the document remains primarily textual.	Microsoft Word, Adobe Acrobat (PDF), Rich Text Format, Postscript
2	Spreadsheet	Spreadsheet files include tables or matrices of values arranged in rows and columns, where each value can have a predefined relationship to the another value(s).	Microsoft Excel, Lotus Notes
3	Presentation	Presentation files include documents or content that is generally visual and intended to be communicated with a commentary. Any file that runs in a slide show format can be considered a presentation file.	Microsoft PowerPoint
4	Database	A database is a data file that uses a proprietary format and environment to structure the data in order to access, search and efficiently organize the data.	Microsoft Access, Hypertext
5	Project	Project files are used to schedule the events required to complete a given project, often in the form of a Gantt chart. Hence, project files are used for planning purposes and are effectively timetables of events and tasks.	Microsoft Project
6	Graphics	The purpose of a graphics file is to construct and display graphical information, using vector-based images or diagrams mixed with bit-mapped graphics. In the field of engineering, graphics files are likely to contain schematic diagrams of apparatus and photographs showing experimental processes.	Microsoft Publisher, Adobe Photoshop, Corel Draw
7	CAD	In the field of engineering, Computer Aided Design (CAD) has developed into a very powerful tool for design and manufacture. CAD files will include 3-dimensional models of parts and assemblies, as well as 2-dimensional engineering drawings.	Solid Edge, AutoCAD
8	Code	This encompasses numerical and visual analysis bespoke software. The code will generally be compiled into a standalone application (*.exe) or run within an integrated design environment (IDE).	Microsoft Visual Basic, Visual FoxPro C, C++
9	Simulation	It is common practice for engineers to use software applications that can perform structural or finite element analysis using 2-dimensional or 3-dimensional models. Any files that contain data created by the user for use with such applications, as well as output files from the applications, are classified as simulation files.	ANSYS, Matlab
10	Applications	Application files are defined as executable files or standalone applications that perform a particular function. This may include standalone code (8) and other proprietary applications. The applications include executables and system files that have been installed within the user's file space.	Executable files, system files

Table IV. Continued

11	Image	A file that contains a single photo, scanned image, drawing or diagram, etc. in any format is defined as an image file.	JPEG, GIF, TIF, Bitmap
12	Audio	A file that contains a single stream of sound of any length in any format is defined as an audio file.	MP3, WAV, Windows Media Audio
13	Video	A file that contains video footage, with or without an audio stream, is defined as a video file.	MPEG, AVI, Adobe Premiere
14	Internet	A file that conforms to the W3C and associated standards for the presentation and representation of data on the Internet. These may also include files that define and support the functionality of Internet websites.	HTML, CSS, JavaScript
15	Data	A file that contains supporting information or output data from another application is defined as a data file. Data files are likely to be text based, but differ from text files in that they are intended to be communicated between software applications rather than users.	Data files
16	Compressed	A file that has been compressed or ‘zipped’ in order to reduce the number of bytes it occupies is defined as a compressed file. Compressed files may contain files of any other format.	Backup WinZip, WinRar

types were incorporated as they were identified during the audit. On average the audit tool was able to automatically classify more than 75% of files within a users file space and in many cases almost 95% (see table 7 in the results section, below).

4. RESULTS

The results of the study are presented in Figures 4 to 9 respectively and discussed in the following sections under the five areas of background, personal strategies, improving access and management, file space audit and file space analysis.

4.1 Background

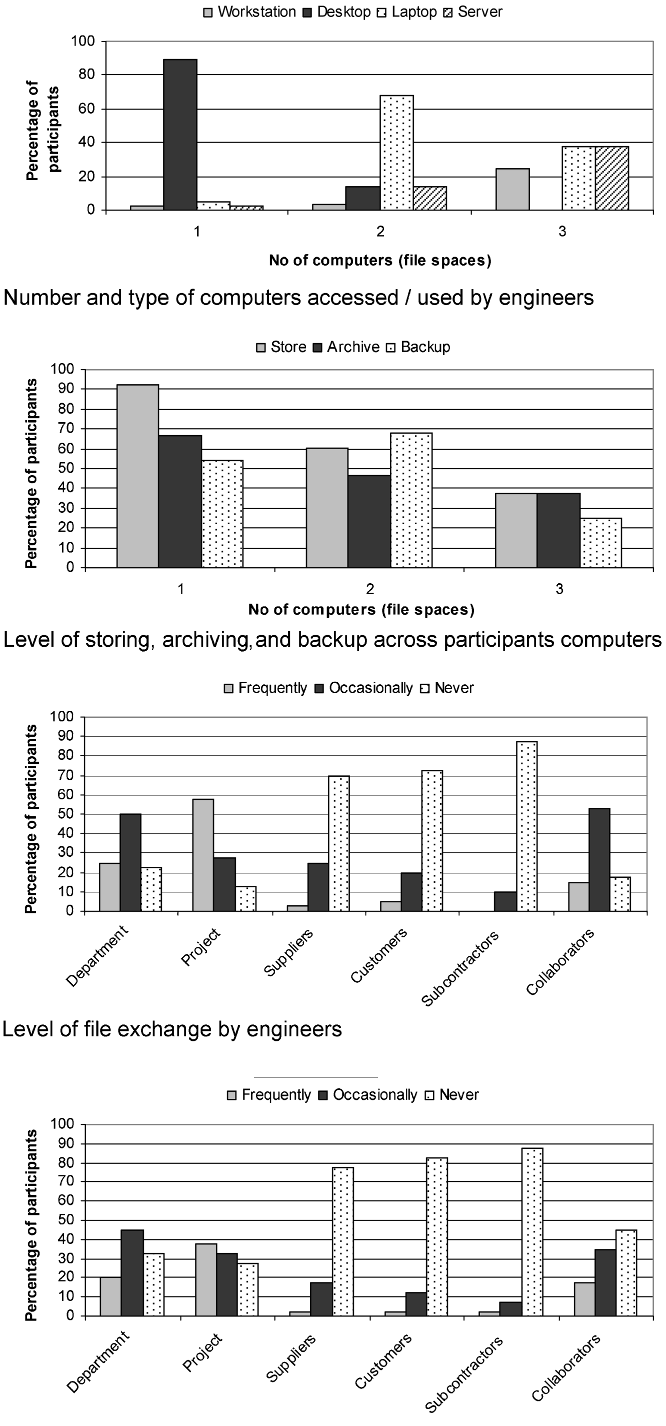
Of the 40 engineers studied, only twelve use a single computer, twenty use two computers, and eight use three computers (part (a) of Figure 4). Unsurprisingly, almost 90% of engineers use a desktop PC as their primary computer, and of the 28 who use two computers, almost 70% use a laptop as their secondary computer. Of the eight engineers who use a third computer, no real preference emerged for its type, although two participants specifically describe accessing UNIX Workstations. Those who accessed the UNIX Workstations did so because they needed to run software which ran only on those computers.

4.1.1 *Archiving and Backup.* For the purpose of this study a distinction is made between the tasks of archiving and backup. Archiving denotes the organization and storage of files for their long-term preservation and typically

Table V. File Classes and Associated File Types (Extensions)

File Class	Software Applications	File Types (extensions)
Text	MS Word, Adobe, Text, Rich Text, Lotus notes	doc, rtf, pdf, txt, ps, dot, one
Spreadsheet	MS Excel, Lotus	xls, imp, wk
Presentation	MS PowerPoint	pps, ppt, pot
Database	MS Access, dBase, SQL	mdb, db*, odb, sql
Project	Microsoft Project, MPMM, Visio	mpp, mpm, vsd
Graphics	Adobe PhotoShop, Corel Draw, MS Publisher,	pub, cdr, psp
CAD	Solid Edge, UniGraphics, CATIA, PRO/Engineer, AutoCAD	sld, prt, dxf, dwg, dwf, prt, asm, drw, igs, mod, dlv, exp, stl, step
Code	Visual Basic, C, C++, PHP, Perl, Lisp, Fortran, Pascal	vbp, vba, Visual FoxPro, c, c++, php, pl, aps, f, f77, h, h++, p, l, pas
Simulation	ABAQUS, ANSYS, LS-DYNA, Maple, MathCAD, Simulink, MAtrixX, LabView, SPSS, Moldflow, CFD, 3 D Studio	fil, mdl, inp, mat, ms, mcd, sav, ctt, vit, mfr, mll, std
Applications	Executables, Installation, Initialization	exe, msi, ini
Image	Variety of standard formats	jpeg, jpg, gif, tif, tiff, bmp, png, wmf, eps, pic
Audio	Variety of standard formats	mp3, mpeg, wav, wma, cda
Video	Variety of standard formats	mpeg, avi, qt
Internet	Variety of standard formats	html, htm, js, jss, php, swf, xsl, asp, cfm
Data	Data files and Comma Separated Variables	dat, csv, xml
Compressed	WinZip, WinRar, Backup files, Archives	zip, gzip, tar, zip, bak, tar, arc, tgz

includes a 'master' version of a file that will be used for future reference including, for example, commercial or legal disputes. In engineering such archives are commonly produced for each project and for a particular customer. In contrast to archiving, the task of backup is a shorter-term activity involving the generation of a “copy” of a file system or drive for the purpose of restoring working files should the need arise. These backups are typically generated at regular intervals and hence organized in a chronological manner. In terms of electronic file storage, primary computers are used by over 90% of respondents to store files, by 70% to archive older files and by 55% to backup files (part (b) of Figure 4). The general trend is a decreasing percentage of participants storing, archiving and backing-up files on subsequent computers, despite the fact that the values are calculated using only the number of participants using a second or third computer.



Part (a) Number and type of computers accessed / used by engineers

Part (b) Level of storing, archiving, and backup across participants computers

Part (c) Level of file exchange by engineers

Part (d) Level of file sharing by engineers

Fig. 4. Background data and levels of file sharing of participants.

