PERSONAL INFORMATION ORGANIZATION AND RE-ACCESS IN COMPUTER FOLDERS: AN EMPIRICAL STUDY OF INFORMATION WORKERS

BY

HONG ZHANG

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Library and Information Science in the Graduate College of the University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Professor Linda C. Smith, Chair
Professor Michael Twidale, Director of Research
Professor Carole L. Palmer
Associate Professor P. Bryan Heidorn
ABSTRACT

The current hierarchical folder system has long been found limited causing various difficulties in organizing and re-finding information on personal computers. Many alternative prototypes have been proposed to replace the current folder system. However, past empirical studies consistently observed that people prefer browsing folders in re-accessing information and only use searching as the last resort. Recognizing the complexity and our limited understanding of personal information organization and retrieval behavior in computer folders, my study was aimed to explore what people need from folders and the affordances and limitations of folders in the different stages of organization and retrieval, and furthermore provide implications for system design. Improved understanding on personal information organization and retrieval on computers is especially important today when personal information management (PIM) has entered public domain and the boundary between personal information management systems and general information systems becomes blurred.

This multiple-case study investigated the participants’ information organization and retrieval behavior in their computer folders at four stages: keeping and discarding, organizing, re-organizing, and re-accessing. The difficulties they had at each stage were identified, and their computer folder structures and contents were analyzed. The participants include six PhD students and six administrative staff in an academic institution, with the former group representing the “research” end and the other one close to the “administrative” end in the activity spectrum proposed in (Bondarenko & Janssen, 2005). The data collection instruments include two rounds of in-depth semi-structured interviews, information re-access task observations, disk scans of several folders, and emails reporting re-access difficulties. The use of the two distinct groups of
participants together with the multiple data sources and data collection methods provided rich and varied data for exploration and at the same time increased the opportunity to do triangulation in data analysis.

The specific research questions include: 1) how do people keep and discard information items on computers, and what are on some people’s “messy” computer Desktops or in some “messy” folders? 2) From an integrative view, how do people organize information in computer folders, and what difficulties do they have in doing this? 3) What are the folder structures and contents like? 4) How do people re-organize folder structure and what difficulties do they have in doing this? 5) What are the tasks and strategies of re-accessing information on personal computers, and what difficulties do they have in re-accessing information?

The result implies that: 1) people need an in-between mechanism for keeping or not keeping, as well as for discarding or not discarding a particular information item; 2) behind all the idiosyncratic folder creation behaviors, four elements are identified in how the participants organize information based on Hjørland's typology of four views: rationalism, empiricism, pragmatism, and historicism; 3) the study identified two extreme types of folders in a spectrum – “genre folders” with no interfile relationships and “project folders” with complicated relationships between files, which can be partly attributed to the impact of using folders as workplaces. With all the three types of relationships identified in PREMIS observed in various “project folders” or folders close to this end, this study found that the various derivative relationships between files and/or groups of files led to the greatest difficulties for participants in finding and identifying files; 4) Behind the general browsing and search behaviors in folders, this
study observed a re-access strategy similar to faceted navigation. The content analysis of the re-access difficulty examples reveals that the four FRBR tasks (finding, identifying, selecting and obtaining) in searching and making use of bibliographies and library catalogues also exist in information re-accessing on personal computers. This helps to define the “re-access” behavior on personal computers, and suggests the importance of the other tasks in addition to “re-finding” information.
ACKNOWLEDGEMENTS

At first, I would like to thank my dissertation committee for their support and encouragement throughout this research process, specifically:

To Linda Smith, my advisor, who was willing to share knowledge and time without reservation; whose understanding and support helped me to learn and grow, and whose detailed advice made this study possible.

To Michael Twidale, my director of research, whose incisive vision and spirit of creativity helped shape this study, and whose enthusiasm and guidance made this journey a fruitful and joyful adventure.

To Carole Palmer, who guided and helped refine the study design, and who taught me respect for methodology and rigorous research.

To Bryan Heidorn, who worked closely with me in the process of research design, and who provided invaluable comments and suggestions at various stages of the study.

The committee members’ different research perspectives tremendously influenced the overall views adopted in this research project.

I must acknowledge my gratitude to the participants in the study. Their cooperation and support through this study made this study possible.
I would also like to thank some faculty members and doctoral students in the Graduate School of Library and Information Science who have influenced my thinking, especially to those I discussed my work with, e.g., Thomas Dousa and Ingbert Floyd, among many others. The discussions and insightful comments helped clear my thought during the exploration.

Finally, thanks to the Graduate School of Library and Information Science for the supportive environment and the financial aid they generously provided for conducting this study as well as through my entire doctoral program.
To my daughter Amy
# TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................... xi

LIST OF FIGURES .......................................................................................................... xii

CHAPTER 1: INTRODUCTION ......................................................................................... 1

  1.1 Background ........................................................................................................ 1
  1.2 Purpose and Research Questions ..................................................................... 3
  1.3 Terms and Definitions .................................................................................... 6

CHAPTER 2: LITERATURE REVIEW .............................................................................. 11

  2.1 Introduction ...................................................................................................... 11
  2.2 Information Keeping ....................................................................................... 12
  2.3 Information Organization ............................................................................... 18
  2.4 Maintenance .................................................................................................... 28
  2.5 Information Retrieval ..................................................................................... 29
  2.6 Context in Information Needs and Seeking Behavior ................................... 35
  2.7 Alternative Structures ..................................................................................... 38

CHAPTER 3: METHOD ..................................................................................................... 46

  3.1 Research Design ............................................................................................... 46
  3.2 Data Collection ................................................................................................. 49
  3.3 Data Analysis ................................................................................................... 55
  3.4 Considerations and Reflections ....................................................................... 57
  3.5 Evaluation Issues ............................................................................................. 64
  3.6 Lessons and Limitations .................................................................................. 65
<table>
<thead>
<tr>
<th>CHAPTER 4: KEEP AND DELETE</th>
<th>.................................................................</th>
<th>67</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Introduction</td>
<td>..........................................................................................................................</td>
<td>67</td>
</tr>
<tr>
<td>4.2 Keep and Delete Behaviors</td>
<td>.......................................................................................................................</td>
<td>67</td>
</tr>
<tr>
<td>4.3 Two Types of In-Between Keeping</td>
<td>..................................................................................................................</td>
<td>70</td>
</tr>
<tr>
<td>4.4 Delete</td>
<td>..........................................................................................................................</td>
<td>75</td>
</tr>
<tr>
<td>4.4 Conclusion and Implications</td>
<td>......................................................................................................................</td>
<td>78</td>
</tr>
<tr>
<td>CHAPTER 5: PERSONAL INFORMATION ORGANIZATION IN FOLDERS</td>
<td>..........</td>
<td>80</td>
</tr>
<tr>
<td>5.1 Introduction</td>
<td>..........................................................................................................................</td>
<td>80</td>
</tr>
<tr>
<td>5.2 Hjørland’s Typology of Four Views</td>
<td>....................................................................................................................</td>
<td>81</td>
</tr>
<tr>
<td>5.3 Four Elements in Information Organization Methods in Folders</td>
<td>..................................................................................................................</td>
<td>83</td>
</tr>
<tr>
<td>5.4 Problems with the Current Folder System</td>
<td>..................................................................................................................</td>
<td>92</td>
</tr>
<tr>
<td>5.5 Conclusion and Discussion</td>
<td>......................................................................................................................</td>
<td>95</td>
</tr>
<tr>
<td>CHAPTER 6: FOLDERS AS WORKPLACES AND ITS IMPACT ON FOLDERS</td>
<td>..........</td>
<td>97</td>
</tr>
<tr>
<td>6.1 Introduction</td>
<td>..........................................................................................................................</td>
<td>97</td>
</tr>
<tr>
<td>6.2 Folder Breadth and Depth</td>
<td>..........................................................................................................................</td>
<td>97</td>
</tr>
<tr>
<td>6.3 Information Organization Beyond Computers</td>
<td>..................................................................................................................</td>
<td>98</td>
</tr>
<tr>
<td>6.4 Two Types of Folders</td>
<td>..........................................................................................................................</td>
<td>99</td>
</tr>
<tr>
<td>6.5 Folders as Workplaces and the Impacts</td>
<td>..................................................................................................................</td>
<td>103</td>
</tr>
<tr>
<td>6.6 Conclusion and Discussion</td>
<td>......................................................................................................................</td>
<td>113</td>
</tr>
<tr>
<td>CHAPTER 7: FOLDER RE-ORGANIZATION</td>
<td>.................................................................</td>
<td>114</td>
</tr>
<tr>
<td>7.1 Introduction</td>
<td>..........................................................................................................................</td>
<td>114</td>
</tr>
<tr>
<td>7.2 Why to Re-organize</td>
<td>..........................................................................................................................</td>
<td>114</td>
</tr>
<tr>
<td>7.3 Re-organization Behavior</td>
<td>..........................................................................................................................</td>
<td>116</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 3.1 Participants Overview ........................................................................................................ 50
Table 5.1 Key Features of Hjørland’s Typology in Folder Structures ................................................ 83
Table 8.1 Reporting Email Components ............................................................................................ 121
Table 8.2 Overview of Circumstances .................................................................................................. 124
Table 8.3 Reason Codes and Contents .................................................................................................. 129
Table 8.4 Mapping between Re-access on PC and FRBR Tasks ............................................................. 135
Table 8.5 Elements That Were Remembered and the Frequencies ...................................................... 137
Table 8.6 Re-access Strategies and the Results ...................................................................................... 139
LIST OF FIGURES

Figure 1.1 Research Questions and the Corresponding Chapters ........................................... 5

Figure 3.1 Data Sources ........................................................................................................ 47

Figure 3.2 Method Diagram ................................................................................................ 49

Figure 3.3 Bondarenko & Janssen Activity Spectrum (2005) .................................................. 51

Figure 4.1 Online-PC-Paper Triple ..................................................................................... 68

Figure 4.2 Keep & Delete on Personal Computers ............................................................... 78

Figure 5.1 Example “Meeting” Folders ............................................................................... 90

Figure 6.1 Numbers of the Top Level Folders (1st interview) .............................................. 98

Figure 6.2 Maximum Depth of Folders (1st interview) ......................................................... 98

Figure 6.3 Part of a “Genre” Folder .................................................................................... 99

Figure 6.4 Part of a “Project” Folder .................................................................................. 100

Figure 8.1 “What Was It” in Two Groups .......................................................................... 126

Figure 8.2 “What Was It” Overview ................................................................................ 127

Figure 8.3 Reasons in Two Groups .................................................................................... 129

Figure 8.4 Reasons Overview ......................................................................................... 130

Figure 8.5 Use Overview .................................................................................................. 136

Figure 9.1 Types of Information in PIS .......................................................................... 151
CHAPTER 1: INTRODUCTION

1.1 Background

It has been widely observed that people have difficulties in filing and re-finding documents on their own computers. When a new document is created or accessed, people have to decide whether to keep it or not to keep it, and, if to keep, where to put it in the folder structure. The cognitive difficulty in making the organization decision can result in procrastination in filing and deleting which can cause disorganization and clutter. With the falling costs and rising storage capacity in various forms and locations and the resultant growing number of documents on personal computers, it has become an increasingly noticeable problem to re-locate information within our own personal information spaces (Barreau, 1995; Erickson, 1996; McKenzie & Cockburn, 2001; Ravasio, Schär, & Krueger, 2004), especially compared the ease of finding information on the Internet with available search engine technology. At many times people do not even bother to look for documents on their own computers and search online again. Both the failure and work-around in re-locating documents indicate problems of current computer systems.

A main reason that causes the organization and re-locating difficulties has been believed to be the current hierarchical folder system. Studies have long recognized its limitations. For example, documents can only be filed at one place in the hierarchy and thus cannot be accessed according to a different criterion (Dourish et al., 2000); filing to a folder can hide information which reduces the visual cues for memory and recognition (Jones et al., 2005), among others. As the result, new and sometimes radically different prototypes such as topical, temporal, and spatial metaphors have been proposed as alternative systems (e.g. Adar, Karger, & Stein, 1999; Freeman
& Gelernter, 1996; Rekimoto, 1999; Robertson, Czerwinski, & Larson, 1998). In addition, recently semantic tags have been used in a variety of experimental systems to provide alternative or complementary ways of organizing and accessing information items (e.g., Arriaga, 2002; Dourish et al., 2000; Oleksik et al., 2009; Quan et al., 2003).

On the other hand, many empirical studies in the past decades consistently observed that people prefer browsing folders to retrieve documents and use search only as a last resort (e.g. Barreau, 1995), even when new advanced search tools are available (Barreau, 2008; Bergman et al., 2008). Recent studies also found that there are some advantages with the current folder system, and some features usually deemed as drawbacks sometimes seem to be beneficial to users. For example, the stable folder structure provides a familiar environment for personal information organization, and the hiding function of folders is good at clearing the clutter that would distract the user’s attention (e.g., Civan et al., 2008; Jones et al., 2005).

These seemingly contradictory findings about the folder system indicate the complexity and subtlety, as well as our limited understanding of personal information organization and retrieval behavior in computer folders. Indeed, we do not fully understand what people really need from the folder system in organizing and retrieving information and “where and how it is inadequate” (Henderson, 2004). There has not been much research that looks specifically at information organization and retrieval practices in the individual’s personal information space on desktop computers and examines particular structures and contents of people’s folders (Whittaker, 2011), although the importance of this knowledge has been noted (Gonçalves & Jorge, 2003). Furthermore, there is a lack of a perspective of integrating various findings from empirical
studies with one another, which is believed to be fundamental to improving our understanding (Ravasio et al., 2004). The folder system – the structure, content, combined with various naming and sorting mechanisms – deserves further investigations “before we discard folders as an outdated relic” (Jones et al., 2005).

(Bergman et al., 2004) lists four categories of questions that PIM research needs to address: (1) What problems do users encounter as they manage their information? (2) What are the main strengths and weaknesses of today’s PIM tools? (3) How should PIM tools be evaluated? And (4) how should PIM be concisely defined and what theories can guide PIM design? Although this observation was made several years ago, the questions are still largely unanswered. This study was motivated by questions (1) and (2) and partially by (4).

At this point when personal information management has entered the public domain of information management with various technologies such as social computing, and at the same time traditional information organization systems have been questioned both within and beyond personal computers, it is especially necessary and urgent to have a thorough understanding of personal information organization and retrieval in the computer folder system.

1.2 Purpose and Research Questions

The ultimate goal of the study is twofold: 1) to deepen the understanding of personal information organization and re-access behavior on their computers, and identify both efficacies and limitations of the current folder metaphor in different stages, e.g., keeping and discarding, organizing, maintaining, and re-accessing; and 2) to provide implications for system design, either for improving the current systems or for new alternative systems.
The specific research questions include:

1) How do people keep and discard information items on computers, and what are on some people’s “messy” computer Desktops or in some “messy” folders?

2) From an integrative view, how do people organize information in computer folders, and what difficulties do they have in doing this?

3) What are the folder structures and contents like?

4) How do people re-organize folder structure and what difficulties do they have in doing this?

5) What are the tasks and strategies of re-accessing information on personal computers, and what difficulties do they have in re-accessing information?

The four aspects of personal information organization and re-access behavior (keeping and discarding, organization, re-organization, and re-access) are basically corresponding to the four components in Barreau’s framework of information storage and retrieval system which includes acquisition, storage and organization, retrieval, maintenance, output (Barreau, 1995). It’s important to note that these four aspects are closely connected and wired together in many cases.

These research questions are illustrated in Figure 1.1 with the corresponding chapters.
Figure 1.1 Research Questions and the Corresponding Chapters

1. How do people keep and discard information items on computers, and what are on some people’s “messy” computer Desktops or in some “messy” folders?  
   *Chap4*

2. From an integrative view, how do people organize information in computer folders, and what difficulties do they have in doing this?  
   *Chap5*

3. What are the folder structures and contents like?  
   *Chap6*

4. How do people re-organize folder structure and what difficulties do they have in doing this?  
   *Chap7*

5. What are the tasks and strategies of re-accessing information on personal computers, and what difficulties do they have in re-accessing information?  
   *Chap8*
1.3 Terms and Definitions

Several terms used in this dissertation are clarified below. In this dissertation, all of them will be used in the digital environment.

1.3.1 Personal information management

This study is mainly in the context of Personal Information Management (PIM) research area. Lansdale defines personal information as the information that the individual owns, for his or her own use. He further defines PIM as the “methods and procedures by which we handle, categorize and retrieve information” (Lansdale, 1988). Peters extends the concept of “personal information” to “every bit of information that represents an actuality within our locally constructed reality” (Peters, 2002) in trying to set a boundary for the PIM research field. Based on Lansdale's definition, Henderson defines PIM as the process of acquiring, storing, managing, retrieving and using digital documents. She emphasizes that PIM is personal “in the sense that the documents are owned by the user and is under their direct control, not that they necessarily contain information about the user” (Henderson, 2004). Barreau developed a conceptual framework in which PIM includes acquiring, organizing and storing, maintaining, and retrieving, and outputting information (Barreau, 1995). Boardman and Sasse further declare PIM as “an umbrella term used to describe the four component sub-activities which are acquisition, organization and storing, maintenance and retrieval of digital objects by an individual for his/her own use” (Boardman & Sasse, 2004). In the CHI PIM SIG2004, PIM is defined to be “the storage, organization, and retrieval of information by an individual for his/her own use” (Bergman, Boardman, Gwizdka, & Jones, 2004). Jones (2007) provides a definition in that PIM
“refers to both the practice and the study of the activities a person performs in order to acquire or create, store, organize, maintain, retrieve, use and distribute the information” for various reasons.

Jones’ definition is adopted in this study, although the focus in this study is on the four components: storing (keeping and discarding), organizing, maintaining (re-organizing), and retrieval (re-accessing). The major difference between PIM and general information management (GIM) is that these activities are conducted by an individual for his/her own use.

PIM has three types of meanings in research: PIM as activity management; as information management; and as the technology or system. PIM System is the tools and environment where PIM practices are performed and by which the behaviors are shaped. PIM systems more often refer to digital information systems “designed for individual personal use” (Jones, 2007), and include systems supporting various information management, communication, and activity management. This study will focus on the information management part within the particular system on current personal computers, which is the folder system.

1.3.2 Personal information repository and personal information space

In (Jones, 2007), Jones identified six senses of information that is “personal” which include: owned by me; about me; directed to me; sent by me; already experienced by me; and useful to me. He further defines “personal space of information” (PSI) which contains “information that is personal (in any of six senses) together with information tools, objects, and constructs used to manage this information.”
Jones also uses “personal information collections” (PICs) to refer to “personally managed subsets of a PIS.” He explains that “(a) PIC includes not only a set of information items, but also their organizing representations, including spatial layout, properties, and containing folders” (Jones, 2007). For example, a PIC can be the papers in a filing cabinet or particular information items under a project-related directory on a computer.

``Personal document space”(Gonçalves & Jorge, 2003) or ``personal information repository” (Adar et al., 1999) has been used in various studies, generally referring to the repository of emails, electronic and even paper files, electronic bookmarks, and visited WWW resources. In this work, PIR contains any form of information saved on an individual’s computer(s), and can be extended to include digital information the individual has seen before and paper form files in a broad sense. Information in PIR is different from other information in that it has been seen and/or organized and stored by the individual (Lansdale, 1988).

Personal information space (PIS) in this study is used mainly in the sense of information collection (especially on the user’s personal computers) with or without certain structure which is influenced by an individual's needs and activities, software tools and hardware environment, institutional factors, and possibly the whole information collection itself. It encompasses not only PIR (which similar to Jones’ “personal collection of information”) but also the structure layout which produces a space instead of the repository itself. The concept of PIS in this study is similar to Jones’ definition of “Personal Space of Information” (PSI) (Jones, 2007) described above.
1.3.3 Information re-access

In this work, information re-access refers to the information retrieval behavior in PIS and its purpose is to find things that have been seen or found before (Jones et al., 2001); (Dumais et al., 2003). In addition to general IR problems, it has to deal with such requests as “Joe's final report he sent me in an email,” or “that music clip wave file I downloaded and listened to recently.” Jones, Bruce and Dumais use “keeping found things found” to describe the problem in web use, email, electronic and paper files re-access.

This concept will be refined in Chapter 8 for the specific tasks it can involve.

1.3.4 Folder system

In this study, the “folder system” or “folders” refers to the current folder hierarchy in file and email organization systems, typically in today’s Windows, MacOS, and Linux operating systems.

1.3.5 Information item

In (Jones, 2007), Jones defines information item as “a packaging of information” from the “information-as-thing” perspective. The examples of information items he gives include paper and electronic documents, email messages, web pages, web bookmarks, etc., and shortcuts or aliases to any of these items. They can be manipulated (e.g., created, copied, stored, retrieved, given a name or tag, transformed, or deleted).
A similar concept is “digital object” or “object” defined in PREMIS Data Dictionary for Preservation Metadata in which object (or digital object) is a “discrete unit of information in digital form” for the purpose of clustering attributes and clarifying relationships (PREMIS editorial Committee, 2008).

This study uses “information item” in the same sense but limited to a file in a file organization system or an email message in an email system because of the focus of this study. This study does not consider web page content unless it is downloaded to the local personal computer.

1.3.6 Workplace

Workplace is “a place … where work is done” (Merriam-Webster). This study uses this term as a metaphor to convey the phenomenon that a folder can be a “virtual” workplace for a task or a project on the personal computer.
CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The development of science and technology, especially with the increasing digitalization and networking, has led to an explosive growth of information. With the web, people could “access more information more easily than before” (Sellen & Harper, 2002) and collect a growing collection of information as their personal digital libraries (Fertig, Freeman, & Gelernter, 1996). Consequently, information overload has become a problem for many people. To make it worse, too much information demand, constant multitasking and interruption, inadequate support for planning, monitoring, and reminding, have resulted in people's cognitive overload (Kirsh, 2000). This situation makes Personal information management (PIM) an increasingly critical research area.

Improved PIM can not only save our time and energy in finding information, it also makes our everyday life and work more efficient and productive by facilitating and stimulating “better and more intensive utilization of information” (Stibic, 1980). PIM is also a fundamental building block to organizational knowledge management. It complements and contributes to the knowledge base and management in an organization (Burns, 2004).

There are mainly two categories of studies in PIM research: empirical studies and prototype design (Boardman & Sasse, 2004). A fairly large amount of empirical studies have been conducted in both paper-based and digital environments in pursuit of deeper understanding of people's behavior and needs. People's PIM behavior involves many subtle and complex
psychological factors, which has led to the analysis from the perspective of psychology or cognitive science. On the other hand, many prototype systems have been proposed in order to better support PIM activities.

Sections 2.2 to 2.5 summarize the studies in the four information organization steps (keeping, organization, maintenance, and re-access) and especially in the folder-based hierarchical systems on personal computers. Section 2.6 briefly reviews the related research in information seeking behavior. After that, several typical alternative prototype systems are reviewed in section 2.7.

Some of the reviewed issues will be revisited in Chapter 9 with findings and implications from this study.

2.2 Information Keeping

People acquire new information by active search or more passive encounter (Jones, 2004). There are several important issues involved in this stage.

2.2.1 Reasons to keep

The reason to keep can be traced back to the characteristic of human behavior that is to acquire and keep items of value, which is embodied in both the physical and digital domains (Boardman & Sasse, 2004). But with respect to a particular information item, people could have various reasons. For example, the individual believes it might be potentially useful at some time in the future (Jones, 2004), and by keeping it he/she will “be able to retrieve and use it in the future” (Lansdale, 1988). Another identified reason in addition to retrieval is that people store items as
reminders of the tasks they have to perform (e.g. Barreau & Nardi, 1995; Malone, 1983; Boardman & Sasse, 2004). In many cases people keep documents even though many of them are publicly available for a variety of cognitive or emotional reasons (Shapiro & Hughes, 1999; Whittaker & Hirschberg, 2001). In a study on office workers' paper archives (Whittaker & Hirschberg, 2001), Whittaker and Hirschberg identified several reasons for this behavior. First, people want to have “immediate access” to the relevant documents when they need to use them. The importance of immediate access was emphasized by the majority of interviewees and they felt that access delays reduced the usefulness of off-site information. The second reason identified is to remind about an important action related to the document, and the documents in outside repositories were not found to support reminding. The third reason is that they do not trust the stability of external stores. It could also be because the document has added value such as personal annotations. People are also found to keep for sentimental and emotional reasons. Most of these items are for potential re-access (Jones, 2004).

A kind of “irrational keeping” or “pack rat” behavior has been observed in people's information keeping. Boardman and Sasse found that participants continued to save bookmarks even though many were never used. A similar behavior was observed in email and paper archives (Boardman & Sasse, 2004). They also found that, among the categories (files, emails, software, Bookmarks, etc.) in the buildup of PIR, file collections were highly valued, followed by email collections, and Bookmarks were valued less and in smaller collections compared to file and email.
2.2.2 Evaluation prior to keep

People can acquire new information actively or passively (Isaacs et al., 2002). No matter whether the information is “pushed” to or “pulled” (gathered) by the individual, he/she has to evaluate the relevance and potential usefulness and decide whether to save or not (Isaacs et al., 2002; Kirsh, 2000. It can be a difficult task for the individual (Jones, 2004).