4.1.2 *File Exchange.* In order to investigate the level of file exchange a distinction is made between the modes of file sharing and file exchange. File sharing is considered to involve the exchange of personal files from one computer to another, whilst file sharing occurs when multiple users (computers) have access to the same instance of a file residing in a particular location. If a user transfers a file from one personal computer to another via methods including email (as an attachment), file copying, or USB memory sticks, this is considered file exchange. If a user shares a file using shared disk drives including Windows file sharing or network file servers, this is considered file sharing. Of particular, interest in this study was the level of exchange and sharing that occurs across departments, within project teams and with suppliers, customers, subcontractors, and partners. The results are shown in parts (c) and (d) of Figure 4 and reveal that the majority of participants rarely exchange or share files with suppliers, customers or subcontractors. However, a significant proportion of the sample (45%) were based in a university research centre, and whilst these projects involve a large number of industrial collaborators, it is arguable that, due to the nature of the work, fewer suppliers and/or customers are likely to be involved. The results suggest that files are most commonly exchanged and shared at a project level, with 60% of participants claiming to “frequently” exchange files with fellow project members. Files are also regularly exchanged and shared across different departments and with collaborators. One overall trend that emerges from the study is that files are more commonly exchanged than they are shared, regardless of whether the recipient is internal or external.

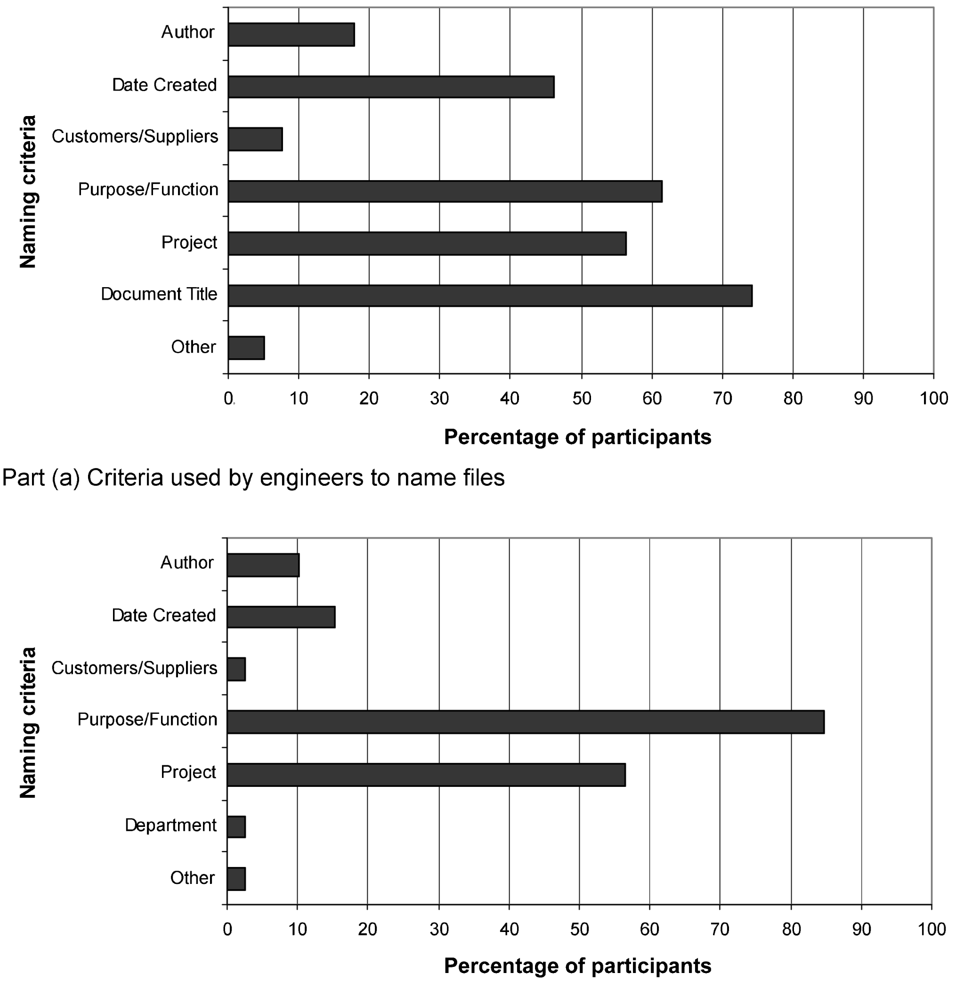
4.2 Personal Strategies

When considering file naming strategies, over 75% of engineers use the title of the document, 60% also use the purpose or function (e.g., maximum stress in wing spar or revised costs for project meeting), 50% the project title (e.g., A380M or LG carrier) and 45% the date. The latter is particularly common for files containing meeting minutes and report drafts. The relative utilization of naming criteria is shown in part (a) of Figure 5. A number of additional criteria were also listed by the participants and include:

—The name of the person to whom the file is being sent, for example, project costs for tony.xls
—A memorable nickname to make the content easily recognizable
—Document revision information, for example, annual report version 2.doc
—A brief description of the file content, for example, output from model of cantilever.ans
—The file type, for example, word document.doc.

Some of the participants also described using different naming criteria for files depending on their level in the directory structure.

Where directories are considered, two criteria emerge as by far the most common (see part (b) of Figure 5). These are the purpose or function of the files contained in the directory for example, budgets (85% of participants) and the project title (55% of participants). This suggests that directories are commonly



Part (a) Criteria used by engineers to name files

Part (b) Criteria used by engineers to name directories

Fig. 5. The naming criteria used by engineers for files and directories.

organized using a function and project hierarchy. Following these two criteria, the date was the next most prevalent, although only 20% of participants used this. Other criteria listed by the participants for naming and organizing directories are:

—Subject for example, aerodynamics
—Work theme for example, maintenance
—The year (as opposed to the full date)
—File type for example, presentations

As part of the study participants were asked to indicate if their strategy was personal or recommended by the organization. For all of the participants their strategies were largely personal. There was only one exception where the engineer—who had been saving the majority of files on the desktop—had been instructed by Computer Services to clear all files from their desktop and place them in theme-based directories. The prevalence of cognitive reference points relating to the ‘purpose’ and ‘use’ of a file is congruent with results from studies by Kwasnik [1989], where it was shown that these non-concrete dimensions have a greater influence in determining the classification of a document within a personal office.

4.3 Improving Access and Management

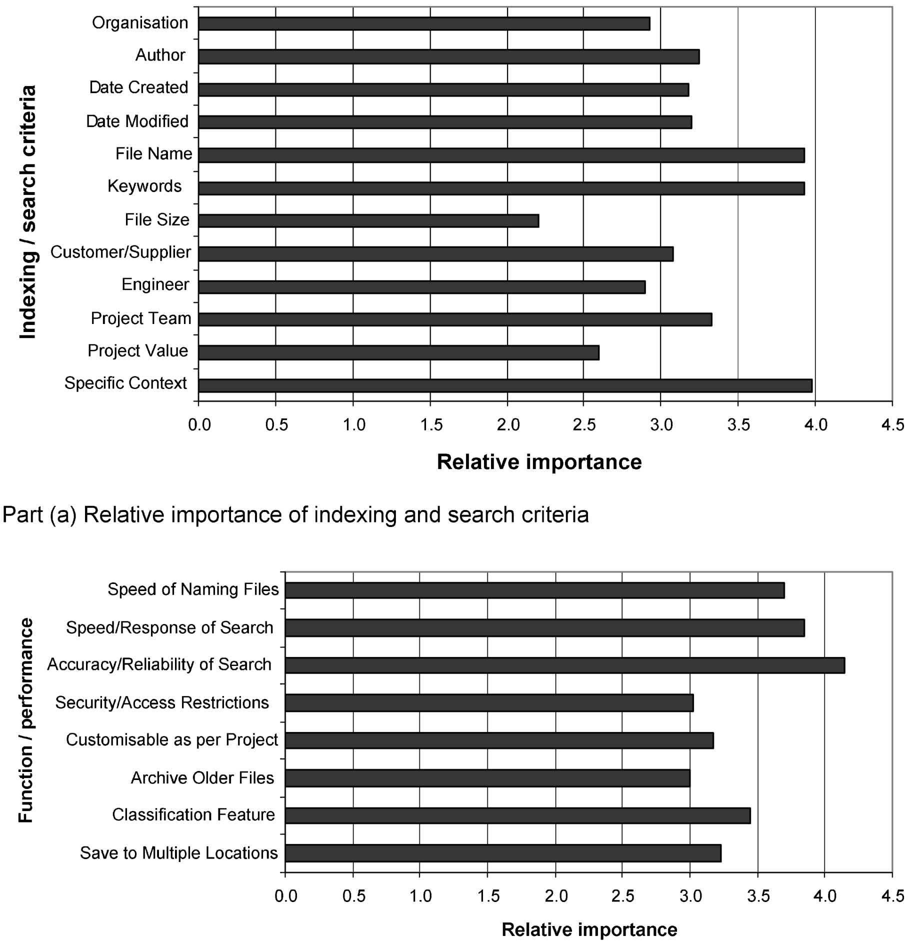
The third section of the questionnaire required the participants to rate the relative importance of file indexing and searching criteria, and the functionality and performance of file management systems on a scale of 1 to 5. The results are shown in parts (a) and (b) of Figure 6. In terms of indexing or search criteria, twelve alternatives were presented, including specific context, project value, project team, engineer, customer/supplier, file size, keywords, filename, date modified, date created, author, and organization.

What is immediately obvious from part (a) is that nearly all criteria are rated above three with the exception of file size and project value which are rated at 2 and 2.5 respectively. The most highly ranked criterion was specific context with an average score of 4.1. This correlates well with the findings in 5.2, where specific purpose or function are commonly used to name and organize files and directories. Of the remaining criteria, filename and keywords, both score an average of four, with ‘project team’ and ‘author’ also scoring above average.

In part (b) of Figure 6, the relative importance of functionality and performance of the file management system are presented. In total seven aspects were considered: the ability to save to multiple locations, a classification feature, archiving older files, project specific structure, security/access restrictions, accuracy/reliability of search, speed/response of search, speed of naming/renaming file. The results reveal that accuracy and reliability of searching (average score 4.2), as well as the response time of a search are the most highly ranked functions for an electronic file management system. With an average score of 3.8, the speed of naming and manipulating files is also important. In a similar manner to the indexing criteria, all of the functions presented in the study returned an above average score.

4.4 File Space Audit

The results of the audits of the 40 file spaces are presented in Table VI. For each dimension, the average value, highest value, lowest value and standard deviation are presented. As previously described in Section 3.2.2 the audit tool provided the option to include directories and files across multiple root directories. However, of the 40 engineers that participated in the study, only 4 had more than one root directory. In one case, five root directories were used to store files. In general, the directories were all hierarchical in nature, structured over an average of 8.7 levels (with a standard deviation of 2.67). The maximum number of directory levels was 14 and the lowest was 3.



Part (a) Relative importance of indexing and search criteria

Part (b) Relative importance of the functionality and performance of file management systems

Fig. 6. The relative importance of functionality and performance of file management systems by engineers.

In addition to this, the average size of all files stored in each directory is also determined. This shows that the average size of files stored within a single level of the hierarchy is approximately 285Mb. The average size of all files in an individual directory is 7.8Mb and 613Kb for an individual file.

In addition to auditing the properties of all directories and files, it is also interesting to evaluate the percentage of total directories and files residing within each level of the hierarchy (14 levels in the highest case). The overall percentage of total directories and files within each directory level is shown in Figure 7. This reveals a bell shaped distribution curve with a slight skew to the left. It has been previously observed that the mean number of directory levels contained within an engineer's file space is 8.6. Figure 7 also reveals that the majority of files are stored in the third, fourth and fifth levels, with the fourth directory level being the most heavily used. In fact, 54.6% of all directories and

Table VI. Personal File Space Properties

File Space Properties				
Total	Average	High	Low	SD
Root Directories	1.19	5	1	0.71
Directory Levels	8.67	14	3	2.67
File Types	155	774	5	148
Directories	406	2,314	13	431
Files	4,660	30,849	200	5,674
File Size (KB)	2,578,076	12,675,176	5,856	3,432,667
Properties per Directory Level, Directory and File				
Number of Directories per Directory Level	41.8	192.8	4.3	36.8
Number of Files in Directory Level	494.8	2,570.8	40.0	502
Number of Files in each Directory	11.6	33.6	3.07	6.91
Size of Directory Level (KB)	287,885	1,260,063	1,171	372,463
Size of each Directory (KB)	7,836	51,328	145	11,706
Average File Size (KB)	613	4,865	29	871
Duplicate Directory and Filenames				
Directories with a shared name	161	1,035	0	237
(% of Total Directories)	(31.3)	(75.4)	(0.0)	(18.4)
Files with a shared name	1,925	11,614	4	2,922
(% of Total Files)	(32.4)	(83.1)	(1.1)	(19.9)

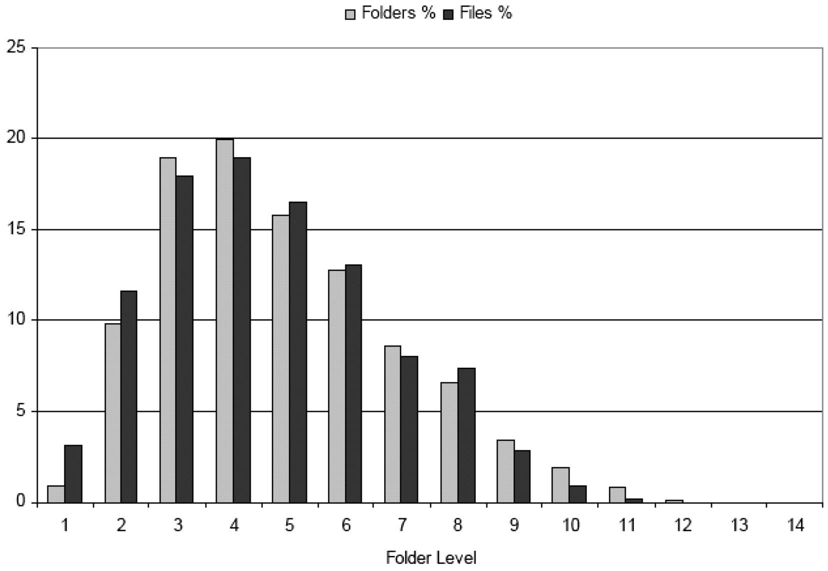


Fig. 7. Cumulative distribution of directories and files over participants file spaces.

53.5% of all the files are contained in these levels. 25.7% of directories and 24.6% of all files are contained in the second and sixth levels, with 15.2% of directories and 15.4% of files maintained in the seventh and eighth levels. The first level of the hierarchy tends to be used primarily for subdirectories hence the proportion of total directories and files contained is very low. The lower levels beyond the eighth level will also tend to contain a very low proportion of the total directories and files, if indeed they exist at all. These observations are

Table VII. Personal File Space Properties

File Class Analysis								
File Class	Files (% Total)				File Size (% Total)			
	Av.	High	Low	SD	Av.	High	Low	SD
Text	20.5	77.0	0.00	18.9	24.3	83.7	0.00	20.5
Spreadsheet	4.27	40.4	0.00	7.38	4.24	47.6	0.00	9.03
Presentation	1.73	7.35	0.00	1.96	7.50	41.6	0.00	9.80
Database	0.57	11.0	0.00	1.70	2.21	47.0	0.00	8.48
Project	0.21	6.28	0.00	0.97	0.10	1.27	0.00	0.28
Graphics	0.38	3.59	0.00	0.74	0.20	1.52	0.00	0.33
CAD	3.99	57.6	0.00	10.5	2.02	31.8	0.00	6.15
Code	7.48	31.3	0.00	8.34	0.76	7.83	0.00	1.45
Simulation	1.22	25.0	0.00	4.22	2.05	59.9	0.00	9.76
Application	1.45	6.09	0.00	1.77	3.09	29.7	0.00	5.53
Image	20.0	47.2	0.00	12.1	10.5	47.9	0.00	10.5
Audio	1.25	35.8	0.00	5.56	3.84	76.4	0.00	13.0
Video	0.19	1.53	0.00	0.30	6.98	78.8	0.00	16.6
Internet	5.91	26.2	0.00	6.89	0.51	4.49	0.00	0.91
Data	4.52	17.5	0.00	4.93	2.69	30.9	0.00	5.47
Compressed	2.47	87.0	0.00	13.4	4.92	49.9	0.00	8.91
Unknown	23.9	100	0.18	21.2	24.2	100	0.00	24.9

further supported by the fact that the proportion of total directories within all levels of the cumulative hierarchy is approximately the same as the proportion of total files, implying that the ratio of files to directories will be roughly constant in all levels of the folder hierarchy at around 12 (Table VI).

4.5 File Space Analysis

The percentage of total files and total file size represented by each class of file is given in Table VII and shown in Figure 8. The results indicate that text files and image files are the most commonly occurring file types, each representing an average of just over 20% of all files. Following text and images, the next most prevalent file types are code, Internet, data, spreadsheet, CAD, presentation, simulation and applications which each represent 5% to 3% of total files. Of particular note is the fact that only 4% of all files were CAD files. This is surprising given the ubiquitous nature of CAD systems in engineering environments, although it could be explained by the critical value of CAD to engineering organizations and the resulting prevalence of well-developed CAD document management systems.

The analysis of file types also includes the relative file sizes of files in the different classes. Unsurprisingly, video files exhibit the highest ratio of size to occurrence (37:1) occupying 7% of the total audited file space, but only 0.2% of total files. Similarly, presentation files, most commonly produced using Microsoft PowerPoint have a ratio over 4:1. The ratio for text files, spreadsheet files and unknown file types is approximately 1:1, whereas images have a smaller ratio of 1:2. Of particular note in Table VII are the large standard deviations calculated for each of the various file classes. These values are included to highlight the wide spread of measured values across the engineers

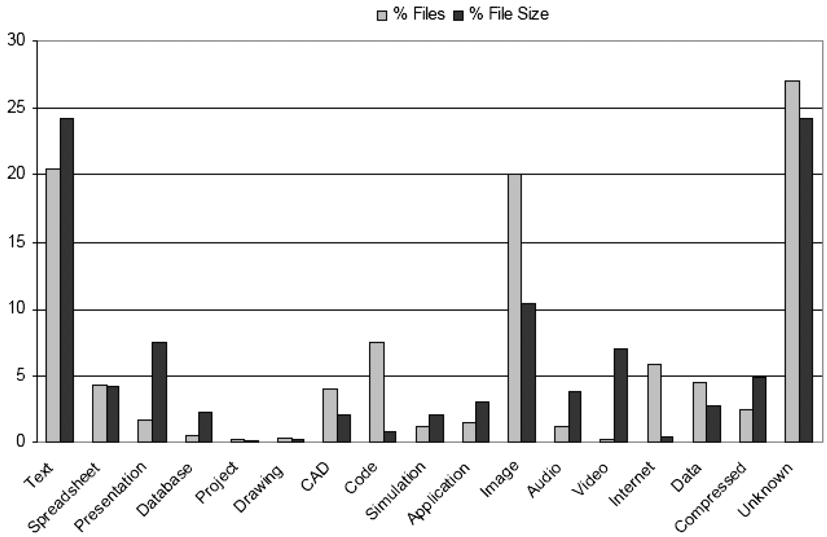


Fig. 8. Relative distribution of file types within participants file spaces.

studied. Whilst it is arguable that large standard deviations are to be expected given the diversity and variety of the engineering tasks and computational tools used by engineers, such large values were not anticipated. One possible explanation for this variation is the inherent variety in the data files produced by, for example, finite element analysis (FEA) software where output files can range from kilobytes for a configuration file, to gigabytes for the results. Furthermore, significant variation is also likely to arise from the widely differing levels of scale and complexity of the systems being engineered. For example, these can range from large special purpose machinery to small household products, and can be modeled using wireframe representations or solid models, the latter generating significantly larger data files. This explanation is supported by the fact that the some of the highest standard deviations observed with respect to the calculated averages occur for CAD files simulation files.