From a psychological perspective, Kirsh analyzes the process of evaluating new information in gathering stage and points out that information is an “experience good” and people may not know how valuable the information is before they use it. For information gathering, it “can be hard to know when you have enough information and when you should continue hunting.” Thus it is hard for people to “develop a coherent demand function for information so to sit down and plan their information gathering strategies” (Kirsh, 2000).

Pirolli and Card draw an analogy between information gathering and food-foraging in their proposed Information Foraging theory. In this theory, people allocate their attention in selecting and pursuing information items according to the balance of foraging costs and benefits, and “optimal forager” would obtain the best result for the effort they exert under the constraints of the environment (Pirolli & Card, 1999).

2.2.3 Problems in keeping and discarding

In a study analyzing the problems accompanied with keeping, Jones points out the problem of information fragmentation caused by separate PIM applications, which is when related information items are scattered in different locations and applications. He also categorizes the
various possible “keeping mistakes” in two groups: false positive (keeping useless information) and miss (not keeping useful information). He further proposes to use the theory of signal detectability to develop tools to solve this problem (Jones, 2004).

Jones also pointed out the dilemma people have in deciding whether to keep or not. People want to keep important items, useful items, and items that might be useful later. But if they keep too many items, it could compete for the limited attention (“attention poverty”) that people have, and possibly cause disorganization, which could keep the valuable information on their computers from even being noticed (Jones, 2004; Jones, 2007). The decision to keep or not is “essential” but “fundamentally difficult” because people have to understand the information item, understand their own information organization system, and sometimes anticipate possible future needs (Jones, 2007).

Similar issues exist when people discard information. In a study looking at a group of workers’ paper documents management behavior in preparation for a move, Whittaker and his colleagues (Whittaker, 2011) observed that people have difficulties in deciding what to discard (p. 16). The discarded documents constituted “once-valuable information that had become obsolete” as well as unread documents (p. 16). The former phenomenon seems unavoidable since “the value of particular information decreases” as “jobs, personal interests, or company strategy changes” (p. 16). The latter one was believed to be the result of two problems: information overload led to partial information processing with some “non-urgent information is never processed”; and deferred evaluation that allowed time for “better informed judgments about information utility” (p. 16).
In proposing the “demotion” principle with which information systems “should allow users to demote unimportant information items (making them less visually salient) so as to reduce distraction” (Bergman et al., 2010), Bergman and his colleagues designed a prototype called “GrayArea” (Bergman, et al., 2009) that allows users to demote unimportant files in a “gray area” at the bottom of the folder. Another related study conducted by Swan and colleagues looks at the containment of clutter in a family home environment (Swan et al., 2008). Although there was no prototype proposed, it is inspiring for the digital world.

2.2.4 Keeping behavior

The keeping behavior has been investigated primarily focused on Web information, e.g., the Keeping Found Things Found (KFTF) project (Jones, 2004; Jones, 2007; Bruce, Jones, & Dumais, 2004), and several other studies (e.g. Marshall & Bly, 2005; Marshall & Jones, 2006). A variety of keeping methods (e.g., send email to self; send email to another person, etc.) are observed in keeping Web information, and a set of functions are identified as the factors that influence the choice of method in different situations, e.g., portability, reminding, etc. (Bruce et al., 2004). A special keeping behavior “clipping” which is defined as “intentionally saving portions of published material” is investigated in a study on how people save and use encountered information (Marshall & Bly, 2005).

However, not much research has been conducted on keeping behavior on personal computers.

In (Jones, 2007), William Jones defines “keeping” as:
Decisions made and actions taken to relate current information (information at hand or under consideration) to anticipated needs. Decisions can include (1) “ignore, this has no relevance to me,” (2) “ignore, I can get back to this later”, and (3) “keeping this in a special place or way so that I can be sure to use this information later.”

Chapter 4 in this dissertation will take a detailed look at the third keeping decision on personal computers.

### 2.2.5 From paper to computer

Paper documents, as the traditional archive form, are still found to be a valuable resource for today's diverse PIM activities such as reminding, annotating, and even long-term archiving (Whittaker & Hirschberg, 2001). The value of paper archives partly derives from the affordances of paper, which include portability, tangibility, universal acceptance, and readability (especially for long documents) (Sellen & Harper, 2002).

But working with paper also precludes the wide variety of interactive functions available in the digital world, for example, spelling correction, email, searching, editing, calculation and translation. It has led to research efforts to reduce the incompatibilities between the paper and electronic domains. For example, (Newman & Wellner, 1992) describes a system using a computer-controlled projector and camera above a desk to incorporate paper-based information into the digital world. (Rao et al., 1994) describes a system called Protofoil using document imaging technology to allow paper documents to enter or return to the electronic world.
Many metaphors in paper-based archives are also adopted in digital systems. For example, the hierarchical folder system currently used in most operating systems is seen as the metaphor of paper file folders. Some researchers proposed interfaces mimic properties of paper in the digital counterpart. For example, the “piles” metaphor (Mander et al., 1992), digital “yellow sticky” reminders, and “Rooms” metaphor (Henderson & Card, 1986) that associates a set of documents with a task in a “room”.

2.3 Information Organization

2.3.1 Why to organize

People’s predisposition to classify and organize seems to be mysteriously “wired” into human nature. Anthropologist Brent Berlin (Berlin, 1992) discovered that in the tribes whose cultures developed “in geographical isolation over tens of thousands of years,” the “indigenous peoples not only create highly developed classification systems, but they do so in strikingly similar ways: with plant or animal ‘families’ divided into nested hierarchies, often using exactly the same categories (Wright, 2008, p22). And “while the details of individual classification systems vary widely, the structure of these taxonomies reveals remarkable similarities” (p.25).

This tendency to organization is also reported by a PIM researcher. When reflecting upon the experience of using a personal electronic notebook Proteus, Erickson noticed the “gradual addition of more and more layers of structure” in which one notebook was broken into sections, and then subsections were added, basically a hierarchical structure, although he “neither wanted nor needed the structure” at the beginning (Erickson, 1996).
“The world has structure and in order to make sense of the stream of experience we impose even further structure on it” (Kwasnik, 1991). Studies show that people need organization, that is, structure, of the documents and rely on the structure to browse to the document that is wanted (e.g. Boardman & Sasse, 2004; Kwasnik, 1991; Malone, 1983). People organize information not just for ease of finding it. It could also be for reminding, understanding, or other reasons such as creating a legacy, sharing resources, confronting fears and anxieties, and identity construction (Kaye et al., 2006). Ravasio et al. (2004) point out that people invest effort in organizing the hierarchical file system structure in order to engrave “the information’s content and context into the system,” and provide “an overview in a single glance.” They also note that “(a)n electronic file system...is also a space upon which people inscribe things they wish to be reminded of” (Ravasio et al., 2004). Jones further points out that people organize as a part of making sense of the information and possible use (Jones & Teevan, 2007).

Structure can also exist in bookmarks. Abrams found that most users thought of the information available on the web as divided into “my bookmarks” and “the cloud of unmapped sources” (Abrams, 1997; Abrams et al., 1998).

2.3.2 Folder structure

There has been very little research investigating specifically how people organize information in folders and what the folder structures are like (Whittaker, 2011) and the few available studies have generated some contradictory findings about the structure of personal file systems (Bergman et al., 2010). In investigating 11 computer scientists’ folder structures in different operating systems, Gonçalves and Jorge (2003) observed deep and narrow hierarchies with the
average folder depth of 8.45 and average folder breadth of 1.84. Boardman and Sasse (2004) observed an average folder depth of 3.3. In a quantitative study looking at the folder structure of 73 university employees using Windows operating system, Henderson and Srinivasan (2009) found broader and shallower folder structures, with average folder depth of 3.4 and folder breadth of 4.1. In terms of the files per folder, Gonçalves and Jorge (2003) and Henderson and Srinivasan (2009) found similarly small numbers: 13 in the former and 11.1 in the latter.

2.3.3 How to organize: factors

Although the ultimate purpose of filing a document is often to retrieve it easily (Lansdale, 1988), research observed that a variety of factors could affect how people organize and retrieve information items (Rao et al., 1994). For example, personal style (Whittaker & Hirschberg, 2001), job type (Lansdale, 1988), the perceived value of information, acquisition-related factors, and particular tools (Boardman & Sasse, 2004) could be determinants of choice in where to put a document and how to look for it later. Kwasnik compiled a list of dimensions by which people classify personal documents in offices. It includes situational (contextual) factors (e.g., physical location; task requirement), the use (e.g., to remind; to give a lecture), circumstances (e.g., an ongoing project), and time issues (e.g., frequency of use or age of document), in addition to semantic (topic) factors (Kwasnik, 1991). Barreau observed that similar dimensions also applied to computer hierarchical file systems (Barreau, 1995). In a study investigating how people tell stories about their digital documents, Gonçalves and Jorge (2004) found that the dimensions most commonly used were time, place, co-author, purpose, subject, other documents, format, exchanges, tasks, storage, and contents. In a study investigating file and folder names and file structure on users’ computers, Khoo et al. (2007) found that the most common types of folder
names were document type, organizational function/structure, and miscellaneous/temporary, and the labels for first-level folders tended to be task-based or project-based (Khoo et al., 2007). A task- or project-based organization method has been recognized as an important need in information organization (Kaptelinin, 2003).

Henderson’s (2009) study is one of the few that investigates the folder creation behavior on personal computers. Henderson categorizes folder creation behavior according to temporal features: “in advance,” “just in time,” and “cleanup” (p. 78). She also analyzes the folder names based on categorization of folder names captured by a file system snapshot program. The categories of folder names include Genre, Task, Topic, and Time. Interestingly, most of the subjects with the organization scheme starting with “task” were found to have high confidence in rigorously following the file system (Henderson, 2005).

Researchers found that different job content, personal strategies, and tools could lead to different document management and thus the role of project or task in digital document management could vary. Kidd (1994) made a distinction between knowledge workers, communication workers and clerical workers.

Instead of fitting people into these categories, Bondarenko and Janssen looked at the two types of activities they perform: research and administrative. And they argue that all information workers have both unstructured research and structured administrative activities in varying degrees in their work. They further reaffirmed in their study that document management is strongly related to task management for all information workers. They noted that the current digital document
management system mainly supports pre-structured procedural tasks while leaving the research type unsupported. They further suggest that “a document management system should place documents into their task-related context” (Bondarenko & Janssen, 2005).

2.3.4 How to organize: piles vs. files

In his seminal article (Malone, 1983), Malone described two strategies of information organization in paper world offices. One is piles of information arranged by physical location, and the other is explicitly titled and logically arranged files. He found that people tend to defer classifying documents, referred to by Lansdale as a “compensating strategy” (Lansdale, 1988), because of the cognitive difficulty in classification, which may lead to piles of paper documents in a typical messy office.

Consistent with Malone’s observation of “neat” and “messy” offices (Malone, 1983), Abrams et al. identified four bookmark management strategies: no-filers, creation-time filer, end-of-session filer, and sporadic filer (Abrams et al., 1998). A similar pattern was also found in email management. For example, frequent filer, spring cleaner, and no-filer which can be further divided into folderless cleaner and folderless spring-cleaner (Balter, 1997; Whittaker & Sidner, 1996. In some other studies, users are categorized into “cleaners” and “keepers” (Gwizdka, 2004), or “pro-organizing” and “organizing neutral” (Boardman & Sasse, 2004). In a cross-tool study, Boardman and Sasse found that most of the people are a “combination of frequent filer, spring cleaner, and no-filer” in digital files, emails, and bookmarks (Boardman & Sasse, 2004), which is consistent with the above observation by Whittaker and Hirschberg on “filers” and “pilers.”
The issue of “piling” vs. “filing” has been further discussed by later researchers. Mander and his colleagues recognized that people use piles to quickly and informally organize the information as complementary to more formal files (Mander et al., 1992; Rose et al., 1993). In a study on personal paper information from 50 office workers in a research laboratory, Whittaker and Hirschberg found that “the distinction between filers and pilers was one of degree,” and “all people filed some information and pile some other information” (Whittaker & Hirschberg, 2001). They also identified several advantages of “piling” over “filing”. It was found that filers amassed more information and accessed it less frequently than pilers and they proposed that premature filing may be the reason. Compared to “filing”, “piling” demands less overhead while providing greater availability which is especially required for ongoing tasks, and “filing” does not always guarantee easy access. But the limitations of “piling” are that, people may need to invest more effort to find a specific item in “piles” than in “files”, and more importantly, it does not scale well (Whittaker & Hirschberg, 2001) and will be eventually “overwhelmed by the sheer volume of papers around the office” (Lansdale, 1988) as an information organization method. In addition to an information organization mechanism in an office, “piling” is also used as a reminding method, which has been recognized as a critical function in PIM (Lansdale, 1988). Although there are debates regarding simulating paper “pilers” in the digital world (Lansdale, 1988), studies by Mander and his colleagues showed that their prototype based on the “pile” concept was helpful (Mander et al., 1992; Rose et al., 1993).

From the perspective of psychology, Kirsh analyzed that users have different preferences and tolerance for the number and type of entry points (invitation to an information item) in their
offices. Those messy offices would have more entry points and more information about activities. But when the volume of information increases, they have to pay a cost in terms of “search time for the profusion and imprecision of their entry points”, although there might be chances for opportunistic discovery of useful information (Kirsh, 2001).

2.3.5 Pros and cons in current hierarchical folders on personal computers

Many studies have found that people have cognitive difficulty in organizing and naming information, be it paper files or electronic files, bookmarks, and emails (Abrams, 1997; Boardman & Sasse, 2004; Malone, 1983; McKenzie & Cockburn, 2001). It is widely believed that the strict and stable hierarchical structure in classification leads to the difficulty in retrieving (Quan et al., 2003) as well as filing (classification and naming for digital items) (Arriaga, 2002; Dourish et al., 1999; Malone, 1983). The inability to do multiple classifications restricts the possible retrieval path.

Another critical problem in traditional file systems is that the mechanism to describe a file is very limited and the relationships between related files are lost. For example, there is no effective way to contain every person's name on a picture (Quan et al., 2003; Soules & Ganger, 2004).

A related problem concerns the granularity of organization in the current file system. Henderson found that many people have the problem of managing different versions of documents (Henderson, 2004), and the current file-based systems treat two versions of documents as two separate files.
In a study trying to learn from paper to inform digital system design, Bondarenko and Janssen (2005) noted that the hierarchical structure of files and folders is not a natural way for organizing information people are still working on. Visual cues of paper documents on a desk as well as textual cues in an email about the document are gone, and thus people have to spend explicit effort to create a context for a document by naming it and choosing or naming a proper folder, which in many cases is insufficient, since the names of files or folders are usually not descriptive enough to aid their recognition. Barreau and Nardi (1995) found that people especially have problems organizing ephemeral information which is information needed for only a short time.

Another issue Bondarenko and Janssen noted is that the depth of the digital file folder structure can go much deeper than in a physical desk, and the structure is not transparent. People have to go through all levels to see the content of a subfolder (Bondarenko & Janssen, 2005). This visibility issue is also noted by Jones and colleagues: “folders can obscure as well as organize” (Jones et al., 2001).

It has been realized that it is important to connect information items in PIS. It is believed that the value of information “depended little on individual pieces of information, but rather on how the pieces were connected and thereby complemented one another” (Ravasio et al., 2004). One limitation of the current hierarchical logical file system is that it could not support networked organization and global view of PIR (Ravasio et al., 2004). It is also limited in supporting automatically and manually assigned context annotations. It has also been identified recently that the file systems should allow people to dynamically rearrange documents and give people the
flexibility to restructure their files in PIS (Bondarenko & Janssen, 2005; Marsden & Cairns, 2003).

Although with these limitations in current folder systems, studies found that users prefer to browse-and-locate information items in hierarchical file systems over direct keyword search (Barreau & Nardi, 1995). Quan et al. found that folders are useful for “file and forget” archiving and for organization schemes with fixed structures (Quan et al., 2003). Boardman and Sasse pointed out that the stable folder systems provide a familiar environment which is a good thing for personal information organization (Boardman & Sasse, 2004).

Sometimes a seeming drawback can be an advantage. For example, it is observed that sometimes information hiding is exactly what people want, tidying some files “out of the way” to enable the remainder to be more visible. Even the drawback that a document needs to be in only one place is sometimes an advantage because that means “we know exactly where to look for it later” (Civan et al., 2008).

2.3.6 Three types of information

Similar to Cole's finding for paper documents, which includes “action information,” “personal work files,” and “archived information” (Cole, 1982), Barreau and Nardi identified the existence of three types of electronic information which include “ephemeral”, “working”, and “archived” (Barreau & Nardi, 1995). They are defined according to short or long “shelf life” and/or relevance to current work:
Ephemeral information has a short shelf life and includes items such as (some) electronic mail messages, “to do” lists, note pads, memos, calendars, and news articles downloaded from databases.

... Working information is frequently-used information that is relevant to the user's current work needs and that has a shelf life of weeks or months. Working information is often created by the user or is the product of the user's work groups. ...

Archived information has a shelf life of months or years, but is only indirectly relevant to the user's current work. It is infrequently accessed. Most archived information represents completed work, including final reports and project histories. ...

A similar pattern has been observed in other studies on physical workplace (Pirolli & Card, 1999), digital documents (Nardi & Barreau, 1997; Ravasio et al., 2004), electronic bookmarks (Abrams, 1997), and screen space (Ravasio et al., 2004). Boardman and Sasse suggested using “active” (including ephemeral and working), “dormant” (inactive, potentially useful), “not useful”, and “un-assessed” (e.g., new emails) to avoid the misleading concept of “archived” (Boardman & Sasse, 2004). Sellen and Harper define “hot”, “warm”, and “cold” documents, where “hot” means those that one uses actively at the current moment, “warm” are those that were just in use or will be used in the nearest future, and “cold” are documents that are not used at the moment (Sellen & Harper, 2002). In a two-year ethnographic study of the management of hot and warm documents within 28 non-clerical information workers, Bondarenko and Janssen describe that: “If people need to search for a document this usually means that the document has been “cold” for a while and its location has been forgotten. Once found, the document is “at hand” (i.e., “hot” or “warm”) and therefore does not need to be searched for anymore (Bondarenko &
Janssen, 2005). They further noted that although “being suitable as long-term storage place for cold documents, a file folder structure does not support the needs of everyday document management of hot and warm documents” (Bondarenko & Janssen, 2005).

Different types of information may have different filing and re-access needs and strategies, and they are supported to varying degrees by current tools. Barreau and Nardi found that, for working information, it is “usually important enough to be organized by location and category in its own folder or location on the desktop. Users have no difficulty finding their working information as they use it repeatedly and thus can easily remember where it is.” “As projects near completion and the information is accessed less frequently, the categorical structure of the information becomes more important than the spatial location for organizing and finding files.” But for ephemeral information, the “central problem of organizing ephemeral information concerns where and how to file information that is needed for only a short time. ...With limitations on the amount of information that can be viewed on a screen at one time, managing large quantities of ephemeral information can be problematic. Users did not have perfect solutions to this problem” (Barreau & Nardi, 1995).

These “three types of information” concepts will be discussed in Chapter 9 with the findings and implications of this study.

2.4 Maintenance

Although people can invest various amounts of time on managing personal information (Burns, 2004), it has been found that many people devote little time to maintain collections beyond
occasional spring-cleans (Boardman & Sasse, 2004), while it takes effort to maintain a hierarchy of folders (Jones et al., 2001). Since people's PIRs are developed over time, and people tend to group information items into a hierarchy of folders as the number increases, no matter whether in files, email, or bookmarks (Jones et al., 2001), the previously constructed hierarchy structure may become unsuitable for upcoming information items. As a result, it has been found that some users “went to great lengths to maintain a single organizational scheme” (Jones, Bruce, & Dumais, 2003).

2.5 Information Retrieval

Information re-access is a special characteristic of information retrieval in PIM. With the prevalence of digital and network information services and business, a lot of living, working, and entertaining activities “involve re-using information that has previously been created or accessed” (Dumais et al., 2003).

2.5.1 Web search vs. search in personal information space

Until very recently there has not been much research that looks specifically at “document retrieval practices in the individual's personal information space on desktop computers” (Ravasio et al., 2004). Although organization is mainly for easy retrieval, the way a document was classified at storage time is not necessarily the same as the way it will be remembered at the time of retrieval (Gonçalves & Jorge, 2004). It has been widely observed that people have difficulty in finding files, emails, and bookmarks on their computers (e.g., Bergman et al., 2003; Dumais et al., 2003). Besides the fast growing amount of information, the limited effort and time people apply and the difficulty they have in organization, and the limitations in current folder systems as described above, the factors leading to the difficulty also include the fact that current
independent applications manage information in their own ways and formats, and have limited
search capabilities (Dumais et al., 2003).

The problem makes information retrieval an important issue in personal information
management. Compared to the powerful search engines in Web searching, the currently available
tools in information retrieval on personal computers are still in a limited stage (Dumais et al.,
2003). Studies found that users are usually more comfortable with web search than searching
their own personal information spaces (Ravasio et al., 2004). Ravasio et al. indicate that a most
important reason is believed to be the simple and easy-to-use interface of search engines where
people do not need to enter the words precisely; and another factor is the different kinds of needs,
purposes, objectives, and evaluation criteria in using web search and on personal computers
(Ravasio et al., 2004). Web searching is mainly for new and unknown information while
searching a personal computer is usually for a particular item the user has seen/saved before.

2.5.2 Information re-access: browse and search

Barreau and Nardi (Barreau & Nardi, 1995; Nardi & Barreau, 1997) have been seen as the
earliest researchers to report findings regarding re-finding on personal computers (Ravasio et al.,
2004). Among other findings, they found that people prefer to browse manually through the
hierarchical directories to get the documents they want and only use searching as a last resort
(Barreau & Nardi, 1995). This “browsing over searching” phenomenon has been observed in
many other studies (e.g., Boardman & Sasse, 2004; Ravasio et al., 2004), and low cognitive and
memory load is an important reason behind it. With browsing, users are able to get familiar with
their PISs (Ravasio et al., 2004) which would enhance the sense of control and make further retrieval possibly easier because of the familiarity.

Boardman and Sasse observed two types of browsing in a cross-tool study: (1) location-based browsing of folders/desktop icons, and (2) sorting/scanning of items, ordered by user-defined metadata such as “name” or system-defined metadata such as “size”. They found that the subjects used browsing-then-sorting in retrieving a file in a folder, and they used search more in email than in files. But still, consistent with other studies, most users were found to use search as a last resort in either collection. A user's quote explains the point: “if you know where you're going, browsing is a lot quicker” (Boardman & Sasse, 2004).

It is also found that users use location-based browsing more for isolated/working information and use search over archived items (Nardi & Barreau, 1997). In a study on how people manage their digital photographs, Rodden and Wood found that, mostly because of the simple browsing feature for photos, participants found it much easier to manage digital photos than non-digital files (Rodden & Wood, 2003).

Teevan et al. observed that most of the search behavior in emails, files, and on the web involved a series of small “orienteeering” steps to find information instead of “jumping directly to the information target using keywords” even when they knew exactly what they were looking for in advance (Teevan et al., 2004).
Overall, low cognitive load that eliminates the need to specify information need or describe the files in advance, the familiarity that builds a sense of where he/she is and to feel in control, and better understanding and trust (Teevan et al., 2004) have been identified to be the reasons for the “browsing over searching” phenomenon.

In the research area of information seeking behavior, it is emphasized that browsing, step-by-step triggers (O'Day & Jeffries, 1993), and mechanisms to support evolving information needs should be supported in a system supporting information gathering behavior (Teevan et al., 2004). Although these points are for information gathering and search in general, they are still applicable in searching PISs.

While these findings suggest that people may prefer location-based finding over logic searching on their computers (Barreau, 1995; Boardman & Sasse, 2004), some other researchers argue that the main reason for the lack of use of search is that the current search tools are not effective and efficient enough (e.g., Blanc-Brude & Scapin, 2007; Fertig et al., 1996; Ravasio et al., 2004). In a study on desktop systems, Ravasio et al. identified several problems in current desktop searching tools, e.g., not easy to handle; “if it's not a hit, it is nothing at all;” and the inability to search by content independent of the file format (Ravasio et al., 2004).

But search is undoubtedly important especially in certain circumstances when users have no idea where a document could be, or they simply get lost, but they have specific characteristics in mind/at hand about the document (Ravasio et al., 2004). As a result, researchers try to solve the problem with improved search tools, for example, enhanced browsing with metadata (Hearst et
al., 2002), searching across formats and tools by unifying index (Dumais et al., 2003), or recording everything and presenting timeline stream visualization (Freeman & Gelernter, 1996).