In addition to considering the occurrence and distribution of file types, the further analysis also considered the modification of files and directories and the duplication of filenames within a file space. The analysis of modification was intended to establish the number of individual directories and files modified within a given time period. For the purpose of this study, modification is identified from the last modified date. However, it is possible that a particular file could have been modified a number of times and such cumulative modifications are not measured in this study. Table VIII shows the mean number of files and directories modified within the previous day, week, month and year. These values are also presented graphically in Figure 9, although data for the year is not included in order to maintain a reasonable scale.

The analysis of modifications for the previous working day reveal that an average of 8 directories and 17 files are modified. In one working week, an average of 22 directories and 50 files are modified. In one month, an engineer

Table VIII. Modification of Directories and Files over Time

Modification Analysis								
Modification	Directories				Files			
Period	Av.	High	Low	SD	Av.	High	Low	SD
Last Day	8.21	55	0	14.0	16.5	253	0	41.0
Last Week	21.9	217	0	40.0	49.5	281	0	64.0
Last Month	58.6	297	1	78.2	183	1,265	5	242
Last Year	295	2,312	8	368	1,531	8,657	60	1,752

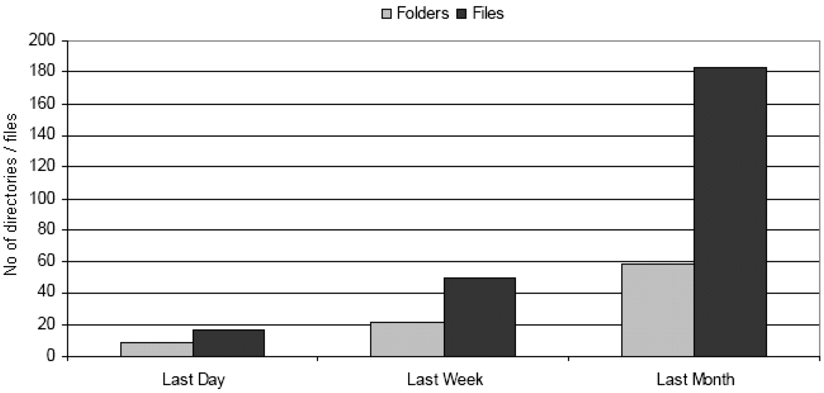


Fig. 9. Number of directories and files modified over recent time periods.

will work with an average 59 directories and 183 files. Over the course of an entire year, an engineer will work with around 300 directories and 1500 files. The relative ratios of files modified in a month to a week and a year to a month are roughly proportional to the respective ratios of working time. For example, the working time in a month could be considered roughly 4 times that in a week, (taking into account days away from the office and holidays) and a working year could be considered to be roughly 10 times a working month for similar reasons. The relative ratios of files modified in these time periods (3.66 and 8.37 respectively) is of a similar magnitude to the temporal differences. In contrast, to this, when the previous week to previous day is considered, the ratio is only 3.1. The likely cause is that over a week, many files will be modified more than once.

The final stage of the analysis involved the identification of directory and filename duplication. The number of duplicate filenames and directory names is shown in the lower portion of Table VI. In the most extreme case observed, over 75% of directories and 83% of files used shared names which equated to over 1,000 directories and 11,000 files respectively. The average results are also surprisingly with 32% of files sharing the same name and 30% of directories using shared or nonunique names. Such statistics suggest a culture of using common names to describe subdirectories and files within a given directory. For example, an individual may have subdirectories for two different projects, both of which contain a directory called design and a file called report.

5. DISCUSSION

For the purpose of this study the findings of the audit are compared and contrasted with the results of previously reported studies and also discussed with respect to what are considered to be engineering-specific requirements.

5.1 General File Management Observations

In line with standard practices for the reporting of the organization of hierarchical, electronic file systems, we summarize the findings in relation to archiving, maintenance, and the use of hierarchical structures, [Barreau and Nardi 1995, 1997] and the psychological needs of users [Lansdale 1988].

1. Archiving and backup—The majority of engineers archive electronic files (70%) and around half backup files (55%). This is generally performed between primary and secondary computers. Following the characterization proposed by Whittaker and Hirschberg [2001], engineers are filers rather than pilers and store their personal electronic files in structured archives. The average size of the directory structure is 2.5Gb. This includes 406 subdirectories and 4,660 files. This is far smaller than the mean of 8900 directories previously reported by Agrawal et al. [2007]. However, the average file size of 613Kb is nearly three times larger than the mean file size reported by Agrawal et al. This difference can be accounted for by the relatively large size of the data files produced by engineering simulation packages and CAD software, which account for about 4% of the total file space, and audio and video files, which account for about 11% of the total file space. This finding suggests that larger block sizes for file-systems would be preferable since engineers generally store fewer files of large sizes.
2. Maintenance—In an average day, an engineer will modify 17 files from 8 different directories; in a week, the engineer will access 50 files from 22 directories; in a month, an engineer will access 183 files and 59 directories, and in a year, 1500 files from 295 directories. In all, however, the engineers accessed only about 33% of the files stored on their personal computers each year and less than 4% on a monthly basis. This finding suggests that once a file has been created, used, and then archived, it is relatively unlikely to be accessed again. The engineers in the study do not report significant challenges in maintaining their file systems once the files are archived, although this might change as their hard disk drives fill up. One surprising finding, though, relates to duplicate files and directories, with the engineers in this study often maintaining two subdirectories for two different projects, both of which contain a directory called design and a file called report.

   The general trend for engineers is to have two working computers, which for the purposes of this work are described as the primary and secondary computer respectively. Primary computers have a strong tendency to be desktops, whereas secondary computers tend to be laptops. The engineers thus manually maintained up-to-date copies of files on both computers. That is, the engineers must remember to update. There is no built-in capability in the versions of the operating systems used by the participants in the study to remind the engineers to update, nor did the engineers report knowing about or using a software tool to synchronize multiple copies of the same file. Having to remember that files have been changed and thus require synchronization across all copies is not included as a cognitive load in Lansdale's [1988] scheme, although we believe that remembering to do such a task would present difficulties for some users. Surprisingly, the engineers we studied did not comment on the effort they expended in maintaining up-to-date copies of files across computers, although this was not asked of them explicitly.

3. Hierarchical structures and namespace usage—The two most common criteria used for naming directories are i) the purpose or function of the directory, that is, the specific context of the file that it contains (85% of participants) and ii) the name of the project (55% of participants). This finding is consistent with that of Boardman and Sasse [2004] who also found that the project name was a common top-level name for a directory structure. However, in this study "purpose" and "function" are the most widely used, whilst Boardman and Sasse [2004] found project name was the most common. One possible reason for this could be the underlying function-based perspective and practices of engineers (discussed in detail in the next section). This difference might also suggest that naming strategies and directory hierarchies could be highly context-dependent, with generalizations across industries and companies difficult to make. However, the findings of this study—and in particular the use of cognitive reference points relating to the purpose and use—are consistent with Kwasnik [1989], where it was shown that such nonconcrete dimensions have a greater influence in determining the classification of a document. The date a directory was created is also used in some cases (20% of participants), indicating a last-option for file naming when no other suitable criteria seemed applicable.

   The directory structure includes an average of almost nine hierarchical levels with 42 subdirectories and 495 files per level, giving an average of 11.6 files per directory. In addition, it was observed that the third, fourth, and fifth levels of the hierarchy hold the majority of all the files (53.5%), regardless of any hierarchy imposed by the software. That is, the hierarchies are user-defined rather than defined by the operating system or a particular software program. The number of files by name-space depth parallels the finding by Agrawal et al. [2007], who reported that most files were stored within a directory name-space depth of two to four.

4. Psychological issues—Given previous studies on the cognitive load of locating files for engineers [Crabtree et al. 1997] we did not study this aspect specifically. Although they were not specifically asked, the engineers studied did not report whether or not they expended considerable cognitive effort in naming files. The most useful and/or commonly used mechanisms for indexing and retrieving files are context, followed by filename and keywords, then finally project team and author. The personal strategies for naming files based primarily on the title of the document (75% of files) and the purpose or function of the files in a directory (e.g., budgets 85% of the time) suggests

that not much cognitive effort was placed into naming files. This is again consistent with Kwasnik [1989], who concludes that document names are chosen that have the most usefulness for the least cognitive effort.

5.2 Engineering-Specific File Management Observations

Aside from the file-system management aspects, the observed behaviour of the engineers in managing their personal files followed common engineering work practices. Specifically, we find that the way that the engineers approach personal electronic file management parallels both the way that engineers commonly think and the collaboration and knowledge reuse requirements of the practice of engineering.

1. *Engineers follow a functional approach to file naming.* Engineers named files based primarily on the title of the document (75% of files) and the purpose or function of the files in a directory. This functional schema for file naming confirms our intuition that engineers would follow a function based approach to file naming, carrying over their practice of function-based design methodologies [Pahl and Beitz 1999]. Functional reasoning is a default mode of thinking; engineers routinely think of a product model in terms of a product's overall function and a set of sub-functions. This functional reasoning proclivity, evidenced by both standard engineering work practices and the way that the engineers already name their electronic files, suggests that functional file naming (i.e., name the file for what it would be used for) would be a suitable recommendation in a general framework for naming electronic files. In addition to this, it has been shown that 30% of directories use shared or non-unique namespaces. This again reflects the structured and very often prescriptive nature of engineering practices and suggests that a common framework for the naming and organization of directories and files may be a suitable remedy.
2. *File archival to support situated and case-based reasoning.* Engineering design involves the analysis and development of complex systems. A design case can be drawn from a complex series of experiences and decisions. Engineers store electronic files on their personal computers for the purpose of drawing together cases to support their engineering work. The data from this study suggests that design cases will be drawn from a broad range of personal electronic files. In general around 95% of all files used by engineers can be classified as Text, Spreadsheet, Presentation, Database, Project, Graphics, CAD, Code, Simulation, Application, Image, Audio, Video, Internet, Data and Compressed files. Text files and image files are the most commonly occurring file types, each representing an average of just over 20% of all files used by engineers. Following text and images, the next most prevalent file types are code, Internet, data, spreadsheet, CAD, presentation, simulation, and applications, which each represent 5% to 3% of total files. This broad range of files introduces challenges for information systems designers who are creating design rationale capture and retrieval systems [Regli et al. 2000] and case-based reasoning systems [Maher and Gómez de Silva Garza 1997]. The

reuse of design knowledge depends on two things. Firstly, the electronic files that explain why certain engineering decisions were taken must be successfully retrieved. This is the problem addressed by design rationale capture systems. Secondly, the integration of this information into a user-definable design case to support the current reasoning process, which is the problem taken up by case-based reasoning systems. No one system yet exists which can assist in both processes (rationale capture/retrieval and case-based reasoning) simultaneously. Furthermore, the many files that engineers store and access exist for more than co-coordinative practices as described by Schmidt and Wagner [2004]. In particular, they exist to support situated reasoning. That is, the collection of files are accessed to support an action applicable at a certain time in the engineering design work. Hence, information systems will need to support the rapid identification of a collection of situational relevant files and their organization into a customizable design case.

3. *File sharing for knowledge sharing.* More than half (60%) of all engineers frequently exchange and share files with colleagues who are participating in the same project. The data in Figure 4(d) shows that the engineers most frequently share their electronic files with others in the same department (40%) and less often with those working on the same project (30%) or in collaborating companies (30%). File exchange and sharing may also take place with customers, suppliers and contractors, but this was not extensively observed due to the characteristics of the sample population. More importantly, the evidence indicates that there is both a broad and wide distribution of knowledge as data contained in each the electronic files. That is, each engineer retains a deep base of knowledge in personal files (2.5Gb), but shares and accesses a broad range of knowledge with and from others. This pattern of electronic file sharing is consistent with engineering design practice. Communication of knowledge among different domain specialists is a fundamental aspect of engineering. The maintenance of the data files by engineers within a department also serves a secondary purpose, assisting other engineers in the department by filtering and providing information from past projects to those who need the data at specific junctions of the project. This role is what Sonnenwald [1996] called intradisciplinary stars, those who transmit knowledge within a discipline. Given the lack of negative feedback about the file sharing, it is plausible to infer that the cost of maintaining file sharing services is exceeded by the payoff. This should be considered not only in terms of the direct value of the knowledge itself, but also in terms of being able to control the circumstances of disclosure and to maintain personal relationships [Grudin 2001], with necessarily demand higher levels of trust and less misinterpretation in the shared knowledge [Tsoukas 1996]. Additionally, it appears that the engineers prefer to maintain a local copy of electronic files specifically so that they can be referred to at a future date and reused both by the engineer and for file sharing, which makes sense where the reliability and permanence of formal external stores are not always trusted [Whittaker and Hirschberg, 2001].

6. IMPROVING THE ORGANIZATION OF PERSONAL ELECTRONIC FILES

This section is concerned with the development of a remedy for the naming and organization of electronic files that balances the personal nature of data storage with the need for personal records to be accessed and reused by colleagues, possibly far into the future where the original owner is no longer available. Firstly, a summary of key approaches for improving file access and management is presented and critiqued with respect to the previously observed practices of engineers. The most apposite approach is then used as the basis for the development of a strategy for improving the organization and management of personal electronic files in engineering organizations. The possible benefits and potential barriers to the introduction of the proposed strategy within the context of engineering are then discussed.

6.1 Improving the Access and Management of Electronic Files

Improved file management can be achieved by a variety of means and in particular: conventions for file naming, file management systems, search tools (including indexing and classification) and visualization techniques. Each of these are discussed in turn below.

1. *De facto conventions and industry standards for naming files.* These standards either prescribe the name space or restrict the allowable character sets, reserved words and maximum length of file names. Examples of rule-driven name space conventions are those used for the management of large homogeneous sets of data files such as climate and meteorological data (NetCDF, http://cf-pcmdi.llnl.gov), and satellite images [TruEarth 2008]. A common example of restricted character conventions are those of ISO 9660 [ISO, 1988], which define a file system for CD-ROM media which supports Microsoft Windows, Mac OS, and Unix/Unix-like, and restricts both the length of file names and the allowed character set. In the engineering domain, no such standards exist that would be suitable for the naming of electronic files.
2. *Content and document management systems.* Document management systems support the creation, flow, storage, retrieval and archiving of electronic documents [Sutton 1996]. Development began in the late 1980s and early 1990s with the creation of Laserfiche [Compulink Management Center 2008] and PC DOCS [Hummingbird 2008] and there are now a large number of commercial systems in use in engineering design firms. Systems have evolved from initially providing document imaging and cataloguing to providing repositories [Szykman et al. 2000] for all unstructured documents including for example reports, emails, letters, and drawings. The repositories employ a formal design modelling language and standard representations of data to support interfaces for adding, editing and browsing repository entries. A formal design modelling language leads to a more comprehensive description of engineering design works, which in turn provides for meaningful indexing of engineering information. The repositories, however, deal with engineering data as information objects which are independent of the

actual electronic files within which they are contained. That is, an information object within the repository may be comprised of data coming from multiple electronic files. The repositories do not address the management of the files themselves.

3. *Search tools*. Commercial tools such as Google Desktop (http://desktop.google.com) automatically traverse a prescribed set of electronic files to create an index of all files and keywords based on the content of the files. This index can then be used to perform very fast keyword-based searches of files. Research prototypes such as Microsoft Research's Stuff I've Seen [Dumais et al. 2003] support both full-text search like Google Desktop and metadata based search, such as the name of the person (not necessarily the user) who created an information item. Implicit Query facilities [Dumais et al. 2004] allow contextually-similar information to be presented in addition to the "most likely relevant" file. The adoption of full-text search engines for personal computer files is problematic for engineering organizations due to the lack of indexable and searchable text in many types of files used (Table V). Finally, while these tools enable search based on content, they do not as yet support task-based or context-based search [Cuttrell et al. 2006].

   Whilst such full-text search tools can significantly simplify the problem of finding an electronic file, a prerequisite for their effective use is an understanding of the content of the files. To overcome the need for a relatively in-depth *a priori* knowledge of the document set, commercial classification tools have been developed [McMahon et al. 2004]. These tools require that each document be classified, usually by an expert or a keyword search, with respect to a predetermined ontology or taxonomy. These ontologies generally represent domain knowledge or a particular perspective (context) with the aim of providing the user with a common and more insightful view of the information space thereby reducing access time. The classification is then intended to assist in searching for information items. The remedy that we propose includes a simple taxonomy and ontology.

4. *Visualization techniques*. Tools such as Haystack [Karger and Quan, 2004] present views of the file space which emphasize certain dimensions or factors including graphical views of metadata such directory sizes and file sizes, color-coded views of file types, and graphical representations of modification history, and summary views of the data itself. They may also include utilities to notify the user(s) of the files currently being modified or the most frequently accessed files. They are normally incorporated with search tools, that is, present the results of a search for electronic files, rather than present a visualization of an entire set of personal electronic files. As such, they generally deal with the problem of locating a file rather than managing the personal electronic files for sharing.

These approaches all offer potential benefits, but also possess various limitations with respect to the management of personal files and, ultimately, a remedy to overcome the barriers to more effective reuse of personal electronic files across an engineering organization. In the case of standards, there exist no guidelines or recommendations in the engineering design community for the creation of naming conventions for personal electronic files. Standards such as STEP (ISO 10303) Product Data Representation and Exchange provide a means of describing engineering product data, but not the electronic files containing the product data. Where document management systems and product data management systems are available, our experience in consulting for engineering design firms in the UK is that their application tends to be limited to project files, file groups, or published documents and particularly those documents generated through collaborative working rather than personal electronic files. For these reasons, and cost considerations, such systems not generally implemented for personal or user file spaces, which, as the empirical results from this study show, can represent a significant amount of electronic information.

Given the limitations of these approaches and the fact that they can be considered to address the consequences of poor file organization rather than the cause (i.e., inconsistent and unrepresentative naming and organization of the electronic files when first generated), it is proposed to develop conventions and standards for naming and organizing files as a remedy to the reuse of personal information within the context of engineering.

6.2 A Shared Scheme for Directory Organization and File Naming

As discussed previously, the results of the study suggest that the reuse of personal information in an engineering environment (and hence knowledge sharing) can be facilitated through a shared scheme for directory organization and file naming. However, the issues surrounding the development of a generic scheme are complex and unclear. Arguably, the primary purpose of a filename is to represent the subject matter or context of the information contained within the file—to better support recall and recognition activities associated with the file's retrieval [Lansdale 1988]. When naming files, users may deliberate at length to decide a name which best represents the information content, purpose or subject. Filenames which fail to achieve this may make identifying and retrieving particular files unnecessarily time-consuming. It is known in engineering that there exist normative ways of reasoning about engineering problems, such as mapping between function, structure, and behavior of the design work [Gero 1990], functional reasoning [Far and Elamy 2005], and as hierarchical systems. These aspects however, are likely to be dependent on the particular organization, its design process and the artifact being designed. Consider for example, the design of a commercial aircraft vs. a mobile phone. The scale, structure, complexity, and life cycle of the two artifacts are considerably different and are hence likely to require the adoption of different schemes. For these reason, this section discusses the key 'baseline' features that a shared scheme should support, including:

—The requirement of engineers to use more than one computer.

—The requirement to support the identification and recall of files based on criteria that include: specific context, keywords, project title, author and project team.