It is worth noting that the search tools on personal computers have largely improved within the last decades, and the above problems have more or less been addressed in the more powerful search tools. But search cannot solve all the problems. KFTF studies found that people don't realize the existence of a document before it gets expired (Jones et al., 2001), which means that people don't even think about searching for it when they need the document. Searching requires people explicitly doing it to get what they need. Another reason is that, even though a document is found by searching, because there's no context under which the user was when the document was originally saved or created, the user may not understand the document. What the user gets is the single document itself, without other related information provided (e.g., Ravasio et al., 2004). Search cannot support important needs such as reminding, planning, keeping content information, and so on (Bondarenko & Janssen, 2005).

This issue of “why browse” will be discussed further in Chapter 9 with the findings and implications of this study.

2.5.3 Information re-access: other strategies

In investigating Web page revisitation behavior, Jones, Bruce and Dumais found that the participants used other methods more frequently than those explicitly supported by web browsers to revisit web contents. Although the methods are mainly for web page revisitation, some of them also apply to re-access on local computers, e.g., email to self or to others, print out, or
search on Web again. And the functions that they analyzed in choosing the strategies can also apply to local computers, e.g., portability, preservation, reminding, ease of integration, or ease of maintenance (Jones et al., 2001). Some of these strategies and reasons are observed in this study.

Studies also found that people often retrieve files by accessing the software used to create them (Barreau, 1995). Whittaker and Hirschberg suggested that retrieval of documents via their context such as another concurrently worked document should be supported (Whittaker & Hirschberg, 2001). An important factor different from general information management is the idiosyncrasy. For example, requests for “the fat book about computers I skimmed last month,” and “the email about reinforcement learning that I forwarded to Terry last week” could not be satisfied in a general IR system (Adar et al., 1999).

In exploring new retrieval strategies, Gonçalves and Jorge managed to discover common trends in the stories about documents for the design of narrative-based document retrieval interfaces: time, storage, and purpose are the three most common story elements (Gonçalves & Jorge, 2004). Indratmo and Vassileva proposed to present activity traces in order to help users comprehend their document collections better and recall the context of their documents (Indratmo & Vassileva, 2006).

This re-access issue will be revisited and discussed in Chapter 9.
2.6 Context in Information Needs and Seeking Behavior

A frequently encountered concept in PIM research and system design is *context*. Dervin (1996) once declared that in the information needs and seeking (INS) area “there is no term more often used, less often defined and, when defined, defined so variously, as context”. Similarly, there is various usage of this term in PIM and related HCI and IR research.

To clarify the concept, this dissertation separates *context* into two layers: *context of information item*, and *context of information behavior*, with INS research mainly focusing on the latter one. Personal information organization behaviors involve both layers although the former is often the major part. *Context of information item* can refer to “any information external to the contents of the file that may be present while the file is being created or accessed” (Soules & Ganger, 2004), that is, any information related to the interaction. It could include the active applications, other open files, or the current physical location of the user, etc. (Soules & Ganger, 2004). If it is extended to *context of information behavior*, information such as location, user's identity, states of people, groups, computational and physical objects (Greenberg, 2001) need to be considered.

It is widely accepted that information should be understood in its context (Bergman et al., 2003), be it of information item or information behavior. Context in both senses is also important in PIM. Lansdale noted that PIM is affected by personality, preferences, job types, circumstances, tasks, and technology, and “the overt behavior and performance were the result of a trade-off between the constraints of the technology and the underlying psychological processes” (Lansdale, 1988). Studies found that “the context in which a document is acquired or created has much to do with how it is classified, stored, and later retrieved” (Barreau, 1995) and it is often “the way that
humans remember things” (Soules & Ganger, 2004). Most of the influential factors identified by Kwasnik in different information management stages were contextual factors, e.g., situation attributes, disposition, order, value, and cognitive state (Kwasnik, 1991). Barreau observed a similar pattern in electronic files (Barreau, 1995). It has been realized that “cognitive loading is reduced only temporarily through the creation of an information space but the context of the information and its specifics must still be recalled if that meta-level information is not placed in some prominent location or place” (Peters, 2002).

The challenges are what kind of context to save and how to save and use the context with the limitation of available technologies. In practical systems, context of an information item is approximated with information encountered (Baldonado & Winograd, 1997), or documents with similar content, or documents citing or cited by. In the new temporal metaphor systems, it can be the computer state at a certain time and it can be regenerated in retrieval time (e.g. Time-Machine Computing in (Rekimoto, 1999), Rooms metaphor in (Henderson & Card, 1986). (Bergman et al., 2003) proposed to capture external context (link to other items), internal context (annotation), and temporal context (left as it was in last interaction) and make use of these idiosyncratic subjective attributes meaningful to the user alone to make unique PIM instead of general information management (GIM). As for capturing the context of information behavior, we have to wait for technology development, although much effort and progress have been made in related areas (e.g., Kimura, RoamWare).

As a complement to the discussion, a brief introduction of context (of information behavior) research in INS is presented below.
The research on context of information behavior in INS mainly includes objectified and interpretative approaches. The former one is aimed to identify factors that influence information seeking and to build general patterns so as to predict individuals' behavior and needs in certain situations (Talja et al., 1999). Context here refers to objective reality waiting to be identified. It is usually produced in research by naming the entities (actors, structures and attributes) which affect the research object. The interpretative approach, on the other hand, believes that these entities do not exist as such and they are “constituted in researchers' social activity in the same way as the research object” (Talja et al., 1999). There are a growing body of in-between approaches which are constructionist but not explicitly social constructionist and see users’ experiences as the essential phenomena to be explained (e.g. Kuhlthau, 1991; Talja et al., 1999).

In the practical INS studies, context of information behavior includes the factors or variables that seemed to affect individuals' information seeking behavior. For example, socio-economic conditions, work roles, tasks, problem situations, communities and organizations with their structures and cultures, and so on (Talja et al., 1999).

According to (Greenberg, 2001), there are three theories relevant to context in INS: (1) Suchman's situated action in which situations are analogous to context of information behavior and they are fluid and ever-changing; (2) Nardi's activity theory which claims that activity defines context of information behavior and that an activity comprises a subject, an object, and operations. Context of information behavior cannot be inferred simply by enumerating the external set of resources (people, artifacts, settings), for it must also include people's internal and changing states (objects and goals); (3) Fitzpatrick's Locales framework in which a locale (place)
is associated with the site and means for doing work and is the external contributor to context (of information behavior).

A common key point among these three theories is the dynamic characteristic of context of information behavior. It is a “continually evolving and highly situation-dependent construct” (Greenberg, 2001). These characteristics are also observed in folders in this study as described in Chapter 5 and 6.

2.7 Alternative Structures

Since the current hierarchical file folder structure is believed not to support natural ways of organizing and retrieving, alternative structures have been proposed by researchers.

2.7.1 Temporal

The temporal dimension is important especially in personal information space since it involves a lot of autobiographical memory related to information access (Whittaker & Hirschberg, 2001). Rekimoto proposed the concept of Time-Machine Computing (TMC) which is a time-centric approach to organizing information on computers and a user can refer to a document by visiting the past and the future states of computers (Rekimoto, 1999). Based on the psychological research in episodic memory which includes “information such as the location of an event, who was present, and what occurred before, during, and after the event”, Ringel et al. (Ringel et al., 2003) use a mix of personal and public landmark information as memory clues in a timeline-based visualization system of search results by SIS (Stuff I’ve Seen) (Dumais et al., 2003). The time coordinate is also used as an organization and presentation structure of information.
Project **Lifestreams** (Fertig et al., 1996; Freeman & Gelernter, 1996) uses time stream as a main method in presenting all the information recorded in a PIS. It is based on the idea of “save everything” and stores every document created or received by the user. The organization and retrieval are conducted via the stream filters which produce corresponding substreams from the main data stream according to search criteria. These substreams are virtual document organizations (much like semantic file systems) instead of fixed and rigid directories in conventional file systems. And they can continue to collect new matching documents over a period if needed.

The documents in a lifestream are arranged from the past to the future, starting from the tail containing documents from the past. It can contain documents you will need in the future by moving beyond the present and into the future, by which it can serve activity management functions such as reminders, meeting schedule and to-do planning. Some other efforts using time as a main organization mechanism include Plaisant et al.’s LifeLines (Plaisant et al., 1996), Rekimoto's Time-Machine Computing (Rekimoto, 1999), and Microsoft MyLifeBits project (http://research.microsoft.com/en-us/projects/mylifebits/), among others.

### 2.7.2 Spatial

Malone found that people prefer spatial classification over logical one in a paper-based world (Malone, 1983). It has been observed in many other studies (e.g. Kwasnik, 1991), and it is found that people retrieve and order their data spatially even on their computers if they have the opportunity to do so (Ravasio et al., 2004). It is important to note that physical location-based
organization and retrieval is different from “location-based search” (e.g. Barreau & Nardi, 1995) of digital files in the hierarchical file systems - which is called “browsing” in this study - in the sense that the file folder structure on computers is basically symbolic instead of physically spatial. That is, the directory path of a file is not a physical location.

A place where people can use spatial organization on computers is the “desktop” screen where users were found to group the content based on thematic proximity, document type, or other personal criteria, much like a physical desktop (e.g. Ravasio et al., 2004). Several proposed spatial alternative metaphors include the variety of 2D and 3D interfaces (e.g. Data Mountain in (Robertson et al., 1998), or “Rooms” metaphor that allows regenerating context when switching tasks (Henderson & Card, 1986). Visuo-spatial organization is intuitive with low cognitive load. But the effectiveness of these spatial metaphors are still under study (Cockburn & McKenzie, 2001; Cockburn & McKenzie, 2002 since studies found that it does not scale well and it's limited by the number of objects people can perceive (Dumais & Jones, 1985; Jones & Dumais, 1986). There are contradictory findings (e.g., Mander et al., 1992 and Jones & Dumais, 1986) on spatial location (e.g. screen space) as an effective way to organize and search files.

2.7.3 Topical and document attributes, tagging

As described in above 2.3.2, empirical studies identified various elements such as form, purpose, time, topic, and so on (e.g., Kwasnik, 1991). Parallel to these empirical studies, several alternative systems have been proposed to incorporate categorical attributes into the main structure to organize and retrieve information items, such as Haystack, Presto, and Semantic File Systems. Another example is “personalized organization scheme” (Shapiro & Hughes, 1999)
proposed by Shapiro and Hughes in an effort to make personal information organization personalized.

The **Haystack** project (Adar et al., 1999) aims to make a digital IR system a personal bookshelf instead of a library-like system serving all users. It tries to personalize the information retrieval process by gathering and utilizing data about users' interactions with information.

A user's haystack system provides automated data gathering through active observation of user activity, customized information collection organization, and adaptation to individual query needs for the information retrieval process. Haystack scans its corpus to make connections between documents with similar content, and connects documents by tracking the usage history of objects and exploiting metadata links. With the ultimate goal of providing high-quality retrieval, Haystack searching makes use of various search engines and has interfaces to different search engines.

**Semantic File System**: The motivation for Semantic File Systems is to look at data independent from tools and to build a unified data management framework that allows for seamless and uniform access to the data (Gifford et al., 1991). These systems provide several (attribute, value) pairings to files and thus have the ability to classify and search by these attributes.

A key issue in these systems is to obtain useful and meaningful attributes automatically, which would determine the effectiveness of the systems over existing search tools (Soules & Ganger, 2004). Most semantic file systems rely upon user input and content analysis to gather attributes,
which could be extra load for users in both organization and retrieval times. Other possible sources include available file system attributes, user feedback, external context, application assistance, and even user access pattern and multiple copies or versions of a file.

Two challenges for Semantic File System are (1) how to deal with the increase in the number of attributes assigned to each file which would result in too many results in retrieval (Soules & Ganger, 2004); and (2) sharability of document attributes between different individuals (Ravasio et al., 2004).

**Presto:** Similar to Semantic File Systems, the project “Placeless Document” in Xerox Parc and the proposed system Presto is a document attribute-based system with “as many attributes as possible” (Dourish et al., 1999) to solve the many problems in traditional hierarchical file systems. The basic idea is that document attributes (or “properties”) can reflect not only categorizations of documents and document use, but also expressions of desired system activity, such as sharing criteria, replication management, and versioning. Thus the prototype combines document properties as a uniform mechanism for document management and active properties as a way of delivering document services.

Again, how to balance the individuality and sharability is the challenge for the attribute-based system.
A limitation in Presto and also in Lifestreams is that users have to handle and remember the arbitrary and isolated piece of properties and there is not a way provided to make them coherent (Gonçalves & Jorge, 2004).

A few recent systems try to use semantic tags to label documents, and provide better support for multiple classification (e.g., Arriaga, 2002; Dourish et al., 2000; Quan et al., 2003). For example, Dourish et al. proposed Placeless Document system based on document properties which avoids many problems of traditional hierarchical organization systems. It augments the system with active properties which enable the provision of document–based services. It refines document properties into uniform properties and user-specific properties and provides “collection” mechanism instead of traditional “folder” concept (Dourish et al., 2000). In (Quan et al., 2003), Quan et al. proposed a user interface with Web browsing in which users can do multiple categorization with attribute-value pairs. In a prototype called Newdocms, Arriaga removes the “file name” concept, and replaces it with attribute-value pairs in describing files. Similar to (Dourish et al., 2000), “collection” is used for sharing metadata within hierarchical relationships. Oleksik et al. created a tagging system that works with Windows desktop metaphor and can be used as a layer on top of hierarchical file system structure (Oleksik et al., 2009).

While most of these systems rely on manual tagging, there are some studies showing that people are not likely to spend time to assign metadata (Marsden & Cairns, 2003; Rodden & Wood, 2003).
2.7.4 Task- or project-based

Support for a higher level organization based on task or project has been observed to be important in computer systems (e.g., Bondarenko & Janssen, 2005; Ducheneaut & Bellotti, 2003; Kaptelinin, 2003; Ravasio et al., 2004), since tasks or projects are facilities driving and organizing the daily work. Kaptelinin et al. identified problems related to temporary file organization, personal activity planning, and recreating context of a file (Kaptelinin et al., 1999), and addressed these drawbacks with a task-centered design of the UMEA system (Kaptelinin, 2003) which manages different types of information based on project. Ducheneaut and Bellotti designed Taskmaster and recast email as task management facility using embedded task-centric resources (Bellotti et al., 2003; Ducheneaut & Bellotti, 2003). Challenges for these approaches are to deal with all information types and reducing the overhead of management. For the proposed communication-based tools such as ContactMap and TaskMaster, the problem is that the relationship between project and information may not be one-to-one.

UMEA (User-Monitoring Environment for Activities) is a project-based organization system across different forms of information (Kaptelinin, 2003). It organizes different forms of resources into project-related pools, and by monitoring the user's activities adds new resources to active project pools automatically. Its design is informed by activity theory.

The interface of UMEA includes project panel, resource panel, a complete list of projects, and a history area displaying PIM tools and the interaction history of the active project. The challenge is to deal with the inadvertently added resources coming from the situations when the user conducts unrelated information activities such as visiting a news source on the web.
Recognizing that folders “often include information relating to a task and its outcome” (Jones, 2007, p. 50), Jones and his colleagues developed a tool called Personal Project Planner which provides an “integrative organization of information” in planning a project and managing its tasks (p. 51). Personal Project Planner provides an alternative view of the user’s file folder hierarchy and links related information across tools (e.g., email system and file system) with flexible ordering and context annotation mechanisms.

A similar but more open structured view of a file system is proposed by Voida and Greenberg in a prototype called WikiFolders, which is “a hybrid system for annotating file folders that draws upon the strengths of both the hierarchical file system and wikis” (Voida & Greenberg, 2009).
CHAPTER 3: METHOD

3.1 Research Design

The complexity of personal information organization and re-access behavior in computer folders, as described in Chapter 1 and further demonstrated in Chapter 2, and the open ended and exploratory research questions, led to the qualitative research method that this study adopted.

Three considerations guided the research design:

1) Data variation and richness. Because of the complexity of the phenomena under study, this study tried to obtain varied and rich data from multiple perspectives. Specifically, the subjects include two groups of participants who are conducting different types of information activities. The multiple data sources and data collection methods include in-depth interview with guided tours, emails reporting example cases of re-access difficulties, disk scan of several file folders and screenshots of some emails folders, and re-access task observations, as illustrated in Figure 3.1.

2) Evolving nature. Since folder structure and content are evolving and can change over time under different situations, this study conducted two rounds of interviews in trying to capture the changing nature and dynamics.

3) Being specific. The study tried to get specific and detailed data instead of general descriptions in order to provide in-depth data. For example, as a part of the effort of understanding how the participants organize information in folders, this study tried to understand why or when they created a particular folder, because the rationale behind it and its content lies at the moment when the folder was created or re-organized.
With the exploratory nature of this study and the open “how” and “why” research questions, the study adopted a multiple-case study with 12 participants as the instrumental cases (Pickard, 2007, P. 86) while the focus is the participants’ organization and re-access behaviors in computer folders. The purpose is to understand more about the phenomena of people’s folder use in organization and re-access, “maximize discovery of the heterogeneous patterns and problems” and maximize “the researcher’s ability to identify emerging themes” (Erlandson et al., 1993, p. 82), instead of generalizing the findings to a larger population in the statistical sense.

Specifically, the 12 participants include 6 PhD students and 6 administrative staff in an academic environment. Two rounds of in-depth interviews were conducted with each participant in front of his/her computer at three month intervals. During each interview, the participants were asked to give the investigator a tour of their primary information organization systems which include mostly file folder systems and some email folder systems (for several administrative participants).
The participants were asked to talk about the information items and folders, their specific behavior in creating folders and organizing particular folders and information items, as well as difficulties they might have in re-accessing them. Then they were asked to find several information items on their computers while the investigator observed the procedure.

At the end of each interview, a disk scan command was run under the top level folder and other 2 or 3 other selected folders to capture the file folder structures. Based on the first interview disk scan data, at the time of the second interview several files or emails were randomly selected on each participant’s computer to ask them to re-find them. From the start of the study, the participants were asked to report via email whenever they experienced difficulties in re-accessing information on their computers. The reported examples were discussed during the next interviews. The interviews were audio recorded and then transcribed and coded. The emails were coded separately in the analysis. Simple counting was conducted with the disk scan results.

The study mainly looked at each participant’s one computer, although they may have more than one machines. The folders investigated in this study are mainly files folders. But for several administrative participants whose emails are a primary part of their daily information management activities, the emails in the corresponding email systems’ folders were also investigated. Only the files saved by the participants themselves were considered in this study.

Figure 3.2 illustrates the data collection and analysis workflow. The details of each step will be described in the following sections.
3.2 Data Collection

3.2.1 Participants

The participants include six graduate students and six administrative office staff members in an academic institution as the subjects. In order to get informative subjects for this particular study, recruiting emails (see appendix A) were sent to about 40 administrative staff and doctoral students in the academic institution. The email asked their willingness to participate in the study and solicited examples of information re-access difficulties. The recruited subjects share two characteristics based on their responses:

1) Using a fairly large number of files for daily life or for work; and
2) Having experience of re-access difficulty.
Among the six administrative participants, the time they have been working in the institution ranges from a couple of months to more than 20 years. The doctoral student participants include ones at early stage in the program as well as at dissertation finishing stage. Table 3.1 includes the operating systems they were using and the length of time they had been in this institution at the time of the first interview.

Table 3.1 Participants Overview

<table>
<thead>
<tr>
<th></th>
<th>Operating System</th>
<th>Time in the Institute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adm. (N=6)</td>
<td>Windows XP</td>
<td>3 months – 29 years</td>
</tr>
<tr>
<td>PhD. (N=6)</td>
<td>3 Mac OS; 2 Windows XP; 1 Unix with XFS</td>
<td>1 – 6 years</td>
</tr>
</tbody>
</table>

According to (Bondarenko & Janssen, 2005), as illustrated in Figure 3.3, PhD students fit in the “research” extreme (with unstructured tasks) and the administrative participants, although they are not “HR managers,” should fall close to the other end of “administrative” (with structured tasks). It is important to note that these two groups are in the activity spectrum and are not exclusively doing “research” and “administrative” activities. Rather, members of each group do both kinds, but in varying degrees, giving an opportunity for a richer understanding of a continuum of multiple practices. Since “activity type has a large impact on document management” (Bondarenko & Janssen, 2005), these two particular groups of information workers provided richer data and variety which are important for identifying issues in this exploratory study.
Similar to the research design, using these two groups of participants is more for helping identify more issues and providing rich and varied data, rather than for comparison between the two groups. As verified in this study, an issue may exist in both groups but was more prominent in one group to be identified (e.g. similar organization strategy for a fast growing project; the work-in-progress folders issue, as described in Chapter 6).

Figure 3.3 Bondarenko & Janssen Activity Spectrum (2005)

![Diagram showing the spectrum of research, administrative, structured tasks, and unstructured tasks]

3.2.2 Emails

In the recruiting email and the Consent Form, subjects were informed that they would be asked to report example cases via email during the study when they experienced difficulties in re-accessing information items on their computers (see Appendix A for the recruiting email, Appendix B for the reminding emails, and Appendix C for the Consent Form). Reminding emails were sent to them every two weeks if the participants did not submit example cases.
Since people may not remember the details after a long time, it is difficult to collect the data during interview about what specific problems the participants had in re-accessing information and how they re-accessed specific information items, especially given the short recall time during interview. Collecting example cases by emails allows them to be able to submit a case whenever it happens, or recall it shortly afterward. It can also roughly show the occurrence frequency of re-access difficulties to some extent. The three months period between two interviews allowed for a sufficient number and variety of cases reported.

3.2.3 Interviews

Interview with a “guided tour” with the participant’s computer can be a useful method to collect specific data connecting system (the folder system) and people (the participants), which are the two primary components in this study. By looking at what the participants had in their folders, and what they described and/or complained about, many tacit complex factors and relationships behind the files and folders can be captured.

During each interview, the participants first gave the investigator a guided tour of their primary information organization systems. Since the focus of this study is the hierarchical folders as a primary information organization system, the guided tours were on all the PhD students’ and the administrative participants’ file folders, as well as several administrative participants’ email folders.

Although a set of broad questions (see Appendix D) were used to help guide the scope of the conversations, the actual interviews were directed by what was observed on the participants’
computers and what the participants talked about. The participants were asked to explain and describe their particular folder creation and organization of information items, and the difficulties they might have in re-accessing them. The investigator tried to ask concrete and generative questions. For example, “what is this folder/file,” “how/why did you create the folder,” “what are the things you are currently working on,” “why did you put the file here,” “how did you re-access things under this folder,” etc.

The second round of interviews was conducted about three months after the first one. Compared with the first interview, the second one followed a similar procedure to the first interview, although it included more re-access task observations and focused more on the new and changed parts of the folder structures. Difficult example cases reported during the two interviews were discussed in-depth during the second interview.

Each interview lasted from 1 to 2 hours. The interviews were audio recorded and transcribed. The total audio length of 24 interviews is just over 34 hours.

As described in section 3.1, the two rounds of interviews allowed evolving issues to be captured and explained to complement the data collected in the first interview. A second chance to look at the same issues for both investigator and subjects provided an opportunity to clarify, adjust, and complement information collected in the first interviews. It is especially valuable in studies of personal information organization and re-access behaviors because the behaviors are often conducted without much explicit thought. The three months allowed time for interviewees to pay attention to and report information re-access difficulty experiences, and the second round
interview provided a chance to discuss in-depth and clarify the interviewer’s understanding. Files were able to be better selected for the re-access task observations on the second interviews with the information obtained in the first interview and its disk scan result.

3.2.4 Disk scans
During each interview, three or four folders were selected randomly by the investigator to run the disk scan command supported by computer systems. On Windows machines it is the Microsoft DOS command “dir /s /ta /od > fileX” while on Mac and Linux machines it is the Unix command “ls –R –u –lt > fileX” (the “X” in “fileX” in the commands is the sequential number used to differentiate multiple files). The result of the command lists the folder structure and the file/folder names, with last access time. The folders generally include the first level directories (without “/s” on DOS or “-R” on Mac) and the directories for the current working project, a completed project, and for miscellaneous files. The result files were collected after they were produced.

Because of the privacy issue, this study only runs the command under three or four folders to make it less intrusive. The folders that the participants did not feel comfortable with running the command, e.g. personal things folders, were excluded from scanning.

3.2.5 Re-access task observations
During interviews, several files and sometimes emails were randomly selected to ask the participants to find them while the investigator observed the procedure. The selection is based on the interview and the results of the disk scan in the first interview, occasionally also based on the
investigator’s knowledge about the participant’s related information activity. The selected target information items were usually from a currently working project, or completed, archived folder, or miscellaneous items. The re-access procedures in folders were observed by the investigator and spoken aloud by the participants for records.

Because of the lack of familiarity of the investigator toward participants’ information systems during the first round of interviews, re-access task observations were mainly conducted during the second interviews.

Asking participants to re-access a particular file or email during interviews provides an opportunity for the investigator to observe participants’ re-access behavior. It may also complement the variety and number of problem examples obtained from emails, and the participants’ description during interviews about their re-access activities.

3.2.6 Screenshots and others

To supplement the above primary data sources, some screenshots of computer Desktops and email folder structures were taken. Brief field notes were taken for each interview. The hand drawn graphs or illustrations participants did were collected.

3.3 Data Analysis

After the interviews were transcribed, the data were coded and then synthesized. Content analysis was conducted on the reporting emails and disk scan results.
3.3.1 Coding

The interview transcripts and emails were coded in QSR NVivo 8. The coding step generally includes two stages:

1) Stage 1: “open coding” (Corbin & Strauss, 2007) which produced a long flat code list;

2) Stage 2: compared to the research questions, adjusted and moved the result of stage 1 to another tree code list. It is similar to the “axial coding” (Corbin & Strauss, 2007) in that it tried to bring issues together based on the scopes decided by the research questions.

Please see Appendix E and F for the two code lists in NVivo.

The stage 1 started with microcoding (Corbin & Strauss, 2007) strategy, a detailed type of open coding usually used at the beginning of data analysis. During this stage, the investigator tried to pay close attention to details, keep an open mind, and be sensitive and reflective for potential issues. Because it is hard to see all possible “concepts” (Corbin & Strauss, 2007) in the data at the early stage, the “data reduction” time was postponed intentionally. Thus the code list, even in the final tree code list of stage 2, included mainly “what it is talking about or related to,” mixed with some concept codes identified by that time for “what it means.” The purpose was to allow the investigator more time and opportunities in the later stages to be submerged in the data and see things that could not be seen earlier. It is also helpful in that there is less need to go back to large, original transcriptions to find missing data, although I did need to check them sometimes.
3.3.2 Content analysis
The reporting emails were analyzed based on the structure the recruiting and reminding emails provided. Factors were identified and coded within each part. The details are described in Chapter 8.

3.3.3 Synthesis and writing
In order to get a better sense of what the data are really indicating, the coded data were grouped and copied from NVivo into text files, transferring the data from digital to paper form for convenience of reading, annotating, and thinking, in another word, synthesis. Another reason for this instead of using iterative coding in NVivo is that, instead of scattered in different nodes in NVivo, all the related issues were grouped together as important context information for iteration of synthesis.

Data reduction was conducted at this stage by composing a “potential issues list” which includes all the identified issues that were deemed of fundamental importance. Several writing projects were conducted and each of them covers one or several issues from the list.

3.4 Considerations and Reflections
3.4.1 Data collection
3.4.1.1 Participants
When thinking back about it after the data collection was completed, the fact that the participants are in the same academic environment as the investigator is actually beneficial, especially because the main topic of the study is files or emails which are related to the institution and
people within it. These contextual cues are difficult to obtain if the investigator is an outsider, which could cause understanding problems for the investigator. Much subtle contextual information (e.g. who is leaving, who is the supervisor, who is the Dean, etc.) actually influenced particular information organization behavior. The contextual information was also used to help the investigator compose some re-access task observations.

3.4.1.2 Emails

The structure given in the soliciting example email turned out to be especially useful in data analysis (as illustrated in Chapter 8), although data analysis stage showed that it would be better if more specific information could be obtained, e.g., on what they remembered about the target information items. It is a challenge, though, to collect more specific data and at the same time keep the participants’ work minimal.

3.4.1.3 Interviews

Because of the exploratory nature of this study and the adoption of interview design that is close to unstructured interviews, the predefined guided questions used to set the scope are more like an aide-mémoire or agenda, which is “a broad guide to topic issues that might be covered in the interview, rather than the actual questions to be asked” (Wildemuth, 2009, p. 223). This imposed challenges to the investigator in conducting successful interviews, since the conversation during an interview largely depended on the specific situation on site, on what the participant said and what he/she had in the folders. One of the major challenges is to identify issues quickly (Wildemuth, 2009, p. 225) from the observation and conversation. Since how deep the exploration can reach depended on what questions were asked, it was important for the
The general criterion of the investigator was to know as much as possible about the behavior of interest, even though some discussion seemed not directly relevant to the research questions, which could end up raising interesting issues.

Another challenge related to this first one is how the investigator did not get in the way of a participant’s free flow of thinking and information and at the same time explored for more details about the issues that emerged during the interview. As (Wildemuth, 2009, p. 226) described in referencing (Patton, 2002), “when a new topic emerges in the discussion, it is difficult for the researcher to know whether to follow it and risk losing continuity, or to stay on the major theme and risk missing additional useful information.”

3.4.1.4 Disk scans

Using the systems’ built in commands to get the files and folders structure data is a simple, unintrusive, zero cost, and practical method for the participants, especially since the administrative participants were using computers in offices and it was not practical to get the permission to install external recording software. Since most PIM studies, including this one, tend to have privacy concerns, using an available tool can reduce the sensitive level and make the study more acceptable for the participants.
Also because of privacy concerns, the study only scanned three or four folders instead of all the folders on a participant’s computer. If all the folders could be scanned, more quantitative studies would be possible to provide better data analysis.

The study was originally designed to capture the access date for each file within the several folders so that data such as access frequency and distribution can be obtained. But when running the commands, it was found that the access time on computers can be affected by many factors (e.g., backup operation) and further reveals the complexity of the concept of “access” from the system’s perspective. A file can be accessed by the user explicitly or implicitly by software; or can be accessed by a supporting system automatically to fulfill some tasks (e.g., check for virus). It is also not clear if it is an access when a user changed the file name without opening the file, or opened it and found it’s the wrong one and closed it.

At the time when this study was conducted, there was no software available for this purpose. A group of spyware and recording tools were investigated, but none of them can provide all the data this study needed. The biggest problem was it’s almost impossible to have participants installing these tools on their computers.

If such a tool were available, and if it can handle reasonably the above definition of “access” issue, and it is safe enough to be allowed to install on all the participants’ computers, the collected data would make more quantitative analysis possible on re-access behavior.
3.4.1.5 Re-access task observations

One issue that emerged in conducting the re-access task observations was how to describe target information items to the participants before they knew what items the investigator was talking about. It turned out that the “tags” themselves in file and folder names were not enough to let participants know what the files are. This issue will be further discussed in Chapters 5 and 6.

3.4.2 Data analysis

3.4.2.1 Coding and synthesis

One issue that appeared in the data analysis stage is how to treat and understand the seemingly inconsistent descriptions. For example, a participant said “I’m afraid to get rid of things” even though half a minute ago she deleted a file in front of me without hesitation. The study sees such inconsistency as a great opportunity to explore for the underlying reason (“the previous file is not so useful for me” in this case). The implication for data analysis is to recognize that at many times participants tend to give a general answer without mentioning the possible exceptions and refining it to an accurate level. In the above case, the “things” the participant was “afraid to get rid of” did not include the files she just deleted because as she said, “the previous one is not so useful for me.” A refined understanding was obtained by identifying the “exceptions” underlying inconsistency.

3.4.2.2 The use of software in data analysis and its implications

NVivo organizes data sources with a hierarchical folder structure, which includes internal, external, Memos. Internal and Memos are used in this study. Under Internal, there are subfolder
“adm staff” and “phd students”, and each of them includes four subfolders: disk scan, emails, interview, and screenshot.

Coding can be done across these types of data sources. The NVivo Help Documentation explains that, “by creating nodes and coding at them, you can catalogue your ideas and gather material by topic.” It supports two types of coding: “free node” and “tree node.” “A free node is a 'stand-alone' node that has no clear logical connection with other nodes—it does not easily fit into a hierarchical structure.” In this study, the free node mechanism is especially useful at the first round coding when the investigator “had not yet developed a node structure,” had no idea of what the structure is like, and even did not care about the structure at that time. At the second stage of coding in building a tree node structure, NVivo provides a fairly easy mechanism to allow free node coding moved to the tree node structure, although it’s not as easy as drag-n-drop.

Different from typical representation of coding schema, the node names used in this study are usually long descriptive tags. For example, “file-folder names & files next to it – as tags for search and understanding” (because NVivo does not allow “/” in node names, “file-folder” in this example means “file/folder”). More examples are displayed in Appendix F. The naming of the nodes tries to list all the important issues identified in the corresponding data pieces. When new data pieces are coded with the node, more tags may be added to the node name for the newly identified issues. During or after the open coding, node names may be modified so that related targets were grouped together, as partially illustrated in Appendix G. For the same reason of postponing data reduction as described above, many nodes on the same issue with small differences were not merged.
The above descriptions raised two interesting issues. The first is the similarity between the coding process in NVivo and creating a hierarchical folder system, filing, and assigning tags to information items. It is clearly noted in the NVivo Help Documentation, “Tree nodes are organized in a hierarchical structure – moving from a general category at the top (the parent node) to more specific categories (child nodes). You can use them to organize nodes for easy access, like a library catalogue.” This similarity interestingly connected the data analysis procedure of the study with the target of this study – information organization in the folder system. Actually the similarity in a larger context between qualitative data analysis method and information organization method, specifically, faceted classification, has been pointed out by Start (1998). It is profoundly suggestive in thinking about whether people need organization and how people do organization.

As the second issue, the coding procedure in this study indicated that building folder structure and assigning tags might be suitable for different cognitive stages. Building a structure with folders was easy only when a particular “view” was available, but the “view” did not emerge until all the data analysis and synthesis were completed, while tagging was the thing that the investigator wanted to do at the very start and did not have much difficulty with. Hierarchical folders are to structure tags, when the person has a fairly stable understanding or view of the information items, which at many times is decided by a particular situation and context. This can explain why it can be a challenge for a new employee to build up a folder structure when he/she has not much understanding of all the possible information items, as observed in this study.
This observation was supported by another experience in the coding process. After the first participant’s data was done with the open coding, the investigator tried to create a tree structure with the available free nodes. But it was quickly found that the tree structure was over-fitting to that particular participant’s data and not flexible and open enough to represent the new data from the other participants.

In synthesis and writing stage, the searching function in NVivo was very useful for collecting nodes whose name included a particular term (tag). It helped checking for coding consistency at the coding stage and grouping related codes in the synthesis and writing stage.

3.5 Evaluation Issues

Because of this study’s qualitative methodology, the dissertation describes the trustworthiness based on the four criteria proposed by Lincoln & Guba (1985), which include credibility, transferability, dependability, and confirmability.

In terms of credibility, the triangulation provided by the multiple data sources and multiple data collection methods in this study strengthened the validity of the study findings. Data triangulation involved integration of data from multiple sources – in this study, they include the Ph.D. students, the administrative staff, and their computers. Method triangulation involved integration of data collected through multiple methods, which include observation, interviews, disk scan, reporting emails, and re-access task observations in this study.
As described in the above research design, the study was not aimed to generalize the findings to a large population, similar to case studies and many qualitative studies. Other studies are needed to test the generalizability of the findings in a different context and user group. Instead of using generalizability, some researchers propose to use transferability as an evaluation criterion which refers to “the extent to which the researcher’s working hypothesis can be applied to another context” (Wildemuth, 2009, p. 313). The transferability of this study was made possible by the detailed documentation of the data collection and handling process, which “makes it easier for future researchers to judge the transferability of the criteria to other user populations or other situational contexts” (Wildemuth, 2009, p. 314). For example, it is reasonable to apply this study’s findings to some other information workers in similar situations.

For dependability and confirmability, each finding of this study is described with the corresponding raw data (quotes or example cases), thereby confirming the research findings and grounding them in the evidence. The findings can be traced back to the raw data.

### 3.6 Lessons and Limitations

The mistakes and lessons gained from this study include:

1) The investigator forgot to turn on the digital recorder after stopping it during a participant’s phone call, which resulted in loss of valuable data; the investigator ran the disk scan command at the second interview under a different folder by mistake than the first interview, which caused an empty value in some data in Chapter 8.

2) More analysis should be done after the first interview and before the second one, in order to have better knowledge of what content the participant had in folders.
The Limitations of this study include:

1) Email folders were not systematically collected. It is not the focus of the study, and there’s a trade-off because of limited interview time.

2) More participants and more time would provide richer data.

3) Many other related issues emerged without enough data to analyze. They will be the focus of future work.
CHAPTER 4: KEEP AND DELETE

4.1 Introduction

Current computer systems offer clear, simple binary choices in keeping and deleting information: when receiving an information item from somewhere, people can either keep it or discard it. Later on they have the similar binary options of continuing to keep it or of deleting it. People are forced to make the decision in terms of whether and how to keep the information item. As described in Section 2.2, the decision of whether/how to keep can be “fundamentally difficult” in both paper and digital worlds (Jones, 2004). The falling costs and rising storage capacity encourage keeping more information that may be useful later, but too much kept information could compete for attention that should be spent more on important items, and make re-finding information more difficult. As this study observed, the decision to keep or not is often a judgment call at a given moment, and the participants acknowledged that they could make mistakes.

After outlining several special keeping and deleting behaviors observed in this study, the following sections will describe how the participants deal with this dilemma on their computers, which implies that a facility is needed for information items that fall between the categories of keep and not keep.

4.2 Keep and Delete Behaviors

Both groups of participants had their information kept online, in paper form, and on their computers. And their keeping behavior on their computers was found to be affected by the other
two parts, especially in PhD participants group. For example, most of the PhD student participants sometimes printed out an online article directly, without even saving a copy on their computers or only kept the web addresses or reference list file online (e.g., del.ici.ous, Refworks or course system such as Moodle) or on computers. For example, several PhD students said that:

- “I think for most of them I actually took the list and just found them and print them out, for the actual exam, because it’s easier to have all the papers. So most of them I don’t think I actually kept the electronic formats.” [the reading list saved in a file on the computer]
- “I don’t save readings electronically. I print everything out.”
- “I was looking for something very specific in articles. ... Like they might have only one paragraph that was important for me. (so you just read online and take notes?) yeah. (where would you put the notes?) I have a folder called reading notes, in ‘dissertation’.”

The PC, online, and paper worlds became closely connected in their keeping strategies (see Figure 4.1). With the close connections and interactions between online information and information on computers, as one participant said, online websites become “the bigger network space”.

Figure 4.1 Online–PC–Paper Triple
When talking about what they do and do not keep, the target items that the PhD participants talked more about were web downloaded files, while the administrative participants paid more attention to email messages because email was a major part of their daily work, as they declared. Compared to the PhD students, who had their personal, school work and research information on their personal computers, the administrative participants using office computers for their work can have special keeping purposes, e.g., for auditor, for record, or for reference at next year’s or next time’s use.

Both groups had the similar not-keeping criterion of “one time thing,” or things “I don’t think I’d need to come back to,” or “inconsequential things.”

Deciding to keep or not takes effort. This study observed that in a fast growing project, an administrative participant adopted the “keep-everything” strategy and had an email folder including 822 emails accumulated within a little over one month:

“it’s just a lot of correspondence and I can probably delete some of it but it was just one of those things that when the project got started, it would be very complex and it was going to be a very fast project that had to get done right and I just felt like, just save it all and that way I won’t have to worry about whether or not I lost something or misplaced some piece of correspondence that I needed to.”
But the “keep-everything” strategy was not a favored strategy, and when she experienced a re-access difficulty with that big folder, she believed that the keep everything strategy partly caused the problem:

“The fact that I am keeping every e-mail related to this project, instead of keeping selective e-mails, is part of the problem. Although it’s good to have a record of all correspondence, it also creates a much larger volume of e-mails to sift through when looking for one in particular.”

This temporary “keep everything” strategy seems not unique. During a casual meeting after the two interviews, a Ph.D. student participant noted that he kept every file for a fast-growing project without deleting or organizing.

4.3 Two Types of In-between Keeping

One issue that emerged from all the participants’ descriptions is that they had a clear sense of their own “main folders” or “home directory” and tried to separate them from system’s or particular software’s files. These “main folders” or “home directory” usually were under a directory such as “My Documents,” a hard drive, or a folder with the participant’s name or initials. Files or folders on the Desktop were not considered a part of their “main folders,” and several participants in both groups said “I’m not sure if I want to keep it” or similar thing, even though the files were literally saved at the Desktop folder on the computer. By saying “keep” the participants always meant “keep in main folders”.
Within this context, this study identified two special types of keeping between keeping and not keeping, in addition to the important and useful items that the participants want to keep in their main folders. The first type refers to the items that the participants haven’t decided whether or/and how to keep, which are called “possibly useful” items in this study. The other type refers to the ones that the participants do not intend to put into their main folders from the very start and only leave them on desktop or root directory for a while (called “to be deleted” in this study).

Most of the “possibly useful” and “to be deleted” types of files were found being “dumped” either on the desktop or under the user’s home directory (e.g. “My Documents” or “D:” drive). These locations were usually connected to the default downloading or email attachment location. But it is important to note that they did not always leave downloaded files there. As the participants said, for files they were clear about what they were for, they would save them to “main folders” right away. It shows that they tried to keep the files at these locations outside of their main folders in current systems at least temporarily.

It should be noted that in the collected data both “possibly useful” and “to be deleted” files were observed more frequently on the PhD students participants’ file systems than on the administrative participants’ systems. Possible reasons for this can be: 1) the administrative participants’ work-related files can be mostly useful; 2) With the large enough file storage space they do not have the need to differentiate these two types of files out from the main folders; 3) the email system and the associated quota issue forced the administrative participants to file messages to a folder, which changed the issue to be a “delete or not” problem as discussed in Section 4.4; 4) Ph.D. student participants’ activities might be different from that of
administrative participants in that there exist more vague needs of information items for future possible purposes.

4.3.1 “Possibly useful” items

The following examples illustrate the participants’ descriptions of files that were categorized as “possibly useful”:

- “Papers are very good unless you go and read the section about it that says oh, this is .... Kind of figure out how to organize that stuff.”
- “...things I downloaded, for some reason or another I haven’t got around to decide: a, if I need it, b, where it should be filed if I (need it).”
- “These were here because at some point I want to get back to these to make sure there’s nothing important. (you don’t use them?) no, I might just delete most of them.”
- “…This I should figure out what it is. This (another one) I should figure out what it is...”
- “Usually when I put something on my desktop, it’s just because I want to take care of it quickly without thinking about where to put it.”
- “a lot of times the things on the desktop, I actually don’t remember what they are, and I have to open them. ...I have no idea what that is.”
- “(...the many files under My Documents?) I don’t know if I want to file them or I haven’t filed them yet. (Haven’t used them for a while.) No. So it could be cleaned up.”

In a word, “I just wanted to look at it” does not necessarily mean that “I want to keep it.”

These “possibly useful” files that come from email attachment or web download usually went to the default download place, which could cause re-access difficulty. For example, one participant
described his problems in trying to re-find an article he knew he had downloaded. The article had been downloaded to his desktop as the default place. Since the participant does not use the desktop (he was a Linux user), he went through a few searches before he finally found it.

As a solution, a PhD student created a separate folder “webdownload” on the desktop for downloaded files, instead of leaving them mixed with other items on desktop:

“...when we didn’t have any folder, when we downloaded something, it’s on the desktop, and after some time it’s very messy. So I told xx to create it and set it as default folder.

Every download and attachment will go to here automatically.”

Similar to this, another participant packed these files up into a folder after a little deleting and filing at a cleanup time:

“...I just created this folder, that’s called ‘desktop articles’ that’s everything that was on the desktop.”

An administrative participant also created a folder for email attachments:

“if I didn’t know if I’ve saved, I just put in my attachment folder, so in later, I can go back and clean this out and put in various other folders.”

Considering the amount of effort participants have to spend to clean up these files individually, it is not surprising that these “possibly useful” files are a major part of procrastination in organizing one’s personal information. The above pack-up solution implies the need for a special kind of container for the files that are “possibly useful” but may only be checked rarely. On the
Macintosh, the automatic download place utility folder plays a similar role in terms of a different place, but it is designed for system use, without differentiating between useful, possibly useful, and useless items for the user.

4.3.2 “To be deleted” items

The other type of special keeping is the “to be deleted” item. For example:

- “…a lot of things that ended up on the desktop are really temporary, like I’m just looking at it for the time being. For instance, I have the current xxx conference schedule. I don’t need to save that. So at some point, I just put it in the trash.”
- “…these are screenshots. …because I need to show students how to do screenshots on Mac. Actually I can clean them up.”
- “…this actually was because I couldn’t get it to print yesterday. I’m gonna get rid of it actually.”

Several other “to be deleted” files that the participants did not want to keep in their main folders are:

- “If it’s something that I’m going to forget where it is anyway, I won’t bother to download it.”
- “I have the reading list so if there’s anything I need to find, I can find it again, …it’s easy enough to find electronic copies.”
- A conference schedule that “usually goes to the desktop and I usually throw it out once the conference is over with.”
• Files that the participant would rather print out, or use the “online bigger network space” instead of to keep on computer.

These files were mainly observed on the Desktops, mixed with other useful, important, and “possibly useful” ones. Similar to the “packing up” method for “possibly useful” stuff, a participant created a “temp” folder on the desktop to separate and pack up these “to be deleted” items:

“Usually after some time, if there are too many, I will delete it. ...I know everything here is safe to delete.”

Again, a separate container for these files will make cleanup or organization much easier.

4.4 Delete

4.4.1 The need to delete

If we look at keeping as a decision made at the stage of receiving or retrieving an item, then “delete” is a decision of “not keeping” at a later stage, applied to files that have been already saved in main folders, when their usefulness has expired or their uselessness has become clear.

Cleanup, which is “discarding items that I no longer need” was low priority for the participants and usually was triggered by space problems. For example, since the administrative participants’ email system was hosted by the university, one of them explained that:

• “...motivation for cleaning has more to do with space than anything else....I can’t think of a time that I purposefully weeded specific folders because I haven’t felt the need to because I have plenty of room to keep it all, so why not?”
• “To me that’s low priority. I’m not gonna mess with that unless I have spare time. ...What I do is like today is slow day, I need to clean up those files because I got messages (for space quota) they want me to clean it up. ...then I delete it.”

• “Probably these I need to clean up and delete. But I haven’t had time and I don’t care.”

In contrast to that, they rarely had such concern for the files on the computers because the space on the computers was large enough for their work.

For the PhD students, although they were using computers with big hard drives, space could still be a concern to motivate cleaning. As a PhD student said: “19 gigs is still a lot of space, but when it starts getting down to like 10 gigs or lower, then I’ll start to worry and go clean.” With more space, they saved more. Several of the students had videos, audios, or games saved on their computers, as well as big collections of photos.

4.4.2 Difficulty in deleting and the solutions

Related to the “pack rat” inclination, several participants were afraid of deleting or reluctant to delete. For example, a PhD student said that:

“In the moment I thought it would be a good idea to keep it here. And then once things are there, I’m afraid to get rid of them. Whenever I’ve decided to save something, I’m afraid to get rid of it. I don’t know, I might need it.”

Similarly, an administrative participant kept nearly all emails “even for one time use, I archive them. Only delete the spamming ones.”
The decision to delete or not is another judgment call at a given moment and participants realize they “may (have) made mistakes to delete something that I don’t mean to do”:

“that has happened. I believe it has happened. …I wish I kept that, well but I didn’t, and I just move on and figure out some other way to recall what it was …or re-create or whatever’s the specific need at the time.”

An administrative participant talked about the experience of trying to locate a deleted file, and finally “chose to redo it”. Another administrative participant reported having the problem of looking for a sent email after she deleted the sent folder because of quota problem and then had a two week vacation. Several participants keep trash for a while as a way of dealing with this problem:

- “Some system, within 24 hours, it automatically dumps out the trash. That doesn’t work well for me. I need a longer time to realize that an error was not created.”
- “…I rarely empty my trash. So a lot of times I went back to trash to see if I throw away something that I shouldn’t have. …not too much, but I have definitely done that.”

Several administrative participants had personal experience (and knew other colleagues with the same experience) of asking the help desk in the department to recover a file from the backup tape.

This implies that the trashing mechanism on a computer may need to have different levels. Files that were once useful, filed to the main folder and then get deleted are different from other useless or even spamming items, and can be packed and compressed in a way that people can
still recover a file from it if needed, similar to the administrative participants’ backup tape mechanism.

4.5 Conclusion and Implications

Resonating and extending Jones’ finding about the difficulty people have in deciding to keep or not, this study observed the participants’ various compensating strategies and identified finer categories for keeping and deleting (see Figure 4.2). There are many files whose status seems somehow to fall between the binary decisions of to keep or not to keep. Allowing for a status between keeping and not keeping and accepting procrastination may help alleviate the filing or deleting difficulty that can be encountered in certain circumstances.

![Figure 4.2 Keep & Delete on Personal Computers](image)

Providing containers for both the “possibly useful” and the “to be deleted” files will separate them from other files where clear binary decisions can be made. This may help decrease the
clutter that competes for human attention with the more important items. It may also make organization easier by turning individual actions into batch dumping or batch packing actions at clean up time.

With the container for “possibly useful” stuff, for example, the items can look similar to what they are in the current system so people can use them, except that they are in groups.

But at cleanup time, they can easily be packed up to have less visibility and less storage space. Search and even browse functions would then make retrieving items from it much easier.

Similar to other studies’ findings (e.g. Jones, Bruce, & Dumais, 2001), this study found that people sometimes do not save just because finding it later would be more difficult than finding it again on the Web. The large network becomes a part of personal information resources, as shown in the listed quotes in 4.2. The container for “possibly useful” items can be even extended to one for “the items I have seen before”. One participant reported a re-access difficulty experience in email about trying to find something that “I know I read it somewhere, but can’t remember what paper it’s in”. The container would be able to serve this need by doing a search.