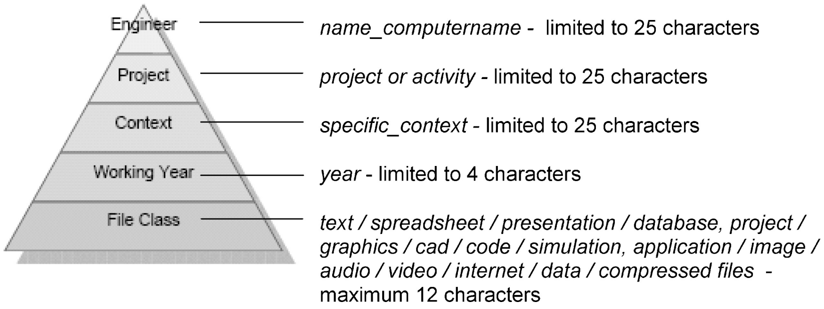

Engineer
Project
Context
Working Year
File Class
name_computername - limited to 25 characters
project or activity - limited to 25 characters
specific_context - limited to 25 characters
year - limited to 4 characters
text / spreadsheet / presentation / database, project / graphics / cad / code / simulation, application / image / audio / video / internet / data / compressed files - maximum 12 characters

Fig. 10. The five level directory structure.

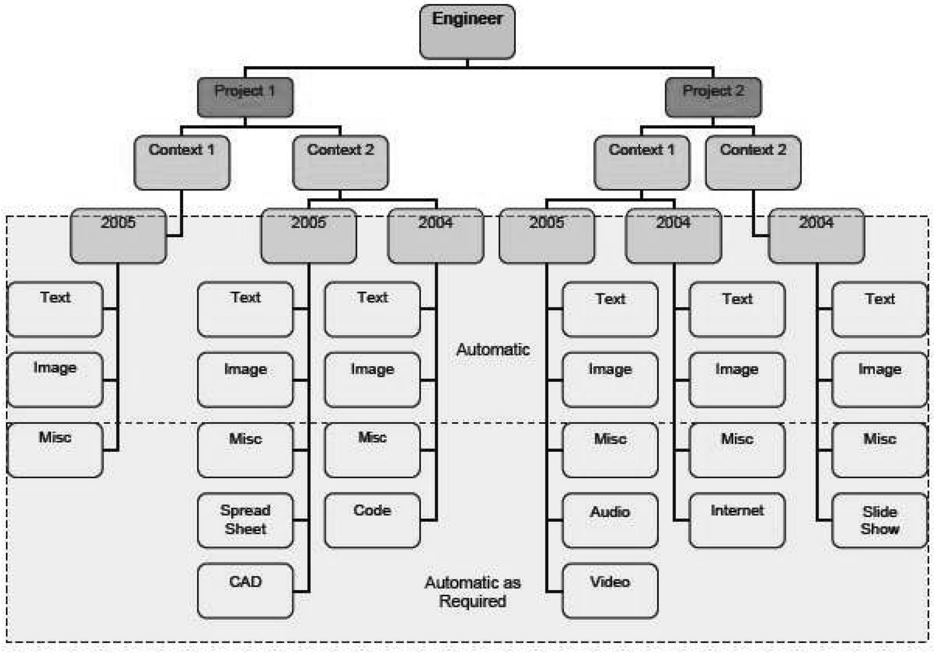


Fig. 11. A hierarchical model of the five level directory structure.

—The practice of engineers to work in levels three, four and five of their directory structure.

—The need to support sixteen fundamental classes of file used by engineers.

—The ability to verify (self-check) the location of a file within the directory structure.

—The need to reduce duplicate filenames.

—The need to support archiving, backup and synchronization of files.

These key features are each discussed with reference to one proposed scheme for directory organization and file naming. This scheme is shown conceptually in Figure 10 and hierarchically in Figure 11.

6.2.1 *Multiple Computers.* In order to satisfy the requirement for using multiple computers the name of the top level directory could be specific to the individual engineer and a computer (drive) or, where file spaces are shared, the project team since design knowledge reuse occurs largely through social knowledge networks [Demian and Fruchter 2006]. For example, in the scheme shown in Figure 10 the uppermost directory of a user's file space might is labeled name_computername. In this manner, files and complete directory structures from multiple computers can be managed without the need for renaming and assimilation of directories.

6.2.2 *Identification and Recall of Files.* The aim of the proposed remedy is to improve knowledge reuse through access to the personal files of others and facilitate rudimentary case-based reasoning (i.e., determining if a file could be relevant to the current problem) simply by reading the name of the file and the directory within which the file is stored. The empirical results are used to guide our understanding of the electronic files as mediators of coordinated and collaborative activity in engineering design [Suchman 2000]. The study has shown that for engineers, the most commonly used cues for identification and recall of files are: specific context, keywords, project title, the name of the author (engineer) and the name of the project team. As a consequence it is desirable for these elements, or part therefore, to appear explicitly in the filename and/or the directory path.

6.2.3 *The Practice of Using Five Level Directory Structures.* As stated in the previous section, there is a requirement to include elements such as specific context, keywords, project title, author and project team in the directory path or filename. However, the study has also shown that the current practice of engineers is to work in the third, fourth and fifth levels of their directory structure. It is therefore desirable to restrict a shared scheme to five levels. Given this constraint and the need to provide a means for organizing sixteen classes of file there remain only three levels of the directory structure available for organizing files with respect to the aforementioned elements. One possible scheme would be to label the second and third levels of the directory structure according to a function-oriented decomposition of the design work, which is congruent with the empirical data found in this study and other frameworks proposed for the capture of rationale about product data [Anthony at al. 2001]. The second level—shown in Figures 10 and 11—should include subdirectories named by each project or activity, which should indicate the component or system being engineered (e.g., bushing or tolerance_ring) or the particular activity (e.g., makefiles or risk analysis). The third level of the hierarchy is for the specific functional context, for example, report or research. In addition, where appropriate, the name should include what the system or component does (e.g., research_dampening). The combination of subject, component, and functional naming has been shown in other studies to be of most value in improving the automated information retrieval of engineering documents [Yang et al. 2005]. This leaves one remaining level of the directory structure (level 4), which should be used to provide an indication of currency. Hence, in the scheme shown in Figure 10 the fourth level of the structure is labeled according to year.

6.2.4 *Support for the Sixteen Fundamental Classes of File.* One way to improve the organization and management of the sixteen classes of file identified in this study is to logically group them according to class. Hence, in the scheme illustrated in Figures 10 and 11, the fifth and final level of the hierarchy provides containers for specific file classes: Text, Spreadsheet, Presentation, Database, Project, Graphics, CAD, Code, Simulation, Application, Image, Audio, Video, Internet, Data, and Compressed files. Separating the files by type facilitates the creation of Product Data Management (PDM) and Product Lifecycle Management (PLM) systems that associate CAD models to related documents, thereby embedding significant semantic content in the product model rather than specific application programs which access these files [Dong and Agogino 1998]. It is suggested that given almost 50% of files can are text and images that subdirectories for these two classes of files be created as standard and the remaining directories be created as and when required.

6.2.5 *Verification of File Location.* Within an engineering environment which is largely team based, there is a requirement to ensure that the file systems provide a mechanism by which files can be organized such that their position within the structure can be unambiguously determined and where necessary verified. Such a task might normally be performed by a moderator or librarian. However, for the purpose of organizing electronic files, it is desirable to provide an inherent mechanism within the strategy that is congruent and consistent with the function-based file naming preference and "way of thinking" about files uncovered by the empirical data reported previously. A prescribed organizational structure for a file hierarchy could assist in providing context about an information item [Karger and Jones 2006], which would then facilitate a shared understanding of the context within which a project was undertaken [Coughlan and Johnson 2008]. This can be achieved by ensuring that the directory naming convention, file naming convention and organization are self-consistent and in some manner are auditable.

6.2.6 *Reduction of Duplicate Filenames.* As stated in the previous section, it is desirable to determine a file naming convention and scheme for directory organization that are self-consistent. It has also been shown that engineers frequently use shared names for both files and directories which frustrates the processes of recall and recognition. One way to address these two dimensions is to adopt a file naming convention that parallels the structure of the directory. For the proposed scheme in Figure 10, this would involve the construction of a filename that included the engineer or project teams' name, the project title, the context, working year and a description of the file separated by, for example, underscores or dashes.

6.2.7 *Archiving and Backup.* In addition to organizing the working directory structure, it is necessary to support archiving, backup and synchronization of files across the multiple computers used by an engineer. As previously stated in 4.1.1, within the context of engineering, archiving tends to be undertaken for either a particular project or a specific customer, whilst backup is performed at regular time intervals for a given file system(s). In general, the tasks of archiving and backup involve the generation of an instance of a drive, a file system or a portion thereof at a given time, either as part of a backup procedure, or at the end of a project or financial period in the case of archiving. Effectively, these processes can be considered to add another level to the top most level of the existing directory structure. For archiving files this could be based on the year in which the project finished or the end of a financial period, whereas for backing up files this might be based on the identifier/name of the computer/drive.

6.3 Potential Barriers to the Introduction of a Shared Scheme

The remedy suggested is an attempt to rationalize the namespace and directory hierarchy for personal collections of electronic files in an engineering organization. While the remedy has not been implemented and evaluated to ascertain the acceptability of a prescribed electronic file collection organizing structure, our aim is to suggest a potential solution to the engineering information management community based on empirical evidence on the use of personal electronic files. Further, there is evidence to suggest the potential for the uptake of such a remedy. Evidence from a study by Jones et al. [2005] reported that their users already attempted to reuse predefined directory structures across similar projects, suggesting that engineers within an organization working on similar engineering projects would likely make use of prescribed directory structures when those structures correspond to the way they think of their projects. Second, Boardman and Sasse [2004] reported that users are likely to devote time and attention to organizing files that they authored, since they feel a sense of ownership over the files. Thirdly, where engineers are considered, many working practices—such as the use of Bills of Material (BOM), and supportive tools, such as PLM systems [Siemens Inc. 2008]—follow hierarchical structures and standardized naming conventions. Thus, whilst it is not a priori obvious that engineers will follow a prescribed paradigm for naming personal electronic files—even if adherence to these paradigms would facilitate both the sharing of the files and the capability to reuse them—a remedy which uses a prescribed naming scheme is consistent with working practices that are widely adopted within engineering.

The remedy is intended to rationalize the storage of personal electronic files under the consideration that these files might later on be of use for others. Given that more than 60% of the engineers reported frequently sharing their files, the issue of managing personal information items to be shared with others is a relevant one. The remedy does not solve the problem of unifying [Karger and Jones 2006] and contextualizing this information by others once this information is found. The unifying problem is being addressed by projects such as Haystack [Adar et al. 1999] whereas the contextualizing issue is being addressed by projects such as Sonic Sketchpad [Coughlan and Johnson 2008] for creative practitioners such as design engineers.

In addition to the potential barriers to the uptake of the proposed remedy, the adoption of a prescribed scheme will impact the distribution of files within the file system. In particular, all files will be contained within the leaf directories (level 5) and will require navigation through four directory levels. However, the additional semantic content provided by the directory levels and associated labels (names) should assist the user in locating relevant files, since conventions for the names of surrounding files and directory names suggesting the content of files provide context for orienteering [Teevan et al. 2004] to find other files. Further, the additional levels (project/activity, context, year and file class) should discourage users from placing large numbers of similar files such as images, in a single flat directory. This remedy differs from Bergman et al. [2006], who recommended that all project-related files should be stored in a single directory. Others studies, such as Boardman and Sasse [2004], find that users may have different organizational needs based on the type of work supported by different tools, and the concomitant electronic files produced, suggesting that neither the single hierarchy or the multiple hierarchy model is necessarily "right." The single convention for the file directory hierarchy may beneficially, however, reduce the duplicate hierarchy problem [Boardman et al. 2003].

For the purpose of developing the proposed remedy, it was determined that there was a requirement to ensure that the directory naming convention, file naming convention and organization are self-consistent in order to provide a means for self-checking (verification) of a file's location. Whilst such functionality is desirable within large collaborative teams, where responsibility for managing files is distributed, the resulting file names are lengthy. For example, the scheme shown in Figure 10 gives rise to a maximum path length of 91 characters excluding directory switches, drive name, and filename. If the filename also includes these various elements, then the filename could be around 108 characters and generate a path (uniform resource identifier) excluding computer and drive information of 200 characters. Whilst these values are significantly below those imposed by ISO 9660 and the operating systems, they will significantly impact on speed of naming and file manipulation. These lengthy paths and filenames could make recognition and recall difficult and increase the cognitive load so that the user can focus on the engineering work rather than on managing electronic files associated with the engineering work [Oviatt 2006]. However, if a prescribed paradigm is followed and the files are listed alphabetically, then recognition and recall arguably become easier through habituation. Further, it may be possible to embody the Name, Project, Context, and Year elements of the file name into the meta data of the file itself, thereby reducing the cognitive load necessary to recognize particular file names.

In summary, it is unlikely that a prescribed convention for naming and organizing personal electronic files will provide an optimal solution by itself for managing files in personal electronic information collections—that is, reducing cognitive load, improving file access and universality. Rather, it is envisaged that the adoption of good practices for naming and organization could deliver benefits in terms of storing personal information to assist in the reuse and sharing of the personal electronic files with others.

7. CONCLUSIONS

This article deals with the management of personal electronic files and in particular the naming and organization of directories and the computer-based files used by engineers on their personal computers. The issues associated with managing electronic files in today's commercial environments—where the volume and diversity of information is continually increasing—are discussed and the need to improve the organization and access of personal electronic files within shared and individual file spaces is described. In addition to this, the lack of guidelines or conventions for naming and organizing files is highlighted and the limitations of a variety of existing approaches and tools are discussed.

It follows that the proposition of this paper is that many of the issues associated with the management of engineers' personal electronic files could be overcome or at least alleviated through the development of shared schemes for directory organization and file naming. Central to the development of any scheme is the need to understand the practices and requirements of engineers for managing personal electronic files. In order to achieve this, a detailed study and audit of the file spaces of 40 engineers is presented. The study involved a survey of the personal strategies of engineers and an analysis of their file spaces with a software-based audit tool developed especially for this study. As this study focused on the management of personal electronic files, the research method did not explicitly consider why the engineers stored the electronic file, or the task(s) that necessitated the retrieval of personal electronic files or their sharing with others. Future studies employing a task-based method as done in the study by Elsweiler and Ruthven [2007] could uncover the rationale behind the behaviors.

The findings of the study reveal that:

- Engineers generally use two computers; a desktop and a laptop.
- 70% of engineers archive older electronic files and around half (55%) back-up files.
- The average size of the directory structure of an engineer is 2.5 gigabytes, including 406 directories and 4,660 files.
- Over 60% of engineers frequently exchange and share files with colleagues involved in the same project.
- Almost 95% of all files used by engineers can be classified as text, spreadsheet, presentation, database, project, drawing, cad, code, simulation, application, image, audio, video, internet, data and compressed files.
- 20% of all files are text files, 20% are image files, and between 3–5% are code files, Internet files, data files, spreadsheet files, CAD files, presentation files, simulation files and applications.
- The most common criteria for naming directories are the purpose or function (85%), the name of the project (55%) and the date (20%).
- The directory structure includes an average of nine hierarchical levels with 42 subdirectories and 495 files per level.
- The third, fourth and fifth levels of the hierarchy hold the majority of all files (53.5%).

- —The most common criteria used to name files are document title (75%), purpose or function (60%), project title (50%) and date (45%).
- —Over 30% of files have a duplicate name. A similar number of directories also share a duplicate name.
- —An engineer will modify 17 files from 8 different directories in a day, 50 files from 22 directories in a week and 1500 files from 295 directories in a year.
- —The most useful and/or commonly used mechanisms for locating files are context, followed by filename and keywords, then project team and author.

What is most interesting is that the personal file naming practices of the engineers followed their function-based reasoning work practice and training, as well as the personal knowledge reuse and sharing required by engineering design. It may not be surprising that electronic file naming practices follow the way that engineers, or any professional, thinks and reasons about their work. Since engineers tend to think functionally about components and systems, it is not surprising that the file names follow this convention.

The understanding gained from the findings of this study are used to develop the key features of a shared scheme for organizing personal directories and naming files. These features are illustrated through a five level directory naming convention that is based on a root directory (highest level of a user's file system) which corresponds to a particular engineer. This is followed by a project level, a context level, a working year and subdirectories corresponding to the sixteen file classes identified in this study. A file naming convention is also proposed, and is self-consistent with the directory naming convention and enables the location of the file within the hierarchy to be cross-referenced (verified). The file naming convention includes the name of the engineer or project team, followed by the project name, the context, the year and a description. For the purpose of backing up and archiving, the five level directory structure is extended to six to include an additional top level (root) directory corresponding to the computer name or the year respectively. Various limitations on the lengths of these elements are also defined in order that the convention conforms to ISO standards and hardware restrictions.

The key features of a shared scheme have been developed from an understanding of the fundamental classes of file used by engineers, a detailed study of current practices for organizing files and the most commonly used criteria for searching and indexing files. It is intended that the scheme be sufficiently generic for engineering tasks per se and also flexible enough to incorporate context specific information for particular users and domains. It is argued that the scheme should afford meaningful navigation around shared file spaces and the file spaces of other users. This will not only improve the access and retrieval of working files but also support more efficient and effective management and in particular, archiving and disposal. It is also proposed that if such conventions are adopted they could be used in combination with indexing, classification and visualization tools to further enhance their overall benefits to engineering work. Finally, although the work reported in this paper is concerned with mechanical engineers, the issues dealt with are relevant to all knowledge based industries (in particular those which involve teams of knowledge workers) and the scheme could be readily adapted to the needs of these other domains through a similar process to the one reported in this research.

REFERENCES

ADAR, E., KARGER, D., AND STEIN, L. A. 1999. Haystack: Per-user information environments. In *Proceedings of the 8th International Conference on Information and Knowledge Management (CIKM'99).* ACM Press, New York, NY, 413–422.

AGRAWAL, N., BOLOSKY, W. J., DOUCEUR, J. R., AND LORCH, J. R. 2007. A five-year study of file-system metadata. *ACM Trans. Stor. 3*, 3, 32.

ANDERSON, C. J., GLASSMAN, M., MCAFEE, R. B., AND PINELLI, T. 2001. An investigation of factors affecting how engineers and scientists seek information. *J. Engin. Techn. Manag. 18*, 2, 131–155.

ANTHONY, L., REGLI, W. C., AND JOHN, J. E. 2001. An approach to capturing structure, behavior, and function of artifacts in computer-aided design. *J. Comput. Inf. Sci. Engin. 1*, 2, 186–192.

BARREAU, D. AND NARDI, B. A. 1997. "Finding and reminding" revisited: appropriate metaphors for file organization at the desktop. *ACM SIGCHI Bull. 29*, 1, 76–78.

BARREAU, D. AND NARDI, B. A. 1995. Finding and reminding: File organization from the desktop. *ACM SIGCHI Bull. 27*, 3, 39–43.

BERGMAN, O., BEYTH-MAROM, R., AND NACHMIAS, R. 2006. The project fragmentation problem in personal information management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'06),* R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries, and G. Olson, Eds. ACM, New York, NY, 271–274.

BOARDMAN, R. AND SASSE, M. A. 2004. Stuff goes into the computer and doesn't come out: A cross-tool study of personal information management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'04)*. ACM Press, New York, NY, 583–590.

BOARDMAN, R., SPENCE, R., AND SASSE, M. A. 2003. Too many hierarchies? The daily struggle for control of the workspace. In *Proceedings of the 10th International Conference on Human-Computer Interaction*. Lawrence Erlbaum Publishers, New York, NY, 616–620.