Although keeping and organizing are related largely in both physical and digital world, this study proposes the possibility to “keep but not organize” certain types of information. Instead of fighting with human nature and limitation in front of vagueness and uncertainty, we might be able to deal with them in a more comfortable way.
CHAPTER 5: PERSONAL INFORMATION ORGANIZATION IN FOLDERS

5.1 Introduction

How people organize information in folders is a challenging question to answer especially because of the idiosyncratic, accumulative, and dynamic characteristics of this behavior. In investigating when or how or why (or why not) people create a particular folder, several factors emerged from the participants’ description and explanation during the interviews, e.g., importance, urgency, amount of items, and access frequency. However, sometimes a participant could not give an explicit reason and simply answered that “it’s just my judgment call at the moment” or “I don’t know. I just felt I need it (a folder).” To make it more complicated, it was often the case that a participant created a folder here when there were a lot of items but did not do it there where there were also a lot of items, or he/she created a folder here for it when an item was important but did not create a folder there when an item was also important, or other similar cases. It is necessary to identify elements behind these varieties and commonalities to connect together and make sense of the pieces, and furthermore provide an overall picture of how people organize information in folders.

With this motivation, the study identified four general elements plus a dimension behind all the idiosyncratic information organization methods in folders. The four general elements of information organization are corresponding to Hjørland’s typology of the four views which include pragmatism, rationalism, empiricism, and historicism. The extra dimension is the time dimension.
After briefly introducing Hjørland’s typology of rationalism, empiricism, pragmatism, and historicism, the following sections will describe the four elements in participants’ folder structures as well as some instances that do not conform to a single element, followed by the description and discussion on the time dimension. The problems in current folder system illuminated by the typology are described and discussed after that.

5.2 Hjørland’s Typology of Four Views

Hjørland started to develop the typology of four epistemological views more than two decades ago which, according to (Hjørland, 2009), includes empiricism, rationalism, historicism, and pragmatism. He proposes the typology of the four views as an analytical tool in research areas such as scientific classification, knowledge organization and bibliographic classification, among others. Hjørland points out the importance of epistemological theories in addressing the problem of how to classify documents by explaining that “(a) few approaches … do represent a specified method, but there is no analysis of the implicit assumptions behind such different methods and there exists no theoretical framework to compare the relative strengths and weaknesses in different approaches to classification” (1998b, p.611). He believes that “(t) here exist a limited number of basic methods of knowledge organization” and the different methods of classification are directly related to several basic epistemological theories (1998a, p.164). He furthermore claims that information scientists can use empirical methods, or rationalistic methods, or historical methods, or pragmatic methods, or any combinations of these methods to organize documents, knowledge, or information (1998a, p.164). This study is not focused on the epistemological level analysis, but as an extended observation of Hjørland’s typology, this study
identified the same kinds of methods in the information workers’ information organization in computer folders.

The typology is still evolving. In (Hjørland, 2009), the four views are defined as:

*Empiricism is the ideal of basing knowledge on observations (and on inductions from a pool of observations) (p. 1523).*

*Rationalism is the ideal of basing knowledge on logics, principles, rules, and idealized models (p. 1524).*

*Historicism is the ideal of basing research on social contexts, on historical developments, and on the explication of researchers’ pre-understanding (p. 1525).*

*Pragmatism is the ideal of basing knowledge on the analysis of goals, purposes, values, and consequences (p. 1526).*

These views are “idealizations,” as noted by Hjørland (2003), and they usually do not exist in pure forms. Instead, the four views are connected. For example, “any kind of pragmatism is limited by constraints set by the real world through empirical evidence” (p. 107). Hjørland (2009) also points out that there are connections between these four views. For example, pragmatism is “closely related to historicism by understanding that observations are contextual” (p. 1526) but pragmatism places more emphasis on purpose.

There are several studies using this typology in knowledge organization systems. For example, Dousa (2008) describes the application of this typology to the analysis of Julius Otto Kaiser’s Systematic Indexing, finding the hybrid nature of epistemological positions in Kaiser’s theory.
In identifying the four elements corresponding to Hjørland’s four views in information organization behavior on personal computers, the key features for each view are delineated (see Table 5.1) based on the definitions given above, and at the same time consistent with the context of personal information organization on computers. These features are used as an operational tool for data analysis and are not intended to be formal definitions.

Table 5.1 Key Features of Hjørland’s Typology in Folder Structures

<table>
<thead>
<tr>
<th>Key Features in Information Organization On PC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pragmatic Element</td>
</tr>
<tr>
<td>Purpose</td>
</tr>
<tr>
<td>Rationalistic Element</td>
</tr>
<tr>
<td>Way we envision our information spaces</td>
</tr>
<tr>
<td>Empirical Element</td>
</tr>
<tr>
<td>Judgment based on observations and inductions</td>
</tr>
<tr>
<td>Historical Element</td>
</tr>
<tr>
<td>Context, situation, and pre-understanding</td>
</tr>
</tbody>
</table>

For pragmatist view, since the basic function of folders is to separate and group, to separate/group will not be seen as a pragmatic element (purpose) in this study.

5.3 Four Elements in Information Organization Methods in Folders

5.3.1 Pragmatic element: purpose

Since the folder structures are personal forms of organization intended to support an individual’s tasks and goals, they are pragmatic in nature. Thus, it is not surprising that there are many examples observed in the data consistent with the pragmatic method.
First of all, it was found that most of the top level folders or the folders under the top level folders are for a particular purpose, e.g. project or course folders for PhD students, and job tasks such as “employee reimbursement” and “awards,” as well as projects such as “TEI workshop” for administrative participants. A participant created a top-level folder of important files and folders for backup purpose: “this is the directory I have to backup no matter what.”

The pragmatic element often exists when participants described what a particular folder is. For example, “(a folder is for a) project or type of work,” “those are all things that I do in support of the …,” “I have a folder just for alumni news,” “all these are about the visit,” “anything involved in CCB (name of a center),” “save here for this purpose, so next time if I go for that purpose, that information will be there,” “for that (field exam), I have a directory,” “(a folder) … was where I was working with … as RA,” etc.

The pragmatic element was also seen in how participants try to re-find information items. When they were asked to re-find a file during re-access task observations, their first responses often include descriptions such as “that would be related to this class,” “sounds like something I did in …,” “that was for …,” etc. One participant explained that looking for items according to what it is for was a way he finds things: “it’s usually an easy little memory device for me, oh I did that paper for this class, it’s in that class folder.”

The impact of the pragmatic element can even be implied in the cases when the pragmatic element is missing, which is there is no particular purpose for an information item: for example, a note a participant took during a talk, or an article a colleague recommended to read. Such items
were often associated with organization and retrieval difficulties. There are four emails by three participants reporting re-access difficulties for this type of documents (see section 8.2.2). As a solution, several participants in both groups used some types of catch-all containers, e.g., miscellaneous folders. A PhD student uses people’s names as such a container for files where the only common clue is by or about the person. Another PhD student created a folder “talk” in order to have a place to put notes on a talk, even though there is only one file under it. An administrative participant did a similar thing with a file that he “didn’t know where else to put it, so,” he created a folder for it. Some participants simply left the files scattered in different locations depending on their judgments at the moment, and possibly a bottom-up organization would play a role to get them organized (see below section 5.3.3).

5.3.2 Rationalistic element: the way we see our information spaces

The participants in both groups have more or less general folders at the top level, e.g., “academic,” “school,” “teaching,” “corporate interactions.” It was more salient in the PhD student group probably because the administrative participants’ jobs are more specifically task-oriented. For PhD student participants, as one of them said, these top level folders correspond to the “big chunks” in their information spaces: “…based on my activities. I have my own study, my dissertation, my teaching assistantship. In form of work, I have these two big chunks.” Another PhD student with 30 top level folders used several quick links (on a Mac) for current main parts, while another student with 10 top level folders said “I mostly think about it just in terms of file folders.” An administrative participant created a top level folder “teaching” and then a course name subfolder even though that is the only subfolder under “teaching” and she did not expect to teach other courses. Another administrative participant decided to organize emails
according to people’s names at the top level and then created project subfolders under a particular person’s folder, because that is how she thinks and recalls these emails.

The rationalistic element can also exist in subfolders. For example, a PhD student chose to organize his readings “based on how people talk about things in the literature.”

All the above examples show that the participants try to impose their way of interpretation, the way of envisioning their information spaces into their folder structures.

5.3.3 Empirical element: judgment based on observations and inductions

The Empirical element in information organization in folders is based on bottom-up analysis. The most prominent phenomenon of this element in the study is that many participants would accumulate files for a while and then decided to create a subfolder when they realized that it was needed. This judgment can be based on not only the similarities between items, as listed above in the definition of empiricism, but also the number of items, importance, access frequency, and so on. For example, an administrative participant said:

“it soon became clear that it’s going to require a lot of communications. So after the first couple of dozen messages just stuck in here, I decided it deserves a subfolder.”

Similarly, a PhD student said:

“At some point, I’ll probably have a bunch of stuff under this directory, then I’ll create a directory for that and move a bunch of things to that directory.”
Because there is no systematic place to “wait,” such files sometimes were scattered in different folders and may not get collected when a subfolder was created later. For example, an administrative participant created a folder for two files about internships, without remembering that there was another file about internship she had received earlier but had left in the parent folder.

In many participants’ home folders/drives (which includes top level folders and files) and subfolders, there existed many individual files, potentially waiting for this bottom-up organization. They were mixed with other files that are intentionally left there, e.g., important ones, templates, more frequently used ones, and the ones put there for reminding purpose. This study observed that two students grouped certain files under their home directories between the two interviews. For example, a student created a “misc” folder at the top level and moved several files that were originally under the home folder into it. Some other examples are described at the end of the above section 5.3.1.

5.3.4 **Historical element: context, situation, and pre-understanding**

The historical element emphasizes context, situation and pre-understanding. It exists implicitly in many folders. A prominent phenomenon demonstrating this element is that folder names do not always carry a literal meaning.

For example, an administrative participant had “TEI WORKSHOP” and “publications” as two folders at the top level among many others. But the contents under them are all invoice vouchers with different clients. They are separate from another top level folder “invoice vouchers”
because they have their own account and the means of payment processing is different from the others. Because a major task the participant was responsible for was various payments and money transactions, there is a “pre-understanding” behind these folders that they are invoice vouchers, even though it was not labeled in the folder or file names.

A PhD student participant had a folder “atlas” (a particular software name). But the content under the folder was her data files instead of the software. Another participant had a folder named with a person’s name, but it was a folder for the project that she had been working on with the person, not a folder about that person or the person as the author. A different participant had an “articles” folder for articles that she might use for her dissertation.

This implicit “pre-understanding” may make perfect sense to the user, as a participant declared, “because that’s something I used, so I know what it means,” but the real meanings with the names are not transparent to others. All these examples indicate that folder names have to be interpreted in the light of the user’s particular understanding, context, and situation, and cannot be interpreted literally or strictly.

Situation factors can change organization behavior. For the “big folder” examples described at the end of section 4.2, the administrative participant and the PhD student participant did the similar thing in keeping a large number of items without a subfolder under the similar situation when in a fast growing project.
Another phenomenon of situation factors influencing folder creation is that sometimes folder creation and structure are not fully controlled by users. They can be determined by system and software. For example, several participants had music files saved by iTunes at a particular “default” place.

5.3.5 Combination element

Although a few organization behaviors observed in this study can be characterized by a single element as described above, many other behaviors in folders do not fit neatly in a single element, and reflect more than one element.

For example, an administrative participant talked about how she refined the folders over the time. She used to have all the years of reimbursement forms together, but then within the last a couple of years she found “it got too big, it became hard for me to manage,” so she break it down according to fiscal year (e.g. “2008”) for the last a couple of years’ reimbursement forms, because “we go our accounting by fiscal years” and “if you want to find some document, like XX comes to me and says I need your help to find certain document, and I could ask what fiscal year it was done, and that’s easy for us to locate it.” Interestingly, a new administrative participant who had been in this institution for 3 months created a new subfolder with the year (“2008”) between the two interviews, and said that the strategy of separating the content with fiscal year was recommended by a colleague. In these examples, the observation that the folder “got too big” is an empirical element, while the fiscal year subfolder can be seen rooted as a pragmatic element. Later on it may be internalized into a rationalistic element since it seems to become a part of the institution-specific knowledge.
As another example, a PhD participant had a folder with the name of the school. Later as he realized that he was going to graduate, he decided to split this school folder into two, with an additional folder named “academic,” so that he could bring everything under the “academic” folder with him after he graduated and archive the other school folder. This example shows that the rationalistic element was connected with a particular situation (historical element) at a certain stage in the program, and was influenced by a specific purpose (pragmatic element) of “bringing with him after graduation.”

Another example is a special phenomenon of folder structure observed in the study. A participant had two folders at the top level as illustrated in Figure 5.1.

Figure 5.1 Example “Meeting” Folders

```
.....
Meetings/
<Dean’s name> /meeting/
```

The second folder is for the meetings that the Dean had asked to arrange. Thus the top level “Meetings” folder is for the other meetings. These general-exceptional folder structures were observed in several participants’ folders in both groups, and the reasons behind it identified in the data include priority, anticipated access frequency, and anticipated volume of items. It is worth noting that since this perception is hidden and mainly makes sense to the user, it could cause confusion to others. For example, a participant who was working on a shared drive for a project was not aware of an existing “meeting” subfolder under a top level folder “advisory
committee” and thus put the files about the advisory committee meeting under a top level folder “meeting and agenda,” which caused confusion for the group members. Both pragmatic and rationalistic elements can be seen in this type of general-exceptional folders.

5.3.6 Time dimension: accumulated over time

An element that seems outside of the typology and is unique in personal information organization systems is the time dimension. The “on-going” characteristic of organization strategy (Boardman & Sasse, 2004) has been noted by researchers. Studies found that people may change their strategies over time, e.g., from regular filers to spring-cleaners (Balter, 1997). Ravasio et al. note that “classification and document storage were considered on-going processes” by the subjects, and “hence no structure was thought of as permanent” (Ravasio et al., 2004).

Consistent with the above findings, this study found that both information items and folder structures accumulate over time. For example, every participant in this study includes folders tagged with “new,” “old,” “archive,” or “work in progress” to indicate the time dimension in folder structures. Although these folders were mixed with current ones, participants use the specific folder naming mechanisms to separate them out.

This accumulation makes the personal information organization system on computer an actual “organic” personal digital archiving system.
5.4 Problems with the Current Folder System

Looking at participant behavior in this framework demonstrates problems and limitations due to functionality and interface constraints of current file organization systems.

5.4.1 Pragmatic element over time: when purpose changes

Although many folders are constructed for a particular purpose, there might be a problem when the purpose changes. For example, a PhD student found that something he worked on for a class became a research area that he wanted to work on. But it was still under the class folder. The participant said: “I keep meaning to change it.” A similar situation happened to another PhD student and he was planning to copy the related files out to separate them.

Sometimes the change could involve more than copying things out. It may need to retrieve items from a perspective different from the available folder structure. For example, two PhD students encountered a similar situation where they needed to access all the readings scattered in different courses and projects. This is difficult in current systems and both participants ended up re-downloading many files again. This may be seen as the need for a mechanism to support multiple classifications. But it is more a need for multiple access points. Using folder names and even file names as tags at retrieval time in the current folder system may help alleviate this problem.

5.4.2 Rationalistic element

The folder system is good at building in a certain type of rationalist view, but the problem is we may have multiple views and perspectives with same set of information items. For example, a
PhD student wanted to be able to look at his files from both research perspective and project perspective: “but those two things kind of overlap, and I haven’t quite figured out what’s the best way to overlap files that are related to a project but also related to my research agenda.” This lack of an ability to do multiple classification is a well-known major weakness in the current folder system. Some systems such as Gmail and Zotero use “labels” or “collections” instead of folders to organize emails, web pages, and articles. But file organization system is more complicated than emails and pdf files because many files on personal computers are editable. The aforementioned multiple classification problem is not easy to solve before a mechanism dealing with the derivative relationships between editable files is provided.

Another limitation of current folder systems is that the strict top-down visualization hides information and forces step-by-step or click-by-click navigation. As a complementary mechanism to a tree view, a “map” view should be provided to give an overview of the folders and access.

5.4.3 Empirical element: waiting mechanism for bottom-up

For the empirical organization method, people need a better “waiting” mechanism to postpone creating a folder before they decide to do so. This study observed many individual files scattered in folders, especially at the top level folder or drive, where some other files that were left there purposely, e.g., some files were left at the top level to remind the participant, or they are important, or frequently used. But these important files are mixed with the other files that were there just because there was nowhere to put them or the participant hadn’t decided whether or not to create a folder for them. As a PhD student commented on one of his folders: “it just confuses
everything in here, because these are archive or reference directories, these are actual content directories, and this is a specific kind of content directory that I used a lot.”

The lack of the “waiting” mechanism can be seen as the cause to the “premature filing” described in (Whittaker & Hirschberg, 2001). The need for this mechanism implies that describing a file by tagging and grouping (e.g., general genre, content, source, etc.) to show “what it is” does not have to be the same as “classifying” it into the folder structure which requires knowing “what it means.” This implication echoes the observation in the method reflection described in Chapter 3.

5.4.4 Historical element over time: derivative relationship tracking and file naming

For the historical element, one problem in the current system is that the folder and file naming is limited in conveying the context. The study found that sometimes the folder or file names used to encode the context may not make sense later even though the participant tried “to be descriptive when I name the files.” As a PhD participant put it:

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

Another PhD student also made similar comments.

A related problem is the file naming mechanism. As an important context encoding method, file names sometimes are controlled by other factors. For example, a participant had a file he worked on with another person who named it in a different way than this participant did; another
participant changed a file she worked on to a different name when she sent it to someone else, “in order to make it more communicative for her so she knows this is mine.”

Although it’s technically and practically impossible to encode all related context into a system, there is some context that a computer system can capture and present, e.g., the derivative relationships between information items in a broad sense which will be described and discussed in detail in the next chapter.

5.5 Conclusion and Discussion

This part of study shows that people’s overall folder organization is a mixture of the four elements and thus a mixture of top-down and bottom-up processes, although the current top-down hierarchical folder system may indicate a pure top-down system. It emphasizes the need to allow these multiple perspectives in any personal information management system. The four elements plus the time dimension give an overall picture that help connecting and integrating the idiosyncratic pieces of observation in this study and make sense of them.

This is the first effort to explore the applicability of Hjørland’s typology in helping make sense of personal information organization behavior on computers. Since the four elements have been observed in some classification system, it’s interesting to see the connections between grass root classification methods and professional knowledge organization methods. Overall, although there are some limitations of Hjørland’s typology, it seems to be a promising analytic descriptive framework for understanding personal information organization methods.
This chapter’s study is a post hoc analysis with available data obtained in the overall study. In future work, larger scale and more focused studies are necessary to explore how the typology can be extended and refined to better suit personal information organization on computers.
CHAPTER 6: FOLDERS AS WORKPLACES AND ITS IMPACT ON FOLDERS

6.1 Introduction

The folders in current file systems on personal computers are not only storage places for information organization and retrieval, they also serve as potential workplaces for corresponding projects or tasks. It has a great impact on personal information management and organization.

After outlining two overall characteristics of the participants’ file folder structure and organization, this chapter will describe two extreme types of folders in a continuum of inter-file relationships as the result of using folders as workplaces, and further suggest that the file system is not just about managing a large number of isolated files, but also about managing the relationships between files.

6.2 Folder Breadth and Depth

Although this study is not for comparing between the two participant groups, the collected disk scan data shows that the administrative participants had broader and shallower folder structures than the PhD student participants. As described in Chapter 3, the disk scan data include the top level folder and 2 or 3 other folders randomly selected from each participant’s computer. Figure 6.1 shows that the administrative participants generally had more top level folders than the PhD student participants at the first interview time; Figure 6.2 shows that the maximum depth of folders among the disk scanned folders at the first interview time. For the convenience of illustrating differences between and within groups, the values of the participants in the two
groups are ordered by size and interleaved. The results at the second interview had the similar pattern.

Figure 6.1 Numbers of the Top Level Folders (1st interview)

![Bar chart showing numbers of top level folders for PhD and Adm participants.]

Figure 6.2 Maximum Depth of Folders (1st interview)

![Bar chart showing maximum depth of folders for PhD and Adm participants.]

6.3 Information Organization Beyond Computers

As described in Chapter 4, the PC, online, and paper worlds are closely connected in participants’ keeping behavior (as illustrated in Figure 4.1). Similarly, this study found that the use of paper (by printing out) and Web services (e.g., del.icio.us, Refworks, or course work system such as Moodle) have become a part of participants’ information organization mechanism. For example, an administrative participant uploaded her pictures from her Mac machine at home to Flickr not for sharing, but for better organization with the tagging mechanism. Another participant used del.icio.us for storing, organizing, and sharing his bookmarks.
6.4 Two Types of Folders

This study found that folders can be described as belonging in a continuum between two extremes that are named “genre folder” and “project folder,” respectively. A genre folder is usually corresponding to a type of files, e.g., “reimbursement forms,” “invoice vouchers,” “meetings,” and “readings.” Figure 6.3 shows a part of an example genre folder of employees’ reimbursement forms according to the disk scan result of an administrative participant’s computer. Genre folders often have the same file format (spreadsheet, pdf, etc.), and the files being of the same form (employee reimbursement, research papers, old drafts, etc.).

Figure 6.3 Part of a “Genre” Folder

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

At the other end of the continuum, a project folder was one created for a open ended or unstructured project, and was often named with the project name or initials, organization names, or the corresponding person’s names. It often had a more heterogeneous collection of forms and formats and a less systematic naming structure. Figure 6.4 is a part of a research project folder from a PhD student’s computer. The project and person names in both figures have been anonymized to preserve privacy.
Figure 6.4 Part of a “Project” Folder

ProjectA/4.txt
ProjectA/AmunChronicle11-17-07onecolumn.pdf
ProjectA/ResMacro_2007_win.sxw
ProjectA/new_members.txt
ProjectA/2008forcompart.txt
ProjectA/formatNewRes.txt
ProjectA/flashdrive_bkup.tar.gz
ProjectA/ResMacro_2007.sxw
ProjectA/macro_errors.pdf
ProjectA/macro_errors.ps
ProjectA/pwd-hg.gpg
ProjectA/drop-reasons.txt
ProjectA/amun_res_2007_upd.sql
ProjectA/amun_res_2007_upd.sqp
ProjectA/XrayTool52_en.sxw
ProjectA/XrayTool52_en.zip
ProjectA/Current-DB.mysql
ProjectA/res_update.mysql
ProjectA/tech-stuff.txt
ProjectA/web_data.mysql
ProjectA/07agenda.doc
A key difference between these two folders is that, in the genre folder, when the participant was pointed to a randomly selected file and asked about what it was, he or she was able to tell what exactly the file was from the file name, without the need to look at the files or folders around it. For example, in Figure 6.3, the file “November 2006 Austin Tx.xls” under “Jen” is the reimbursement spreadsheet for Jen’s trip to Austin in Nov. 2006. However, in a project folder, when the participant was asked about a randomly picked file, he or she often had to look at the other files or folders around it before being able to recognize or remember what exactly it is, although the file was named to indicate what it is. As a PhD student explained during the interview:

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

This “look-around-and-recognize” phenomenon seems partly caused by the fact that the files are related and the relationships between files in a project folder are more complicated than those in a genre folder. In the genre folder “employee reimbursement” two files are simply two separate payment statements. To understand one does not require looking at any other files. The file names were sufficient to describe and understand what the file is. This type of folder is more like a traditional paper archive folder in offices - the source metaphor of the current folder system on personal computers, so it is not surprising that the folder system works well for this type of folders. When asked about how they think about the folder system, the administrative participants were generally more satisfied with it, compared with the PhD students in this study.
On the other hand, under the project folder in Figure 6.4, the relationships between files and/or subfolders are more complicated. There might be data files, code files, reference articles, notes, and some files may have different versions. This type of folder is not as closely analogous to the traditional paper archive folder in offices. It is more like the assembly of materials of work-in-progress on a physical desktop over time. But there is no mechanism in the current folder system to capture and represent the relationships between these agglomerated materials. With the limited representative capability of file/folder naming mechanism, the user has to largely bear the various relationships in mind in order to recognize and understand a file later at the re-access time.

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

We found that the PhD students had more folders at or close to the project folder end than the administrative participants, and the administrative participants had more folders at or close to the genre folder end. This might be related to the above finding that the administrative participants had broader but shallower folder structure than the PhD students, as illustrated in Figure 6.1 and 6.2.
Although there was a difference between the two groups of participants, a participant usually had both types of folders on his or her personal computer. For example, an administrative participant had a project folder “campaign” while a PhD student had a top level genre folder “readings.”