CHAFFEY, D. AND WOOD, S. 2004. *Business Information Management, Improving Performance Using Information Systems*. Addison-Wesley.

CLARKSON, J. AND ECKERT, C. 2005. *Design Process Improvement: A Review of Current Practice*. Springer-Verlag.

COMPULINK MANAGEMENT CENTER INC. 2008. Laserfiche document management. http://www.laserfiche.com/resources/basics/index.html.

COUGHLAN, T. AND JOHNSON, P. 2008. Personal information management for creative practitioners. In *Proceedings of the Workshop: Personal Information Management*. ACM Press, New York, NY.

CRABTREE, R. A., FOX, M. S., AND BAID, N. K. 1997. Case studies of coordination activities and problems in collaborative design. *Resear. Engin. Des. 9*, 2, 70–84.

CULLEY, S. J., COURT, A. W., AND MCMAHON, C. A. 1992. The information requirements of engineering designers. *Engin. Des. 18*, 3, 21–23.

CULLEY, S. J. AND MCMAHON, C. A. 2005. Perspectives on knowledge management in design, engineering design. Theory and practice: A symposium in honour of Ken Wallace. Cambridge, Engineering Design Centre, University of Cambridge, UK.

CURTIS, G. AND COBHAM, D. 2005. *Business Information Systems: Analysis, Design and Practice*. Harlow, Financial Times Prentice Hall.

CUTRELL, E., DUMAIS, S. T., AND TEEVAN, J. 2006. Searching to eliminate personal information management. *Comm. ACM 49*, 1, 58–64.

DEMIAN, P. AND FRUCHTER, R. 2006. An ethnographic study of design knowledge reuse in the architecture, engineering, and construction industry. *Res. Engin. Des. 16*, 4, 184–195.

DIETEL, J. E. 2000. Improving corporate performance through records audit. *Inf. Manag. J. 34*, 2, 18–24.

DONG, A. AND AGOGINO, A. M. 1998. Managing design information in enterprise-wide CAD using 'smart drawings'. *Comput. Aid. Des. 30*, 6, 425–435.

DUMAIS, S., CUTRELL, E., SARIN, R., AND HORVITZ, E. 2004. Implicit queries for contextualized search. In *Proceedings of the International Conference on Research and Development in Information Retrieval*. ACM Press, New York, 594.

DUMAIS, S., CUTRELL, E., CADIZ, J., JANCKE, G., SARIN, R., AND ROBBINS, D. 2003. Stuff I've Seen: A system for personal information retrieval and reuse. In *Proceedings of the International Conference on Research and Development in Information Retrieval*. ACM Press, New York, 72–79.

EDMUNDS, A. AND MORRIS, A. 2000. The problem of information overload in business organizations: A review of literature. *Int. J. Inf. Manag. 20*, 17–28.

ELSWEILER, D. AND RUTHVEN, I. 2007. Towards task-based personal information management evaluations. In *Proceedings of the 30th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (*SIGIR'07*). ACM, New York, NY, 23–30.

ERICKSON, T. 2006. From PIM to GIM: Personal information management in group contexts. *Comm. ACM 49*, 1, 74–75.

FAR, B. H. AND ELAMY, A. H. 2005. Functional reasoning theories: Problems and perspectives. *AI Engin. Des. Manu. 19*, 75–88.

GERO, J. S. 1990. Design prototypes: a knowledge representation schema for design. *AI Mag. 11*, 4, 26–36.

GROTH, K. AND EKLUNDH, K. S. 2006. Combining personal and organizational information. In *Proceedings of the SIGIR Workshop: Personal Information Management*. ACM Press, New York, NY, 2006.

GRUDIN, J. 2001. Desituating action: Digital representations of context. *Hum.-Comput. Interac. 16*, 269–286.

HICKS, B. J., CULLEY, S. J., ALLEN, R. D., AND MULLINEUX, G. 2002. A framework for the requirements of capturing, storing and reusing information and knowledge in engineering design. *Int. J. Inf. Manag. 22*, 4, 263–280.

ISO 9660. 1988. International Organization for Standardization Information processing – Volume and file structure of CD-ROM for information interchange.

JESSUP, L. M. AND VALACICH, J. S. 2008. *Information Systems Today: Managing in the Digital World*. Pearson Prentice Hall, Upper Saddle River, NJ.

JONES, W. 2007. Personal information management. In B. Cronin, *Annual Review of Information Science and Technology*, 41.

JONES, W., PHUWANARTNURAK, A. J., GILL, R., AND BRUCE, H. 2005. Don't take my folders away!: Organizing personal information to get things done. In *Proceedings of the Extended Abstracts on Human Factors in Computing Systems (CHI'05).* ACM Press, New York, NY, 1505–1508.

KARGER, D. R. AND JONES, W. 2006. Data unification in personal information management. *Comm. ACM 49*, 1, 77–82.

KARGER, D. R. AND QUAN, D. 2004. Haystack: a user interface for creating, browsing, and organizing arbitrary semistructured information. In *Proceedings of the Extended Abstracts on Human Factors in Computing Systems (CHI'04).* ACM, New York, NY, 777–778.

KWASNIK, B. H. 1989. How a personal document's intended use or purpose affects its classification in an office. In *Proceedings 12th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'89)*. ACM.

LANSDALE, M. W. 1988. The psychology of personal information management. *Appl. Ergonomics 19*, 1, 55–66.

LYMAN, P. AND VARIAN, H. L. 2003. How much information. http://www.sims.berkeley.edu/how-much-info-2003.

MAHER, M. L. AND GÓMEZ DE SILVA GARZA, A. 1997. Case-based reasoning in design. *IEEE Expert 12*, 2, 34–41.

MCALPINE, H., HICKS, B. J., HUET, G. AND CULLEY, S. J. 2006, An investigation into the use and content of the engineer's logbook. *Design Studies 27*, 4, 481–504.

MCMAHON, C. A., PITT, D. J., YANG, Y., AND SIMS WILLIAMS, J. H. 1993. Review: An information management system for informal design data. In *Proceedings of the 1993 ASME Computer in Engineering Conference and Exposition*, K. H. Law, Ed. San Diego, CA, 215–226.

MCMAHON, C. A., LOWE, A., CULLEY, S. J., CORDEROY, M., CROSSLAND, R., SHAH, T., AND STEWART, D. 2004. Waypoint: An Integrated Search and Retrieval System for Engineering Documents. *J. Comput. Inf. Sci. Engin. 4*, 329–338.

OVIATT, S. 2006. Human-centered design meets cognitive load theory: designing interfaces that help people think. In *Proceedings of the 14th Annual ACM international Conference on Multimedia (MULTIMEDIA'06)*. ACM, New York, NY, 871–880.

PAHL, G. AND BEITZ, W. 1999. *Engineering Design: A Systematic Approach*. Springer, Berlin, Germany.

RACHURI, S., SUBRAHMANIAN, E., BOURAS, A., FENVES, S. J., AND SEBTI FOUFOU, R. D. S. 2007. Information sharing and exchange in the context of product lifecycle management: Role of standards. *Comput.-Aid. Des.*

RAVASIO, P. R., SCHÄR, S. G., AND KRUEGER, H. 2004. In pursuit of desktop evolution: User Problems and Practices with modern desktop systems. In *ACM Trans. Comput.-Hum. Interac.* (TOCHI) *11*, 2, 156–180.

REGLI, W. C., HU, X., ATWOOD, M., AND SUN, W. 2000. A survey of design rationale systems: approaches, representation, capture and retrieval. *Engin. Comput.*, *16*, 3-4, 209–235.

SCHMIDT, K. AND WAGNER, I. 2004. Ordering systems: Coordinative practices and artifacts in architectural design and planning. *Computer Supported Cooperative Work* (*CSCW*), Vol. 13, Springer, 349–408.

SIEMENS INC. 2008. Teamcenter Classification – Facilitating product and process reuse using Teamcentre's classification and reuse capabilities. http://www.siemens.com/plm/.

SONNENWALD, D. H. 1996. Communication roles that support collaboration during the design process. *Des. Studies 17*, 3, 277–301.

STEWART, T. A. 1994. Surviving information overload. Information Technology/Special Report, *Fortune* (July).

STEWART, T. A. 1997. Intellectual capital: The new wealth of organizations. *Knowl. Manag. 67*.

SUCHMAN, L. 2000. Embodied practices of engineering work. mind. *Culture Activity* 7, 1/2, 4–18.

SUTTON, M. J. D. 1996. *Document Management for the Enterprise: Principles, Techniques, and Applications*. John Wiley & Sons, New York, NY.

SZYKMAN, S., SRIRAM, R. D., BOCHENEK, C., RACZ, J. W., AND SENFAUTE, J. 2000. Design repositories: Engineering design's new knowledge base. *IEEE Intell. Syst. Appl. 15*, 3, 48–55.

TEEVAN, J., ALVARADO, C., ACKERMAN, M. S., AND KARGER, D. R. 2006. The perfect search engine is not enough: A study of orienteering behavior in directed search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Press, New York, NY, 415–422.

TSOUKAS, H. 1996. The firm as a distributed knowledge system: A constructionist approach. *Strategic Manag. J. 17*, 11–26.

TRUEARTH. 2008. TruEarth® 1km satellite imagery filename convention. http://www.truearth.com/prod_1km/filenames_content.htm.

WARD, M. 2001. A survey of engineers in their information world. *J. Librarianship Inf. Sci. 33*, 4, 168–176.

WHITTAKER, S. AND HIRSCHBERG, J. 2001. The character, value, and management of personal paper archives. In *ACM Trans. Comput.-Hum. Interac. 8*, 2, 150–170.

YANG, M. C., WOOD III, W. H., AND CUTKOSKY, M. R. 2005. Design information retrieval: a thesauri-based approach for reuse of informal design information. *Engin. Comput. 21*, 2, 177–192.

Received March 2007; revised February 2008, May 2008; accepted July 2008