6.5 Folders as Workplaces and the Impacts

The genre-project continuum in folders can be partly attributed to the fact that file folders on computers can be used as workplaces in addition to serving as storage for organization and retrieval, which in a physical world are usually two different facilities. In a workplace, people need to assemble needed and possibly useful resources so that they are easily at hand both for efficient use and as a reminder that they may need to be used. They also need to keep work progress and status traces. Thus many working files are kept and accumulated under a project folder which helps to convey a status and track of work progress. A simple example is the old version files many participants kept under some folders, which were useful for tracking when the user was working on the document. After the document was done, the old versions often became useless. In a traditional paper archive, these old versions usually will not be kept in a folder. But within a computer folder, the user may not delete them because he or she may want to keep them as archive or context, or simply does not bother to spend time to delete them, as observed in this study.

This use of folders as workplaces makes “context” information an important component in folders. The current folder hierarchical structure and naming mechanism provides a way for users to build in metadata about files and relationships between files and/or groups of files to
some level. In addition to the finding that the participants organized and named to explicitly engrave information item’s content (e.g., author or genre) and context (e.g., “what it is for” and “who asked me to do it”) into the folder system, this study also found the implicit contextual or workflow information, recalling studies in (Ravasio et al., 2004), (Henderson, 2009), and (Jones & Teevan, 2007). For example, an administrative participant intentionally kept more files in order to keep the way of doing a particular task:

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

This indicates the “documentation” function of a folder in recording what the user did and how he/she did it, although other people may have difficulty in deriving process and rationale from the related documents. In another case, a file’s existence under a particular folder informed an administrative participant of the status of the corresponding paperwork in the workflow, because the specific file process procedure and convention is implicit in the folder structure.

Despite such creative uses of the folder system, the hierarchical structure and naming mechanism remains limited in conveying context, and people often have to largely rely on their memory. In this study, several students spent a while to recognize or find a file. A student explained that “I just don’t remember what I did in the class.” Two students had difficulty in finding particular files because they forgot the way they worked on the files – using LaTeX, which had caused the files to be stored in a different location.
It is not feasible to try and consider all possible context and useful relationships, but there are some that seem to be important and recur enough to merit attention. In the PREMIS Data Dictionary for Preservation Metadata working group, three basic types of relationships are identified between digital objects: structural, derivation, and dependency (2008). This study observed examples of these three relationships, as well as others. For example, a kind of structural relationship exists between a dissertation file and the files of the individual chapters; a configuration file and the main program file have a kind of dependency relationship. An example of another type of relationship is that between a website registration receipt and the purchased article. Among all these types of relationships, however, this study found that various derivative relationships are the ones that the participants found the most difficult.

6.5.1 Derivative relationship

PREMIS defines a derivation relationship as one between digital objects where one is the result of a transformation performed on the other. In this study, derivative relationships are used in a broader sense in that one derived file can also be copied or adapted from the other. The derivative relationships between files can have different levels of complexity. Two types of derivative relationships are identified and described here: versioning relationships, and the more complicated derivative relationships.

6.5.1.1 Versioning: current and final version

Similar to (Henderson, 2004), this study found that managing different versions was a troublesome problem for the participants. Participants used file/folder names to differentiate
versions. For example, a participant used “final_” and “final_final_” as a prefix trying to indicate the final version of her dissertation. Another participant used “stage 1” and then “stage 1 new” to differentiate between folders of two versions.

Participants had a common need to make the current version and especially the final version very distinct from others. An administrative participant wanted the final version to “popup” form others so she “would know what I need to look at first, instead of looking through 20-25 folders or files.” A PhD student said:

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

Folder or file naming strategy to note the versioning relationships did not work well when the final version had to be put in a different folder for a particular purpose, since the connection to the version chain became disconnected. For example, several administrative participants worked in their own directory and then put the final version on a shared drive/folder to share with colleagues, and the names of the final version files became less useful. A PhD student put all final papers together in a folder for a particular use. But later on, the version under the original folder was changed without updating the final folder, which caused confusion when the participant tried to find the real final version some time later. Recalling the discussion in 5.4.4, the file and folder naming mechanism in current folder systems is overloaded and has to describe “what it is,” “what it is for,” plus the relationships with other files or folders. To make it worse, file names sometimes are controlled by other factors.
6.5.1.2 Complex derivative relationships

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

It is difficult in current hierarchical structures to represent complex derivative relationships between files, other than by adjacency to mean “somehow connected”. A participant reported a re-access difficulty in trying to identify a particular excel spreadsheet among four spreadsheets which have a complex derivative relationship between them:

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

Although there are only four files in the folder, the relationships between them were so complicated that they were beyond the participant’s capability to remember from their file names.

In some other cases, the relationships could be between information items, and could be between projects since projects evolve. A PhD student had over 38 top level folders, and most of them were projects or course folders that served for a large study with complicated derivative
relationships. She felt her files were not well organized and had frequent file re-access difficulties:

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

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

6.5.1.3 Citation relationship

Citation relationship is not included in the three PREMIS relationships, but it can be related to derivative relationships when the quotation includes a part (e.g., text) or the whole (image) of another file. Citation relationships also caused some difficulty and participants noted they would like a mechanism to link the citing and cited items together. It can be a paper citing or including quotations from several other articles, or a reading note reviewing one or several articles, etc. A student reported that:

One thing that I’ve struggled with is finding pieces of information to include in my dissertation. ... I was looking for a complete citation for a particular book. But here's the thing, I could only remember the author and that I had cited it in another paper I wrote. It was not difficult for me to find the information once I remembered what paper I cited it in.
Another student reported a re-access difficulty case involving a citation mixed with a versioning issue:

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

I did eventually find the citation I was looking for, but it wasn't in the file I thought it was in. Turns out I had not used the citation for the final paper, but I then recalled that I had been experimenting with a new citation software that used BibTeX files as a database. The BibTeX file wasn't in the directory with other Word documents, but stored in another directory. I couldn't recall the file name, but knew it would have a .bib extension and the approximate date I last used it. I was able to perform a search using Spotlight that found all the .bib files, which I then sorted by date to find the right file.

During an interview, another student expressed the hope that “ideally I can specify which readings are used for which papers.”
6.5.1.4 Overview and discussion of derivative relationships

It is important to emphasize that the complexity of inter-file relationships, including the various derivative relationships, falls in a spectrum similar to the genre and project folder spectrum. For the many folders observed in this study, the complexity of inter-file relationships under a particular folder may locate somewhere between the two ends. The point is that the system support should target to the whole spectrum instead of the simplest end.

The various derivative relationships are a part of context information about particular files. These relationships in this study are used in a general sense. They not only include source and target files, but may also include the involved actions and related objects, e.g., who sent it to the user, where the user got/downloaded it, whom the user sent it to, what the user did on it, etc. A lot of these details can be captured automatically by computers. Such information is very useful for people retrieving and understanding their information items. The traces and bidirectional links to connect files together, with various derivative and citation relationships, remind us of the association trails in Bush’s Memex (Bush, 1945).

6.5.2 Three spaces: work-in-progress, main folders, and archives

Using folders as workplaces imposed other needs upon the system than traditional organization and retrieval, e.g., reminding and prioritizing, planning, quick access, and version tracking, among others, since the basic functions of a workplace are to have related documents readily accessible and at the same time keep track of work progress. Related to this, the study observed several participants’ efforts in trying to separate their workplaces from main folders.
Similar to Cole’s finding with paper documents which includes “action information,” “personal work files,” and “archived information” (Cole, 1982) and Barreau and Nardi’s “ephemeral,” “working,” and “archived” information in the digital world (Barreau & Nardi, 1995), we saw a clear three-way pattern in most of the administrative participants’ email systems. For example, an administrative participant had emails for events in three places: the inbox for incoming items and things that need to be done or followed up; these are then filed to inbox folders after they are taken care of; later they are dragged to a local folder (as archive) after they are done because of quota limitations in the email account. Local folders as archives are “only for occasional reference”, the participant noting “I don’t want it cluttering my express (the email account folders)”. A folder with the same name under the inbox in the email system is for current email. This pattern is not as clear amongst PhD participants, probably because of the different types of work done.

Both groups showed a similar pattern in their files: a main folder, work-in-progress files or folders, and the archive. Participants in both groups try to keep these separate. They either have folders in archive status or folders specific for archived files. These can appear under any folder such as that for a project or for a genre, and may be indicated with words such as “old”. Two features are observed for the archived files/folders: intended low visibility (e.g., “I don’t need to see it anymore. It’s history to me”), and keeping the original structure. Most participants do not change the structure of a folder after it is archived, either not bothering to change or intentionally keeping the original context.
Two participants in both groups have separate work-in-progress (or “currently working”) folders or files from the main folders and archives. For example, a PhD student said:

*It’s like the current and old. That’s how almost all my folders are organized. Current stuff is in the main thing and old stuff is in the archive.*

And at the same time, this participant uses several quick links supported by the Mac to point to the folders that he was currently mainly working on, among the mentioned “current stuff”.

Another PhD student using a Mac did a similar thing, while a third student using a Mac used the desktop for a short term and unimportant project, but went to a subfolder deep in the main folder every time for her currently heavily used working folder for her dissertation.

Among Windows users, one PhD student used the desktop for her current working projects (e.g. dissertation, job hunting, applying for visa for her mom, building a video with her daughter’s pictures, etc.). After a project was done, she moved it to her main file folder. Another student’s current working course folder was not organized according to semester because “these are the classes I’m actively working on”. Similarly, an administrative participant used the desktop for a project that is “still going on, but develops slowly” over a long period to remind her and not lose sight of it.

Another administrative participant created a “work-in-progress” folder in both the file system and email system with a similar purpose. But the design did not work as smoothly as expected. There were version problems and somehow it caused confusion in what it exactly included and where a current/final version was located between the work-in-progress folder and her main
folder. Again version control problems arise and a better mechanism is needed for supporting working space which allows for iteration of versions separated from the final one, and the use of other related working files, especially when the number of work-in-progress projects increase and more threads need to be tracked and reminded.

6.6 Conclusion and Discussion

This study observes the impact of using folders as workplaces on the folder content and structure. It showed how the folders that people created on their personal computers could be usefully understood as fitting in a continuum from “genre folders” with no inter-file relationships to “project folders” with complicated relationships between files. The study found that various derivative relationships between files and/or groups of files are the ones that caused the participants most difficulties in finding and identifying files at re-access time. Recalling what is described in above Chapter 5, this chapter emphasizes the problem of overloaded file and folder naming mechanism in current folder system.
CHAPTER 7: FOLDER RE-ORGANIZATION

7.1 Introduction
This chapter describes the findings about why, why not, and how the participants want to re-organize their folder structures.

7.2 Why to Re-organize
Although the participants were found to be generally satisfied with their file organization in serving their purposes, the study found several situations when a folder does not make sense later.

7.2.1 The current structure decays
One type of situation is that although two folders were originally created with clear definitions and rationale that the user “thought would make sense”, they became overlapped and “not separable” over time, as a PhD student commented:

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

An administrative participant had a similar problem: two folders’ contents became overlapped over time. During the guided tours at interviews, the participants sometimes commented that she/he should create a subfolder somewhere or some folders were not well organized. But at least
during the study, they didn’t change them or even have intention to change them (see below section 7.3.1).

7.2.2 New purpose

Sometimes the participants could have new needs and uses for the same items. For the example described in Chapter 5, the two PhD students were at the stage when they had to get all the readings scattered in various courses and projects folders, and ended up re-downloading many of them instead of re-organizing the available ones, simply because collecting them from the various places was “too much effort.”

In this study, four students had some course projects that evolved into proposal projects. For example, a student had a proposal under a course folder because “it wasn’t intended to be a dissertation proposal when I started it, which is why it’s in here,” and now “probably it’ll get moved.” One of the four students kept the proposal files in the course folder and thus the folder became the dissertation work folder, while the other three students decided to copy the related content over to a separate folder and keep the original course folder intact. This “keeping original folders intact” strategy was mentioned by most students as well as several administrative participants when they used old content for a new purpose.

Sometimes new rationalistic or pragmatic element emerged and the participants wanted to move things around to fit the new important perspectives. For example, a student changed his overall folder structure when he started the PhD program because:
...the reason I changed it is because it didn’t let me make distinctions that now are really important for me. So like my first distinction that I drew was that I needed a <universityname> folder just for all my <universityname> stuff.

And later, a while before the study, he changed it again because:

* I realized ...I’m gonna graduate from here some time so I need to separate this stuff (academic) out. ...I moved a bunch of stuff (from <universityname> to academic).

Probably because of the importance of the new purpose, most of the participants re-organized or adopted some workarounds in these situations.

### 7.3 Re-organization Behavior

#### 7.3.1 Why not to re-organize

Even when folder structures decay in their usefulness, participants may not always re-organize them. Re-organization seemed to be a low priority and conducted only when it is deemed a really big issue, although two participants moved a subfolder to top level between the two interviews. An administrative participant was leaving the job at the second interview and she did a big re-organization with her folders which she said she wouldn’t do if she stayed and the re-organization is “only for transition to the next person.” Her usual way of re-organization was to fix the problem when coming across it, instead of in a systematic way. A PhD student said that “I have done [re-organization] a little bit, occasionally, but not very often.”
The reasons that the participants chose not to re-organize include: the participant had been familiar with the structure even though it has some problems; the folder content was not important; the participant did not re-access the folder frequently; or the problem did not affect the work. For example, a PhD student explained that:

*I think one of the reasons that I don’t go back and re-organize things is that, once they get in a certain way, like I can remember how they are, even if they don’t really make sense.*

A related observation is that participants usually did not organize when a project was done and archived.

7.3.2 Small refinements and adjustments all the time

On the other hand, it seemed that small refinement and adjustment in the folder structure happened all the time for each participant during the several months in this study. For example, creating a subfolder, moving a file to a subfolder, moving a folder to another place, and so on, were seen as a part of the evolution of the folders and were observed on every participant’s computer during the two interviews. As an administrative participant said:

*I have to figure out what would be easier for me to find things. We added more staff, ... it became hard for me to manage with the more people that we hire.*

Another administrative participant said that:

*The forms folder I moved from mine to here. ... but sometimes things just evolve. ... such as attachments for email. I didn’t have an email attachment folder before, it was all in my*
inbox. But my inbox got so big that it made me create a file so that wasn’t there. And to archive events, every year have to add a year so that’s constantly changing.

7.4 Incomplete Re-organization

The study found that most re-organization activities were not complete. It was typical that a participant created a folder for a new file at some time, but did not collect other files scattered in other folders into that folder. For example, when a participant and the interviewer were looking at an email folder <person name> and talking about an earlier message by the same person, the participant said “I didn’t save it. I deleted.” But later when she looked at another folder, she found the message was filed there because when she received that message, she hadn’t yet created the folder <person name>.

This incompleteness is common across many participants’ email and file folders, and can cause re-access difficulties. For example, during a re-access activity, a participant went to the corresponding subfolder directly and didn’t find the target file. After she looked around under the upper level folder, she found it and moved it into the subfolder. Some other examples:

- …I think I had SCI in my old computer (under folder “research”), and copied it over, and then I thought it would make more sense to have, because then I was doing research in two different departments, so I thought it makes sense to have a <departmentname> directory (for this project). But then obviously I never got around to actually moving the other one. …it’s the same project.

- …I also remembered I did some re-organization of something, and I moved one copy of Shannon’s paper to this “research/papers/information theory” directory, …so I have a copy in
there but I still have another copy on my desktop, too. ...I actually made two. I made a “papers”
directory which is under “research” instead of under the top level, and then “information
theory”. ...There’s still stuff in there (a top level folder “reading”), too. I haven’t actually re-
organized anything. I just created some new directories when I started to save. ...stuff I have
downloaded I haven’t moved .

• ...I think it (a file saved under home directory instead of the appropriate folder) was not
long after I got this computer and I hadn’t set up those directories yet. And like the many other
things, I just haven’t moved it, it’s just sitting there. .

• ...To make things more confusing, I also have a directory for annual review 2007 under
the “academic” folder, if I recall correctly, that’s actually the correct folder for it, ...I think I
made the other directory (under “school” instead of “academic”) and then decided that it would
be under here more appropriately and didn’t go back to delete the old directory. .

• ...This is all stuff that I haven’t re-organized. This is all the legacy from the last
organizational system. From before I did the re-organization to this model. So at some point, I
need to go through this, but. ...these are things I just haven’t got time to go through it.

• ...I thought it might be here, but. The thing is that I saved it before I had this academics
folder.

Sometimes lack of importance or not using it frequently is a reason for incomplete re-
organization:

...one of the reasons I never got it around to straighten it out is because I’m not working
on that any more, I’m just doing a couple of things kind of finishing. ...Like if I’m
working on it a lot, it would really bother me, and I would fix it. My motivation to fix it is too low because I’m not using it very much.

Overall, the reasons for re-creating instead of re-organizing and incomplete re-organizing might be the “too much effort” caused by the lack of “waiting” mechanism at the filing time (see Chapter 5) and the lack of a good re-collecting mechanism across folders with a particular element “tag”.
CHAPTER 8: RE-ACCESS BEHAVIOR AND DIFFICULTIES

8.1 Overview

This chapter investigates the participants’ re-access behavior and difficulties by analyzing the re-access difficulty cases they reported via email, the corresponding discussion about the cases, and the re-access task observations during interviews. Because the recruiting email and the reminding emails (see Appendix A and B) have the same general structure, each case includes the corresponding five parts: circumstances, importance/urgency, what were remembered, re-access strategies, and found or not (see Table 8.1)

<table>
<thead>
<tr>
<th>Code</th>
<th>category</th>
<th>Questions in emails</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Circumstances</td>
<td>What were the circumstances surrounding the file? E.g., what is it about and what was it for, how often did you use it, was it important for you, and where did you get it?</td>
</tr>
<tr>
<td>B</td>
<td>Important/urgent?</td>
<td>Why were you looking for it? Related to your work/study or something else? Was it important or urgent?</td>
</tr>
<tr>
<td>C</td>
<td>What were remembered?</td>
<td>What could you remember about the file? E.g. filename, source, actions you took on it (edit, copy, save, create, move, etc.) and/or the rough date, keyword, topic, format, etc.</td>
</tr>
<tr>
<td>D</td>
<td>Re-access strategies</td>
<td>What strategies did you try to find it?</td>
</tr>
<tr>
<td>E</td>
<td>Found or not?</td>
<td>Did you find it?</td>
</tr>
</tbody>
</table>
Among the total 45 reported email cases, there are 42 difficulty cases and the other three are example re-access behavior descriptions (participants were asked to describe how they re-access a document when they did not have any difficulty examples). Among the 42 difficulty examples, 25 of them are from the PhD students, 17 from the administrative participants. The following sections will report the results on the above 5 parts.

8.2 Circumstance

From the description in the emails, complemented with further explanation during interviews, “circumstance” data is analyzed and coded in four aspects, although certain aspects do not apply in some cases:

F. What is it:
   1. a particular file
   2. a particular version of file
   3. a set of files
   4. a citation (the citing file)
   5. an email message

G. Situation:
   1. Was for specific purpose or not?
   2. Saved by self or software or else?

H. Reason:
   1. Possibly misremembered?
   2. Forgot where it was
   3. Didn’t know where it was.
4. Know where it was, but not accessible
5. Know where it was (roughly), but didn’t know which one: Forgot the way of working (e.g. using a software)

I. Use (length/duration, or rate – temporal point)
   1. Rarely used – not in a while
   2. used a lot – not in a while
   3. not in a while
   4. rarely used – recently
   5. used a lot – recently
   6. recently

The overview of the circumstances is illustrated in Table 8.2. As the table shows, when the problem in a case is wrong version (i.e., “what was it” value is F2 “a particular version”), the other circumstance factors become irrelevant or not important at all. The table also shows that for the cases where the target document is an Email message, they do not have G, H, and I factors because they do not apply.

Because the different degrees that the participants were involved and responsive during that period, some participants reported more cases than others. It does not necessarily mean that they had more difficult experiences than the other participants. Thus individual difference in terms of the number of difficulty examples is not considered in this study.
Table 8.2 Overview of Circumstances

<table>
<thead>
<tr>
<th>id</th>
<th>F</th>
<th>G1</th>
<th>G2</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PhD1</td>
<td>1</td>
<td>Not specific</td>
<td>Self</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>PhD1</td>
<td>3</td>
<td>Specific</td>
<td>Self</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>PhD1</td>
<td>1</td>
<td>Not specific</td>
<td>Browser</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>PhD1</td>
<td>1</td>
<td>Specific</td>
<td>Self</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>PhD2</td>
<td>2</td>
<td>Specific</td>
<td>Self</td>
<td>2 (which computer)</td>
</tr>
<tr>
<td>6</td>
<td>PhD2</td>
<td>1-&gt;4</td>
<td>Specific</td>
<td>Self</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>PhD2</td>
<td>1</td>
<td>Specific</td>
<td>Software</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>PhD2</td>
<td>2</td>
<td>Specific</td>
<td>Self</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>PhD2</td>
<td>1</td>
<td>Specific</td>
<td>Self</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>PhD2</td>
<td>4</td>
<td>Specific</td>
<td>Self</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>PhD2</td>
<td>1</td>
<td>Not specific</td>
<td>Self</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>PhD2</td>
<td>1</td>
<td>Specific</td>
<td>Self</td>
<td>1, 2</td>
</tr>
<tr>
<td>13</td>
<td>PhD2</td>
<td>1</td>
<td>Specific</td>
<td>Self</td>
<td>1, 2</td>
</tr>
<tr>
<td>14</td>
<td>PhD2</td>
<td>2(wrong version)</td>
<td></td>
<td>Self</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>PhD2</td>
<td>2(wrong version)</td>
<td></td>
<td>Self</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>PhD3</td>
<td>1</td>
<td>Specific</td>
<td>Self</td>
<td>5</td>
</tr>
<tr>
<td>17</td>
<td>PhD3</td>
<td>1</td>
<td>Specific</td>
<td>Self</td>
<td>5</td>
</tr>
<tr>
<td>18</td>
<td>PhD3</td>
<td>3</td>
<td>Specific</td>
<td>Self</td>
<td>1, 2</td>
</tr>
<tr>
<td>19</td>
<td>PhD3</td>
<td>3</td>
<td>Specific</td>
<td>Self</td>
<td>2</td>
</tr>
<tr>
<td>id</td>
<td>F</td>
<td>G1</td>
<td>G2</td>
<td>H</td>
<td>I</td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
<td>-------------</td>
<td>----</td>
<td>---------------------------</td>
<td>---</td>
</tr>
<tr>
<td>20</td>
<td>PhD4</td>
<td>3</td>
<td>Self</td>
<td>2 (changed the structure)</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>PhD4</td>
<td>3</td>
<td>Not specific</td>
<td>2 (multiple places)</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>PhD4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>PhD5</td>
<td>4</td>
<td>Specific</td>
<td>1,5 (forgot the way of working)</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>PhD5</td>
<td></td>
<td>2 (wrong version)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Adm1</td>
<td>1</td>
<td>Self</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>26</td>
<td>Adm1</td>
<td>1</td>
<td>Self</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>Adm1</td>
<td>1</td>
<td>Self</td>
<td>2 -&gt; 1</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>Adm1</td>
<td>1</td>
<td>Self</td>
<td>2 -&gt; 1 misremembered details</td>
<td>6</td>
</tr>
<tr>
<td>29</td>
<td>Adm2</td>
<td>1</td>
<td>Self</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>30</td>
<td>Adm2</td>
<td>1</td>
<td>Self</td>
<td>2 -&gt; 1</td>
<td>1</td>
</tr>
<tr>
<td>31</td>
<td>Adm3</td>
<td>5</td>
<td></td>
<td>2 Drag-n-drop error</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>Adm3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>Adm3</td>
<td>1 (detached)</td>
<td></td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>34</td>
<td>Adm4</td>
<td>5</td>
<td></td>
<td>2 Drag-n-drop error</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>Adm4</td>
<td>1</td>
<td></td>
<td>Lost, re-created it</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>Adm4</td>
<td>5</td>
<td></td>
<td>Lost, with some missing</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>Adm5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>Adm5</td>
<td>1</td>
<td>Self</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
Table 8.2 (cont.)

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>39</td>
<td>Adm5</td>
<td>1</td>
<td>Self</td>
<td>2</td>
</tr>
<tr>
<td>40</td>
<td>Adm5</td>
<td>1</td>
<td>Self</td>
<td>2 Someone else took it</td>
</tr>
<tr>
<td>41</td>
<td>Adm6</td>
<td>1</td>
<td>Software</td>
<td>3 Save to default folder</td>
</tr>
<tr>
<td>42</td>
<td>Adm6</td>
<td>1</td>
<td>Software</td>
<td>3 Save to default folder</td>
</tr>
</tbody>
</table>

8.2.1 “What was it”

The numbers of “what was it” listed in the above table can be illustrated in Figure 8.1 and Figure 8.2.

Figure 8.1 “What Was It” in Two Groups

There is a case tagged with “N/A” because the participant forgot the details and thus no detailed information was provided.
8.2.2 Situation

In 4 of the 24 example cases of the PhD participants, the target documents are not for specific purposes, e.g., for a project, a course, or another type of task. The typical characteristic of this type of documents is: there was usually no corresponding folders for such files when they tried to save them, and they were usually put at a place based on the judgment call at the moment, which could be different at different times and caused re-access difficulty in retrieval time. For example, a problem a PhD student had “on more than one occasion” was the various electronic notes he took, as described in Section 5.3.1. Since he had no systematic place for these files, although he remembered the rough time he created it, the file format “plain text” with file extension “.txt” and possibly part of file name with “note” or “notes” and searched with it, he could not always find them.
For the administrative participants, this issue of “for specific purpose or not” seems not to apply in their examples. It might be because almost every information item they had in daily work was for a specific task that falls in their job content.

There were four cases in which the problems were caused by the fact that the document’s location was the default saved place by software (e.g., Microsoft Word, web Browsers) instead of the places the participants thought they were. Default folder, “temp” folder, and previously saved folder, mixed with the folder that the file was originally opened from, sometimes caused confusion when participants click “save” or download. In a very frustrating case, a PhD student reported she opened a Word file from email attachment in her web email system. Without remembering she needed to “save as” a local file at first, she started to work on that. After two hours’ work, she clicked “save” and closed it. Then she found she couldn’t find it anyhow. After wasting almost another hour looking for it, she had to redo the work again.

8.2.3 Reason

8.2.3.1 Reasons for re-access difficulties

According to the reported emails and the discussion during interviews, there were 5 types of reasons that caused the difficulties, as listed above under category H. The overview of the reasons is illustrated in Figure 8.3 and Figure 8.4, followed by details for each type of reason. The code numbers and contents listed above in H are repeated in Table 8.3 for easy reading.
As Figure 8.3 shows, more types of reasons were identified for the PhD student participants than for the administrative ones.

Figure 8.3 Reasons in Two Groups
1. **Possibly misremembered (1 case)**

There was one case in which the participant declared that she was sure she misremembered the existence of the digital document, although the reasons (1,2) and (1, 5) also include such suspicion or judgment.

2. **Forgot where it was (17 cases)**

Most of them were simply because “forgot where it was”. In some cases this led the participant to suspect that he/she misremembered some details or even the existence of the digital object.

Several of them were related to reasons such as: didn’t know “which computer” it was among multiple computers; could not find partly because re-organized the folders; scattered in multiple places; someone else took it (on a shared folder).
3. Didn’t know where it was (4 cases)

Another type of reason is that the participant did not know where it was, because of the above saved by software reason (the list G2). The cases are different from the first case in that the participant did not know the location, while in the first case, the participant knew it originally but forgot it when trying to retrieve it.

4. Know where it was, but not accessible (2 cases)

There are 2 cases reported by a PhD participant that the difficulties were accessibility, which was related to information on multiple computers or devices. The participant found that the file she had was not the editable format and had to wait till being at home to access the editable version.

5. Forgot the way of working (3 cases)

There are 3 cases reported by two PhD participants that the problem was because the participant forgot the way of working, e.g. using LaTex which caused the relevant file(s) were saved in a different folder.

1, 2. Forgot where it was, and possibly misremembered (6 cases)

There are 3 cases that the reason involves 1 and indicates 2. For example, a participant could not find a file on her computer she thought she had, suspecting that she may never have the electronic version file.
1, 5. Misremember and forgot the way of working (1 case)

There is 1 case in which the reason involves both 1 and 5. The participant “was looking for a particular article citation” that he knew he had used in a particular document for a course. But he couldn’t find it in the bibliography of the document he thought it belonged in and the several earlier drafts under the same course folder. Eventually he recalled that he had been experimenting with a citation software that used BibTex files for an earlier version of the paper, and the BibTex file was stored in another directory mainly decided by the software.

Overall, together there were 8 cases in which the participants suspected the possibility of misremembering. This highlights the human nature limitation of false memory and the corresponding need imposed by the limitation. The browsing procedure in folder structure is not simple judgment in “locating and targeting”, instead, it involves a lot of suspecting, doubting, verifying, and correcting through the process just because of this human memory problem. During the interviews, it was observed that the participants kept correcting and confirming their memories while they are browsing their folders.

8.2.3.2 Reasons for forgetting

A finer look at the reason 2 “forgot where it was” reveals the following reasons in terms of why they forgot:

1) The item was not a “part” of the organization structure, either because it had no specific purpose (e.g., notes) or there was “no place for it” at the filing time.

2) The participant re-organized files and then lost the track backward to find an old file.
3) Drag-n-drop error misplaced items. Although only several administrative participants reported this problem in emails and during interviews, both groups’ participants may have this problem as well.

Several PhD students surprisingly found files under totally irrelevant folders when they were giving the “guided tour” of their folders and files, and it was very possible that they were caused by drag-n-drop error.

4) Complicated derivative relationships are beyond the capability of our memory, as described in the previous chapter.

5) The “save” and “save as” caused confusion as reported in a couple of reported emails. It caused a much painful experience reported by a participant during interview. She was asked to do some translation for a project, and she opened the Word file from a Web email system but forgot to do “save as” to save to her computer as she usually does:

When I opened this word file, I started to translate it. I didn’t realize it. Usually I would do “save as” to save to my desktop to work on it. That day, it’s quite late at night, I didn’t do that. I didn’t know I didn’t do that. So I started to work on this, worked in a couple of hours and finished the translation. Then I closed it, and I thought I was ready to send it. But it’s gone.

The participant looked at the system’s temporary folder, then searched for it, finally traced to the default place for saving an email file, but couldn’t find the one with her translation anyhow, although she can find the original one she downloaded.
8.2.3.3 FRBR tasks in re-access on personal computers

Interestingly, the above reasons as well as element F2 (looking for a particular version of a file) cover all four generic tasks that “are performed by users when searching and making use of national bibliographies and library catalogues” as defined in FRBR: find, identify, select, and obtain (IFLA study group on the FRBR, 2009). In this report (p. 8), the four tasks are defined as:

- **Using the data to find** materials that correspond to the user’s stated search criteria (e.g., in the context of a search for all documents on a given subject, or a search for a recording issued under a particular title);

- **Using the data retrieved to identify** an entity (e.g., to confirm that the document described in a record corresponds to the document sought by the user, or to distinguish between two texts or recordings that have the same title);

- **Using the data to select** an entity that is appropriate to the user’s needs (e.g., to select a text in a language the user understands, or to choose a version of a computer program that is compatible with the hardware and operating system available to the user);

- **Using the data in order to acquire or obtain** access to the entity described (e.g., to place a purchase order for a publication, to submit a request for the loan of a copy of a book in a library’s collection, or to access online an electronic document stored on a remote computer).

The mapping between FRBR tasks and re-access behaviors on personal computers is illustrated in Table 8.4.
Table 8.4 Mapping between Re-access on PC and FRBR Tasks

<table>
<thead>
<tr>
<th>Re-access on Personal Computers</th>
<th>FRBR Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>H2 (forgot where it was); H3 (didn’t know where it was)</td>
<td>Find</td>
</tr>
<tr>
<td>H5 (knew where it was, but didn’t know which one)</td>
<td>Identify</td>
</tr>
<tr>
<td>F2 (looking for a particular version of file)</td>
<td>Select</td>
</tr>
<tr>
<td>H4 (knew where it was, but not accessible)</td>
<td>Obtain</td>
</tr>
</tbody>
</table>

8.2.4 Use

The use factors are identified with temporal properties corresponding to WordNet “temporal property (a property relating to time):

- Temporal point (arrangement of events in time)
- Duration, length (continuance in time)
- Pace, rate

Since “pace, rate” facet does not explicitly appear in the email examples, it was not included in the above coding list.

Not every case identified this “use” factor. Among the 41 valid example cases, there are 10 cases for which this factor cannot be identified or is irrelevant. Not all the time factor facets appear in a case, and the data shows that “temporal point” is a major facet. A majority of the cases (19) involve the temporal point facet “not (accessed) for a while”, while there are 9 cases involving “recently” saved/created documents. It is worth noting that the “temporal point” facet
can appear alone while the other facet (duration, length) is more of a supplementary facet in several cases.

8.3 Importance/urgency

Except in one case in which the target document was “not that important” and two where the targets were “somewhat important,” all the target documents in the other example cases were reported as important. Some of them are also “urgent.”

8.4 What were Remembered

The identified elements from participants’ description on “what were remembered” in 41 reported cases (36 files and 5 emails) are listed in Table 8.5 in descending frequency order.
<table>
<thead>
<tr>
<th>Order</th>
<th>What was remembered</th>
<th>Target</th>
<th>Examples</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Actions that were taken on the target document, may include Source and/or Target</td>
<td>Files &amp; emails</td>
<td>Created; edited/worked; copied; saved; moved/filed; detached; received; cited; downloaded</td>
<td>39</td>
</tr>
<tr>
<td>2</td>
<td>General content, specific topic (e.g., keywords) including time and place</td>
<td>Files &amp; emails</td>
<td>“describes the work of the xx office”</td>
<td>32</td>
</tr>
<tr>
<td>3</td>
<td>Time related to the actions or uses</td>
<td>Files</td>
<td>“9 months ago”</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>Format</td>
<td>Files</td>
<td>.doc, .pdf, .xsl, .txt, .bib</td>
<td>24</td>
</tr>
<tr>
<td>5</td>
<td>Genre</td>
<td>Files</td>
<td>Note; CV; reading; php code; photos; dissertation; flyer; agenda; bibliography; data for dissertation; questions; paper; spreadsheet; research statement; field exam reading list</td>
<td>22</td>
</tr>
<tr>
<td>6</td>
<td>Uses with the target item</td>
<td>Files</td>
<td>“for a course”; “for arranging a meeting”; “for a grant”</td>
<td>17</td>
</tr>
<tr>
<td>7</td>
<td>File name</td>
<td>Files</td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>8</td>
<td>directory it was possibly in</td>
<td>Files &amp; emails</td>
<td>“a couple of possible places”; “possibly general folders it may be in”</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>Sender</td>
<td>Emails</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>Author</td>
<td>Particular files</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>Subject line</td>
<td>Emails</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

It is important to note that the list and the frequencies are not meant to be conclusive and generalized. Because the list was identified from the data instead of being provided to the participants to shape their answers, it was very possible that the participants remembered more
than they included in the emails. It is also worth noting that some of them (e.g. sender) could be misremembered, as described above.

Looking further into the lists may indicate why the participants still had difficulties even though they remembered some elements. Among the most frequently remembered elements:

1) “Action” is not used by current systems to help re-access, although the action information can be automatically captured by the system;

2) The participants mostly only remembered the “general content” which were not helpful for the participants to identify keywords to search with. For example, an administrative participant was looking for “a flier that we normally give to incoming students at orientation, which describes the work of the xx office.” When asked if searching would be helpful, the participant replied in email that “(i)n this case, no. There were no unique keywords/ phrases that would have made searching a useful strategy. Such a search would have generated very many results.”

3) The temporal information was mostly rough time, and there is no way in current system to browse across folders with timeline (see more discussion in the next section).

4) Similar to 3), there is no mechanism to browse across folders with format (see more discussion in the next section).

5) There were 17 cases including use or purpose information. It was not clear if these uses or purposes corresponded to particular folders, thus it was not clear if there was any connection between the difficulties and the fact that the participants remembered the element or not.

6) For the file names, only two cases in which the participants knew the file names, 8 cases remembered “part” of the file names, and the other 3 can recognize or guess the file names.
7) In all 5 cases of emails, the participants remembered the senders, with only one knew a possible keyword in the email subject. For example, a participant described in the reporting email that “I remembered the sender’s name. I knew that I had moved it from my inbox to a folder. It was related to a meeting that took place several weeks prior but I didn’t know the words in the subject line.”

8.5 Re-access Strategies and Results

Among the 41 valid example cases, there are 5 types of re-access strategies identified. The strategies and the corresponding results are listed in Table 8.6.

<table>
<thead>
<tr>
<th>Table 8.6 Re-access Strategies and the Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Found</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Browse alone</td>
</tr>
<tr>
<td>Search alone</td>
</tr>
<tr>
<td>Browse - Search, Search at end</td>
</tr>
<tr>
<td>Browse - Search - Browse, Browse at end*</td>
</tr>
<tr>
<td>Use “open recent” function in software</td>
</tr>
<tr>
<td>N/A</td>
</tr>
<tr>
<td>total</td>
</tr>
</tbody>
</table>
* Note: this is a case that the participant kept browsing during several weeks and tried searching without finding the document. Eventually, when he was looking for something else, he “stumbled upon” a place and found the document there. This browsing step – even though it was for something else at the later time, made the overall strategy “browse – search – browse.”

Many more cases used browsing as a re-finding strategy than used search. There are together 13 cases that included “search” in the strategies, and 31 cases included “browse” in the strategies. Most of the cases (11 out of 13) including “search” strategy also used “browse,” while there are 20 cases that used “browse” strategy alone. It is important to note that “sorting” many times is combined with browsing and participants may not explicitly specify it when they described the browsing strategy. There are 3 cases that explicitly listed “sort” in browsing as their strategies.

The “open recent” strategy in the table is reported by a PhD student participant:

(Y)esterday I knew that I saved a new version (named with a new date appended to the file name). But when I went to my folder where I keep my dissertation it wasn’t in there! I didn’t freak out because I KNEW that I had saved it, I just didn’t know what folder it was in. So (I) just used the open recent function in my Word processor. Turns out it was in my conference folder. I think that when I did the save as I assumed it would automatically be saved in the same folder that the original file was opened from, but clearly the default is set to something else.

This can be seen as a small example of faceted browsing with action (“open”), format (Word file), and time (“recent”) to help re-finding a file supported by a software.
During interviews, the study observed another interesting “selective browsing” strategy which is similar to faceted browsing with format. Several administrative participants used the “Open” window of Microsoft Excel and specified the file type to browse across folders for Excel files only. This strategy of browsing across folders is especially useful for several administrative participants who had mainly Word files and Excel spreadsheets – and a fairly large number of each - on their computers. Compared to the “sorting with Type” method in the “official” browsing tools such as “My Computer” or “Windows Explorer”, this method is better targeted in this case – you don’t have to sort under each folder when you are navigating across folders. The drawback of this method is the missing search function which only exists in the official browsing tools (in WindowsXP. It has been added in Windows7). Another problem is that, if we misremember the file type, it may require more time to locate the file, which was observed during the re-access task observations at the interview with one administrative participant – she thought the target file was a Word file and could not find it. After a while, she found that it was a PDF.

Considering that browsing would become less useful and effective when the volume of items increases, this “selective browsing” or “faceted browsing” strategy can be seen as a remedy to this problem.

There are several other finer strategies behind the broad browsing and searching strategies. The first one is remembering the way of working. For example, there is a case for “Browse + Search, Search at end” strategy where after browsing for a while without finding it, the participant remembered the way of working: he used LaTex for an early version. He then searched with the
LaTeX file format (in this case, .bib) and found it. Another case with “Browse alone” as strategy was a similar situation. The participant did not find the document and then remembered his way of working with it – using LaTeX – and then went to another folder to find it.

A general but tacit strategy is inducing from topic to purpose or project which usually maps to a specific folder, as described in Section 5.3.1. Although the number of email cases is not big, the more detailed explanation and discussion during interviews showed that this is a frequently used working strategy that applies to both groups, especially the PhD participants.

Finally, “search again on Web” as another re-finding strategy was identified in the interview data although not in the reported email cases.
CHAPTER 9: CONCLUSION AND DISCUSSION

In this chapter, the findings reported in the previous chapters are summarized at first connected with general information organization and retrieval. Then the related studies described in Chapter 2 “Literature Review” are revisited and compared with the findings of this study. A following separate section outlines the implications for system design. The limitations and possible future research will be described at the end.

9.1 Summary and Discussion of Findings

Personal information organization and retrieval on computers is different from general information organization and retrieval in several aspects.

At first, different from clear curation policies and activities in creating digital library collections, the study found that the participants have difficulty in making the dichotomous decisions in keeping and deleting some types of information, as often manifested in the “messy” desktop or folders on their computers and later on re-access difficulties. It suggests that people need in-between facilities for keeping and not keeping, as well as deleting and not deleting, in order to better accommodate the vagueness and uncertainty of certain information.

The file folders on current personal computers are not only an information organization and retrieval mechanism, they are potential workplaces for projects or tasks. It has a large impact on what are saved in folders and the organization of files in folders. Extra features are thus necessary to accommodate extra needs for a workplace, e.g., keeping the work progress, keeping
the complicated relationships between files and/or groups of files, instead of treating each file as a separate item as in a library collection.

Different from a digital collection where the metadata are usually defined in certain schemas and created by metadata librarians, on personal computers people usually do not create complete metadata for their files. The folder organization is often not systematic especially because reorganization behavior on current personal computers is often incomplete.

Finally, retrieving an item on personal computers is actually re-accessing it since the item has been “found” and “kept” before, as represented in the name of the project Keeping Found Things Found (Jones et al., 2004), while in general information systems the target information item can also be a new unknown item.

However, although the way people organize information in computer folders may vary person by person, the study suggests that the perspectives with which they make sense of and organize things are remarkably similar, and are similar to the ones adopted in professional knowledge organization systems, e.g., rationalistic, empirical, pragmatic, and historical. The framework provides an overall picture of how people organize information in computer folders, and at the same time indicates a common ground between personal information organization and general information organization.

The study identified the two extreme types of folders in a continuum: the “genre folders” with no interfile relationships and the “project folders” at the other end which include complicated
relationships between files and/or groups of files. The two types of folders connect the concepts of traditional library collections and archival collections with the folder metaphor in the paper and digital worlds. Furthermore, examining the three types of relationships described in PREMIS reveals that the various derivative relationships between files led to the greatest difficulty in finding and identifying files and deserve extra support from the system.

For the information re-access behaviors, even with the small set of the difficulty examples, it was demonstrated that re-access on personal computers includes the four generic tasks in searching and making use of bibliographies and library catalogues as listed in the FRBR framework, i.e., finding, identifying, selecting, and obtaining. It refines the concept of retrieval on personal computers, justifies the use of the term “re-access” instead of “re-finding” in this dissertation, and lays out the tasks that the system should support. Re-access strategies similar to faceted browsing in general information retrieval systems are also observed in this study.

In terms of the larger context, similar to general information systems such as library online cataloging systems, personal information organization and re-access are affected by the Web and paper world, e.g., two participants organized personal pictures or bookmarks at flickr or delicious; several participants sometimes searched again on the Web instead of looking for an item on their own computers. In addition, with similar traditional hierarchical organization structure, and similarly being challenged by new technologies such as tagging, and even similarly questioned as to “whether or not we need organization,” research studies in both personal and general information organization & retrieval areas can benefit each other in improving understanding and providing implications for system design.
9.2 Discussion of Findings in Relation to Other Studies

9.2.1 Keeping and discarding

Similar to the study on GrayArea (Bergman et al., 2009) described in section 2.2.3, this study observed people’s need to differentiate between unimportant and important information. However, instead of focusing on the differentiating mechanism under each folder, this study identified the need in the level of personal information space at the stage of keeping, and refined the observation of these “unimportant” files. Specifically, people intentionally leave certain files separate from their main folders and do not see them as “kept” items even though the files literally sit on their computer desktops or in particular folders. There are two different types of files in this special group that demand different in-between keeping mechanisms: “possibly useful” and “to be deleted.” This study suggests that the “low value but expensive cleaning cost” of these files in the current binary keeping mechanism exacerbates the difficulty of deciding “keeping or not,” results in the many “messy” desktops or folders, and explains the reluctance of “cleaning” them up. The study also identified the similar need at the discarding stage which can address the difficulty of “discarding or not”.

Furthermore, these in-between mechanisms represent the investigator’s observation that the current dichotomous mechanisms dominating information management systems need to be refined to provide increased granularity.
9.2.2 Why to organize and why people prefer browsing

The available studies described in section 2.3.1 mainly identified the following reasons of why we organize information:

1) Ease of finding;
2) Reminding;
3) Sharing;
4) Emotional reasons, e.g., fears and anxieties, identity construction, creating a legacy, etc;
5) Describing content and context;
6) Way and by-product of making sense of information;
7) Providing an overview of a personal information space.

In addition to the observations confirming the above findings, this study identified several new as well as refined reasons:

8) We need to group related information for a particular project or task, and we need to see what we have for a particular work as we use a folder as the workplace;
9) Confirming Jones’ point that organization is both a way and a by-product of making sense of things (Jones & Teevan, 2007, p. 35-56), this study observed that people make an effort to impose their interpretation and view of their information space into the folder structure. The connection between making sense of information and a particular organization structure is not only manifested in this study’s findings (the study’s target), but also reflected in this study’s method, and in Star’s research on the similarity between a research method and faceted analysis, as described in Chapter 3. It may also explain the anthropologist’s findings on the tribal classification systems as described in section 2.3.1.
In a more specific look at this process of “making sense of information in/as organizing”, this study indicates that people embed their rationalistic, pragmatic, empirical, and historical views into their folder structures and information organization, although different people see different things under each perspective. This *particularity* leads to and is demonstrated in their particular way of organization.

Although people have the need to access an item from multiple access points, it does not always mean that they want to change their folder structure or do multiple-classification with the target information items. A fundamental question emerging from this is what is the relationship between classifying and describing, e.g., when do we need classifying and when do we need describing? What are the purposes? And how should the system accommodate or separate these two mechanisms? These questions will be explored in future work.

In addition to the above findings at the organization stage, the study recognizes that the question of “why people organize” is also closely related to the question of “why people prefer browsing to searching” in the re-access stage because browsing *sometimes* is based on a particular organization (browsing does not necessarily require classifying. It can be based on a particular describing mechanism. This is related to the above mentioned future research on the relationship between classifying and describing). Available studies observed people’s preference of browsing to searching even though more powerful searching tools are available (e.g., Whittaker, 2011). With the findings on information re-access stage in this study, the following identified reasons
can help explaining why we need a particular structure of information and why we prefer browsing:

1) We may be looking for a group of files instead of a single one;
2) We need the tacit context built in the current folder system, in addition to the partially “tagged” content and context, to help us not only re-finding files, but also selecting or identifying among a group of files;
3) Because we are human, and we may remember vaguely, misremember, or forget the clues to the target file/files (which happens a lot as observed in this study), we need a reassuring and correcting mechanism at each step of approaching the target(s);
4) Hiding behind a folder is a way to clear the clutter and deal with information overload.

9.2.3 How to keep/organize: “piles” and “files”

As described in section 2.3.4, “files” and “piles” in the paper world have been investigated in many studies trying to inform system design on computers. Specifically, the concept of “files” has been connected to the current folder metaphor on computers, and there has been debate on whether we need a digital “piles” metaphor or not. Although recognizing the similarities in the general level, this study’s findings indicate the need to emphasize more on these concepts’ differences between paper and digital realms. The digital folders, although the concept came from the paper folders and are named “folders,” are functioning much differently from the traditional office folders in the paper world. They serve as workplaces in addition to the information storage and organization utility. In addition, this study also suggests that the “piling” phenomenon observed in the paper world as well as on computers does not necessarily mean that
we need “piles”. Instead, it is a compensation strategy out of certain needs and deserves a refined look. The identified two types of in-between keeping strategies represent such effort.

9.2.4 The three types of information

The three types of information identified by Barreau and Nardi (1995) and the various similar patterns observed in other studies (see section 2.3.6) involve different temporal properties and sometimes relevance property. According to WordNet, temporal property (a property relating to time) includes three dimensions:

(1) temporal arrangement or order (recency);

(2) duration and length; and

(3) pace or rate (frequency).

For example, according to the definitions of the ephemeral information, working information, and archived information provided by Barreau and Nardi (1995) quoted in section 2.3.6, the “ephemeral information” involves “duration and length” (“has a short shelf life”), the “working information” involves frequency (“frequently-used information”) and “duration and length” (“has a shelf life of weeks or months”), while the “archived information” involves all three dimensions as well as relevance. Some types proposed in other studies may include the first dimension of temporal property, the usefulness, etc.

This study also observed three types of information among those that the participants deemed as their own folders and files: “work-in-progress,” “main folders,” and “archive” which are similar to Sellen and Harper’s “hot,” “warm,” and “cold” documents, and consistently involve “recency” and then “frequency.” Combined with the findings in the keeping stage, the overall diagram of
information types can be illustrated in Figure 9.1. The “work-in-progress” and “main folders” are explicitly separate on some participants’ computers while not on some others’.

**Figure 9.1 Types of Information in PIS**

9.2.5 Re-access behavior

Although factors such as time, use, and topic have been identified in many studies on how people rationalize their organization and tell the story of information items, as described in 2.3.3, little research connects them with the re-access stage and investigates how they are used or whether they are useful in re-access, especially when people have difficulty in finding information. In this study, the example cases described in the reporting emails demonstrate that although often people mentioned the time or topic attribute of the target items, the remembered contents often were too rough and general to be of help in locating the items in their folders and perhaps in any systems relying on explicitly represented or tagged systems. Some remembered attributes (e.g., genre, use, and actions) are rarely codified in the current folder systems. Overall, in the difficult
cases, there was a remarkable mismatch between what features were remembered and what were built in the folder system. Even the last resort – searching cannot help.

On the other hand, it is worth noting that although these difficulties were reported, most of the participants said that they only had difficulties occasionally and the folder system served their needs quite well. It was verified in the observations of re-access task observations and the several positive re-access experiences described in reporting email. Problems and failures are usually heard loud and clear, successes are kept quiet. A big part of such successes is the affordance for various mundane innovations or workarounds, as observed in this study. For example, the various forms of “faceted browsing” observed in this study within a particular application across folders indicate such affordance in folders, although from the system design perspective it indicates that the system should explicitly support various faceted browsing mechanisms.

Considering the similar access strategies and similar access tasks (the four FRBR tasks) between web-based information systems and personal information management systems, it suggests the need and possibility of connecting the PIM research with the research area of information organization and management.

9.2.6 Reflection on the research method and perspectives

In explaining the reasons that help explain “why we know so little” about personal information management behavior on computers, Whittaker (2011) pointed out the difficulty in gathering data because of the intrusiveness of the study and privacy issues. In this study, the adoption of a multiple-case study as research method, and specifically the use of system built-in commands,
multiple data sources and data collection methods, especially the selection of the two groups located at or close to the two ends of the activity spectrum, demonstrate a low cost, less intrusive, but productive research method.

This study also identified the challenge of analyzing and describing people’s personal information management behaviors on computers because many of them are non-categorical phenomena. This study made a fruitful effort by identifying two end types of folders in a spectrum (‘‘genre folders’’ and ‘‘project folders’’) in investigating the participants’ folder structures and contents.

Furthermore, because of the complexity of the research target, this study brings in multiple perspectives in the investigation, e.g., human computer interaction (HCI), knowledge organization, digital preservation, and information access and retrieval. These multiple perspectives provide flexible and powerful lenses in exploring different aspects of people’s information organization and retrieval behavior on computers. However, in reflecting upon the study method after the study, it is realized that the use of these multiple perspectives is not coincidental, because the personal information management system on computers, and probably also true for any information management system, is a HCI system, a knowledge representation system, a knowledge organization system, an information access and retrieval system, and a digital preservation system.
9.3 Implications for System Design

The improved understanding described above at the same time provides implications for system design which are outlined here:

1) Providing different interactive “views” of personal information space. As implied by the rationalistic element in organization method (Section 5.4.2), for example, a “map” view would provide an overall picture of the personal information space and zoom-in and zoom-out can be used in navigation instead of the current step-by-step (click-by-click) navigation. Other variations of a map view such as an interactive heatmap view would provide more powerful and informative visualizations that people may need. As indicated in looking at the folder from the pragmatic perspective (Section 5.4.1), and related to the difficulty the participants had in re-organizing files (Section 7.4), a faceted navigation view with folder or file names as tags may serve better different people’s needs at different stages;

2) As suggested in Chapter 4, better mechanisms for “possibly useful” and for “to be deleted” as well as different levels of deleting can help reduce the clutter on personal computers and further make re-access easier;

3) As illustrated by the empirical element in organization method (Section 5.4.3), a better “waiting” mechanism is needed for items that rely on the “empirical” organization method;

4) The system should capture certain metadata that are important for re-accessing and understanding information items, specifically, the actions taken on the item and the derivative traces between files, as suggested by the historical element in Section 5.4.4 and then described and discussed in more detail in Chapter 6;

5) Since the folder system is serving as workplaces in addition to storage and organization places (as described in Chapter 6), better work environment support in folders is needed. For
example, as discussed in Section 6.5.2, a better mechanism is needed to separate and connect “work-in-progress” items from “main folders” and “archive”, and corresponding functions are needed for particular types of information (e.g., highlighting, planning, and backup for work-in-progress items);

6) With the participants’ “mundane innovations” of conducting “faceted browsing” in folders (browsing across folders with one or several elements), as described in Section 8.5, a more powerful “faceted browsing” mechanism deserves system designers’ consideration;

7) As described in Section 4.3, the participants want to separate system files from their own files, which should be taken into account in system design;

8) Considering the errors that the current drag-n-drop mechanism is causing in folders (Section 8.2.3.2), a better mechanism to do drag-n-drop is needed to avoid mistakes;

9) Several issues indicate the problem of the current “default saving” mechanism, e.g., Section 4.3 on keeping “possibly useful” information, Section 8.2.2 on the difficulties it caused in re-accessing files, and Section 8.2.3.2 on the confusion between “save” and “save as” that caused a big problem for the PhD student. All these suggest that a better mechanism for the “default save” places is needed;

10) As described in Section 8.2.3.1, the participants relied on the folders to reassure and correct faulty or vague memories during the process of re-accessing. This deserves more explicit and careful considerations in designing a better personal information management system.

9.4 Limitations and Future Work

Limited by the time and scope, many issues were left out of the discussion of this dissertation, such as a quantitative analysis of the folder structure and a micro level analysis on the folder and
file names and their relationships with tags. This study is from a particular perspective (folder structure and content), for a particular purpose (exploration), and with a particular focus (the use of folders in the four organization stages). Some topics are not covered. For example, information fragmentation is not discussed in this dissertation although there are some findings on this issue. The main reason is that the available studies especially William Jones’ study discussed this issue thoroughly. Another reason is that I personally think that information fragmentation is more of a hardware issue or a system architecture problem, and can be better handled by technologies such as cloud computing.

Because of the exploratory nature of this study and the small sample, the implications of the study need to be verified with larger samples. Limited by the time and scope in the dissertation work, there are some questions unanswered and issues unexplored:

1) Further studies are needed to extend and refine Hjørland’s typology as it applies to personal information management.

2) As described in previous sections, further investigation is needed around the relationship between classifying and describing, which would help answer the questions such as how to use tagging in a classification system, the folder’s affordances, and how to represent context.

3) Related to 2), for the specific metadata about files on personal computers, large sample studies are needed to verify and refine the elements that people remember best, especially for “purpose” or “use” of the item. It would be interesting to see if there exist some “prototype concept” elements, e.g., genre (what it is) and purpose (what it is for).

4) Large sample studies are needed to explore re-access tasks beyond the FRBR tasks, although it might be difficult to collect the data except observing over a period of time.
5) This study provided implications for system design. Cooperative studies with HCI visualization techniques can further test, develop, and implement prototype mechanisms with these implications.

6) More studies are needed to investigate the applicability of the theories about information seeking behavior in personal information organization and re-access behavior.
REFERENCES

Department of Computer Science, University of Toronto.


APPENDIX A: RECRUITING EMAIL

Hello,

I am writing to invite you to participate in my dissertation study that investigates the problems people have in re-locating information on their computers and explores the possible reasons related to the file systems on personal computers. This research study is intended to deepen the understanding of information access on personal computers, and provide implications for system design to make information more accessible on computers.

You can find detailed information about the study in the consent form at:

https://netfiles.uiuc.edu/hzhang1/Share/HongZhang_Consentform.html

which has been approved by IRB. We need six administrative staff members and six graduate students to participate in the study.

If you are willing to participate in the study, please answer the following questions and your responses will help us select the participants. It will take you no more than ten minutes.

a. Do you use a lot of computer files for daily life or for work?

b. Did you have the experience that you couldn’t find a particular file on your computer?

c. If you did have such experiences, could you describe an example case including the
following information?

i. What were the circumstances surrounding the file? E.g., what is it about and what was it for, how often did you use it, was it important for you, and where did you get it?

ii. Why were you looking for it? Related to your work/study or something else? Was it important or urgent?

iii. What could you remember about the file? E.g. filename, source, actions you took on it (edit, copy, save, create, move, etc.) and/or the rough date, keyword, topic, format, etc.

iv. What strategies did you try to find it?

v. Did you find it?

Thank you very much. I am looking forward to your response. A reply by December 14 would be most helpful.

Best regards,

Hong Zhang

Ph.D. Candidate

Graduate School of Library and Information Science

University of Illinois at Urbana-Champaign
Hello everyone,

Thank you very much for your previous responses. I'm sending this email to ask if you had any access difficulties during the last several weeks. If you have, please describe the situation covering the below five questions; and if you haven't had any difficulties during the last two weeks, please describe a most recent file access experience on your computer guided with the same questions as below.

Thank you very much for your participation. Have a nice day,

Hong

--

Questions for re-accessing information situation:

1. What were the circumstances surrounding the file? E.g., what is it about and what was it for, how often did you use it, was it important for you, and where did you get it?

2. Why were you looking for it? Related to your work/study or something else? Was it important or urgent?

3. What could you remember about the file? E.g. filename, source, actions you took on it (edit, copy, save, create, move, etc.) and/or the rough date, keyword, topic, format, etc.
4. What strategies did you try to find it?

5. Did you find it?
APPENDIX C: CONSENT FORM

You are invited to participate in a research study that investigates the problems people have in re-locating information on their computers. My name is Hong Zhang, and I am a Ph.D. student in the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, working with Professor Linda Smith. This research study is intended to deepen the understanding on information re-access on personal computers, and provide implications for system design to make information more accessible on computers. The participants include six administrative staff members and six graduate students. You must be 18 years old or older to participate.

If you agree to participate in this study, you will be asked to provide example cases during a two month period about the difficulties you have experienced in re-accessing a file on your computers. Emails will be sent every two weeks to remind you if you do not submit example cases. If you do not have difficulties in re-accessing during that two weeks, you will be required to describe the most recent file access experience on your computer. You will also be interviewed twice in your office or a place you like, with the second one conducted two months after the first interview. Each interview will last about 1.5 hours and will be audio recorded. During each interview, you will be asked to show several directories on your computer and answer questions about the general content and access frequency and strategy. Then depending on whether you feel comfortable with it, you will be asked to either run the list command (i.e. “dir /s /ta /od” on Windows and “ls –R –u –lt” on Unix) under each of these directories, or show the investigator these directories so rough numbers of the files, directories, and the depth of directories can be noted. After that, you will be asked to re-access several example files (selected
according to the information you provide in the interview) as an experiment. Finally you will be
asked about your personal view of your file system on your computer. You will be asked to talk
aloud during the interviews and they will be audio recorded. If you have an office, you will be
asked to briefly describe the layout of the information in the office. Notes will be taken to
supplement the audio recording if necessary.

All people’s names in audio recordings, emails, and the list command results will be replaced
with codes during transcribing or coding. Only the Responsible Project Investigator and the
Investigator will have access to the data collected (audio recordings, the transcriptions, emails,
and field notes). Data collected in the study will be used confidentially. Pseudonyms will be used
in the investigator’s dissertation and related conference presentations and journal publications.

Your participation in this study is voluntary, and you may discontinue participation at any time
during the study without penalty. You can request turning off the audio recorder at any time, as
well as destroying and deleting a list command result. You can skip any files or directories that
you do not want to be included in this study, and you can decline to answer any questions with
which you feel uncomfortable. The decision to participate, decline, or withdraw from
participation will have no effect on your grades at, status at, or future relations with the
University of Illinois.

There are no known risks in this study beyond those of everyday life. However, it is possible that
you might feel uncomfortable with showing the directories/files names and talking about the
general content in them, although you can skip those directories and files or decline to answer any of the questions. You might also feel uncomfortable with being audio taped.

A $50 gift card from a local grocery store will be provided as compensation at the end of the study in recognition of your participation. The amount will be decreased proportionally to time period if you withdraw prior to completion of the study. Your participation in this study will contribute valuable information to deepen our understanding and provide implications for system design which will make our information more accessible on computers.

If you have further questions about the research, you may contact:

Linda Smith, Prof.
Graduate School of Library and Information Science, UIUC
112 E LIS Bldg
Mail code: 493
501 E. Daniel St. Champaign, IL61820
217-333-7742
lcsmith@uiuc.edu

Or:

Hong Zhang
Graduate School of Library and Information Science, UIUC
212 LIS Bldg
Mail code: 493
501 E. Daniel St. Champaign, IL61820
217-369-9417
hzhang1@uiuc.edu

If you have any questions about your rights as a participant in this study, please contact the University of Illinois Institutional Review Board at 217-333-2670 (collect calls accepted if you identify yourself as a research participant) or via email at irb@uiuc.edu.
Sincerely,

Hong Zhang

Ph.D. candidate in GSLIS, UIUC

Your signature indicates that you have read and understood the consent form, you certify that you are 18 years old or older, and voluntarily agree to participate in this study. You may withdraw at any time after signing this form. You will be given a copy of this form.

Signature of participant

Date
APPENDIX D: GUIDED QUESTIONS FOR INTERVIEWS

The guided questions for question answering part in interviews include:

1. Please simply introduce your background, daily studying and working activities and content, your computer use (at the first round interviews), and any changes since the last interview (at the second round interviews).

2. Please show me and explain your file system on your computer. How do you generally organize your disk (e.g. software on one disk and data on another)? Does your study/work/daily life involve a lot of file re-access? The number of first level directories, their names, and the general content, purpose. The general depth of the directories.

3. What are the projects/tasks and corresponding directories you are currently working on? Could you show me the directories? The number of subdirectories and the depth. What are the files paralleling with the subdirectories? The number of them. What are the files you are currently working on or using a lot? What are the other files at the same directory? Did you change the structure (create, delete, move a directory) recently? Why? How do you re-access a file in this project? Did you experience any difficulty in finding a file related to this project/task? Are there any connections between the directories/files and any other directories/files?

4. What are the projects/tasks you have finished? When did you finish it? Could you show me the corresponding directory? Did you access a file within it after the project/task was finished? What did you do to access it? Did you change the structure after the project/task was finished? Did you add, delete, or move a file within it? How do you re-access a file within it generally? Did you experience any difficulty in finding a file related to this
project/task? Are there any connections between the directories/files and any other directories/files?

5. Where do you save the files that have no relationship with your study or work, or the files having no corresponding directory available? Could you show me the directory? The number of subdirectories and the depth. The files paralleling with the subdirectories. Did you access a file within it recently? What did you do to access it? After you built the directory and subdirectories originally, did you change the structure ever? Did you add, delete, or move a file within it? How do you re-access a file within it generally? Did you experience any difficulty in finding a file under the directory? Are there any connections between the directories/files and any other directories/files?

6. Do you have any archived material? Do you think you will re-access it sometime? If not, do you still want to keep it? What will be the general strategy you would use if you will re-access it? If you actually did access a file within it after the archive was created, what did you do to access it? Are there any connections between the directories/files and any other directories/files?

7. (For subjects having offices) Could you explain a little about the paper files and piles in your office? Do you have electronic versions on your computers? What are they for? How frequently do you use them?

8. (For subjects having failed example cases) What do you think is the possible reason for that?

For the experiments, the target files will be selected based on the information obtained during the interviews. Subjects will be asked questions such as “did you have a file about XX (e.g. LaTex)”,
“did you submit a paper for the conference (e.g. SigIR)”, “what is the data file your current project is using”, and so on.

For the personal views of PIS structures, subjects will be asked questions such as “forget about the directories on your computers, what your information on your computer is like as you think about it”, “you can draw a graph of it or a part of it”, “what bothers you in your file system structure”, “what do you think is missing in the current hierarchical file system”, and “what do you hope the structure of the information on your computer is like”.
APPENDIX E: SOURCES IN NVIVO
# APPENDIX F: FREE NODE LIST IN NVIVO

<table>
<thead>
<tr>
<th>Name</th>
<th>Sources</th>
<th>Reference</th>
<th>Created On</th>
<th>Created</th>
<th>Modified On</th>
</tr>
</thead>
<tbody>
<tr>
<td>File - naming tells something</td>
<td>1</td>
<td>1</td>
<td>8/13/2009 1:39 PM</td>
<td>HZ</td>
<td>8/13/2009 1:39 PM</td>
</tr>
<tr>
<td>File name - from other people</td>
<td>2</td>
<td>2</td>
<td>7/31/2009 11:56 AM</td>
<td>HZ</td>
<td>8/2/2009 1:0</td>
</tr>
<tr>
<td>File name - lost the control relationship although it got the keywords</td>
<td>1</td>
<td>2</td>
<td>8/2/2009 12:40 AM</td>
<td>HZ</td>
<td>8/2/2009 2:1</td>
</tr>
<tr>
<td>File name - purpose - vs what it is - external naming requirement</td>
<td>5</td>
<td>8</td>
<td>7/15/2009 10:37 AM</td>
<td>HZ</td>
<td>10/26/2009 1:0</td>
</tr>
<tr>
<td>File names can serve as folder function - group</td>
<td>2</td>
<td>2</td>
<td>7/18/2009 1:36 PM</td>
<td>HZ</td>
<td>8/3/2009 3:0</td>
</tr>
<tr>
<td>File relationship - which reference used for which paper</td>
<td>3</td>
<td>3</td>
<td>7/14/2009 3:19 PM</td>
<td>HZ</td>
<td>8/13/2009 1:0</td>
</tr>
<tr>
<td>File relationship - use - old file as a template or reuse</td>
<td>3</td>
<td>6</td>
<td>6/16/2009 12:20 PM</td>
<td>HZ</td>
<td>8/8/2009 2:1</td>
</tr>
<tr>
<td>file - no place to put it - small tasks marginally related to an ar</td>
<td>4</td>
<td>3</td>
<td>7/30/2009 11:50 AM</td>
<td>HZ</td>
<td>8/12/2009 3:0</td>
</tr>
<tr>
<td>file relationship names &amp; files next to - as tags for search and understand</td>
<td>3</td>
<td>5</td>
<td>7/30/2009 11:50 AM</td>
<td>HZ</td>
<td>8/12/2009 3:0</td>
</tr>
<tr>
<td>files - derived in two folders - started as course work and end u</td>
<td>1</td>
<td>1</td>
<td>7/26/2009 12:00 PM</td>
<td>HZ</td>
<td>7/26/2009 3:0</td>
</tr>
<tr>
<td>Filing - one thing relates to multiple folders</td>
<td>5</td>
<td>6</td>
<td>7/16/2009 6:21 PM</td>
<td>HZ</td>
<td>8/12/2009 5:0</td>
</tr>
<tr>
<td>Filing - person who I am working with (for sl)</td>
<td>1</td>
<td>2</td>
<td>8/11/2009 1:30 PM</td>
<td>HZ</td>
<td>8/11/2009 2:0</td>
</tr>
<tr>
<td>Filing - sth - new of current file system - no systematic</td>
<td>1</td>
<td>3</td>
<td>7/14/2009 12:22 PM</td>
<td>HZ</td>
<td>7/14/2009 3:0</td>
</tr>
<tr>
<td>Filing - where to file</td>
<td>1</td>
<td></td>
<td>8/1/2009 1:50 PM</td>
<td>HZ</td>
<td>8/11/2009 2:0</td>
</tr>
<tr>
<td>Filing - where to file according to email message</td>
<td>1</td>
<td></td>
<td>8/1/2009 12:19 PM</td>
<td>HZ</td>
<td>8/11/2009 2:0</td>
</tr>
<tr>
<td>folder - a place that is difficult to find</td>
<td>1</td>
<td></td>
<td>7/14/2009 1:49 PM</td>
<td>HZ</td>
<td>7/14/2009 1:0</td>
</tr>
<tr>
<td>folder - a way to accommodate everything with no specific purpose - a</td>
<td>1</td>
<td></td>
<td>7/14/2009 1:04 PM</td>
<td>HZ</td>
<td>7/14/2009 1:0</td>
</tr>
<tr>
<td>folder - archive vs. topic</td>
<td>1</td>
<td></td>
<td>8/12/2009 6:19 PM</td>
<td>HZ</td>
<td>8/12/2009 5:0</td>
</tr>
<tr>
<td>folder - as archive</td>
<td>3</td>
<td>7</td>
<td>7/14/2009 8:09 PM</td>
<td>HZ</td>
<td>8/13/2009 12</td>
</tr>
<tr>
<td>folder - as reminder - as trace of old machine</td>
<td>1</td>
<td></td>
<td>7/14/2009 10:43 AM</td>
<td>HZ</td>
<td>7/14/2009 10</td>
</tr>
<tr>
<td>folder - constantly in work vs. current - activity used need to be use</td>
<td>1</td>
<td></td>
<td>7/14/2009 1:30 PM</td>
<td>HZ</td>
<td>7/14/2009 2:0</td>
</tr>
<tr>
<td>folder - create as plaintext</td>
<td>2</td>
<td></td>
<td>2/2/2009 12:00 AM</td>
<td>HZ</td>
<td>7/10/2009 6:0</td>
</tr>
<tr>
<td>Folder - create when need it - maybe wrong anticipation and me</td>
<td>6</td>
<td>10</td>
<td>6/16/2009 6:00 PM</td>
<td>HZ</td>
<td>8/9/2009 2:1</td>
</tr>
<tr>
<td>folder - current way of org may not work later for re-accessing</td>
<td>1</td>
<td></td>
<td>8/4/2009 1:59 PM</td>
<td>HZ</td>
<td>8/9/2009 5:0</td>
</tr>
<tr>
<td>folder - dynamics</td>
<td>2</td>
<td></td>
<td>7/16/2009 6:33 PM</td>
<td>HZ</td>
<td>7/31/2009 3:0</td>
</tr>
<tr>
<td>folder - finals version</td>
<td>5</td>
<td>5</td>
<td>7/31/2009 1:02 PM</td>
<td>HZ</td>
<td>8/12/2009 5:0</td>
</tr>
<tr>
<td>folder - forgot about it and used another one for related things</td>
<td>5</td>
<td>6</td>
<td>6/16/2009 1:22 PM</td>
<td>HZ</td>
<td>8/12/2009 2:0</td>
</tr>
<tr>
<td>folder - general vs. specific purpose</td>
<td>2</td>
<td>2</td>
<td>6/16/2009 12:40 PM</td>
<td>HZ</td>
<td>7/24/2009 10</td>
</tr>
<tr>
<td>folder - main directories</td>
<td>1</td>
<td></td>
<td>7/14/2009 9:24 PM</td>
<td>HZ</td>
<td>7/14/2009 2:0</td>
</tr>
<tr>
<td>folder - misc test positions discarded</td>
<td>1</td>
<td></td>
<td>7/14/2009 11:00 AM</td>
<td>HZ</td>
<td>7/14/2009 11</td>
</tr>
</tbody>
</table>
APPENDIX G: TREE NODE LIST IN NVIVO

<table>
<thead>
<tr>
<th>Name</th>
<th>Sources</th>
<th>References</th>
<th>Created Date</th>
<th>Cre</th>
<th>Modified By</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delete - not delete</td>
<td>2</td>
<td>2</td>
<td>9/19/2009 12:18 PM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>Delete - retrieve from trash</td>
<td>2</td>
<td>3</td>
<td>9/18/2009 12:18 PM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>Delete - when running out</td>
<td>2</td>
<td>2</td>
<td>9/18/2009 12:18 PM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>1. Organization-structure</td>
<td>0</td>
<td>0</td>
<td>6/14/2009 12:35 PM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>2. File</td>
<td>6</td>
<td>0</td>
<td>6/14/2009 12:35 PM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>3. Folder</td>
<td>6</td>
<td>0</td>
<td>6/14/2009 12:35 PM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>4. Info fragmentation</td>
<td>0</td>
<td>0</td>
<td>6/14/2009 12:35 PM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>5. Info on web</td>
<td>0</td>
<td>0</td>
<td>6/14/2009 12:35 PM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>6. File System Structure</td>
<td>0</td>
<td>0</td>
<td>6/14/2009 12:35 PM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>7. Email files - relation</td>
<td>6</td>
<td>13</td>
<td>9/18/2009 12:01 PM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>8. Email organization</td>
<td>2</td>
<td>3</td>
<td>9/18/2009 12:01 PM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>9. Similar email and files</td>
<td>6</td>
<td>11</td>
<td>9/18/2009 10:03 AM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>10. Email - inbox</td>
<td>3</td>
<td>3</td>
<td>9/18/2009 12:19 PM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>12. Factor that matters the most</td>
<td>1</td>
<td>2</td>
<td>9/18/2009 12:21 PM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>13. Indicators of information</td>
<td>0</td>
<td>0</td>
<td>12/27/2009 4:08 PM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>14. Keeping context - activity tree</td>
<td>1</td>
<td>1</td>
<td>1/17/2010 4:23 PM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>15. Related needs</td>
<td>0</td>
<td>0</td>
<td>9/18/2009 9:07 AM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>16. Re-Organization</td>
<td>0</td>
<td>0</td>
<td>9/18/2009 9:07 AM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>17. Unit in file system - file or folder</td>
<td>2</td>
<td>1</td>
<td>9/18/2009 12:23 PM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>20. Work environment - backup specific</td>
<td>2</td>
<td>3</td>
<td>10/26/2009 12:14 PM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>21. Working on files - later file it</td>
<td>1</td>
<td>1</td>
<td>10/26/2009 12:14 PM</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>22. RE-ACCESS - USE</td>
<td>0</td>
<td>0</td>
<td>6/14/2009</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>23. Vork Environment</td>
<td>0</td>
<td>0</td>
<td>6/14/2009</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>24. Human Factor</td>
<td>0</td>
<td>0</td>
<td>6/14/2009</td>
<td>Hz</td>
<td>Hz</td>
</tr>
<tr>
<td>25. Comfortable level</td>
<td>1</td>
<td>1</td>
<td>6/14/2009</td>
<td>Hz</td>
<td>Hz</td>
</tr>
</tbody>
</table>